Risk Management in Financial Institutions
Risk Management in Financial Institutions

Edited by
Shahin Shojai and George Feiger
The views expressed in this book are the views of the authors and contributors alone and do not reflect the views of Euromoney Institutional Investor PLC. The authors and contributors alone are responsible for accuracy of content.

While all reasonable efforts were made to confirm internet references at the time of publication, there can be no guarantee that referenced articles will remain on the internet, that their content has not or will not change and that their content is accurate.

Referenced articles on the internet may require, or may require in the future, registration, subscription or the payment of a fee to access all or a portion of an article’s content. The reference to an article in a footnote does not constitute an endorsement of any website or its content or a recommendation to subscribe or register with any website to view such article and under no circumstances shall Euromoney or the author be responsible for any fees that may be occurred in connection with accessing or viewing such article.

This book contains exhibits that are based, in part or in whole, on information obtained from third-party sources. All reasonable efforts have been made to seek permission for the use of such material. Euromoney and the author bear no responsibility for the accuracy of information from third-party sources and assume no obligation to correct or update such information at any time in the future.

Note: Electronic books are not to be copied, forwarded or resold. No alterations, additions or other modifications are to be made to the digital content. Use is for purchaser’s sole use. Permission must be sought from the publisher with regard to any content from this publication that the purchaser wishes to reproduce (books@euromoneyplc.com). Libraries and booksellers and ebook distributors must obtain a licence from the publishers (books@euromoneyplc.com). If there is found to be misuse or activity in contravention of this clause action will be brought by the publisher and damages will be pursued.

Typeset by Phoenix Photosetting, Chatham, Kent
This book is dedicated to our families and the many students we have had the pleasure of teaching over the years.
# Contents

Foreword xv  
About the editors xix  
About the contributors xxi  

1 Managing the risks in financial institutions 1  
*Shahin Shojai, Applied Thinking FZE, George Feiger, Aston Business School*  
Overview 1  
Introduction 1  
MPT models do not work in practice 2  
Because of irrational investors? 2  
Securitisation, mortgages and CDOs 3  
Structured products, liquidity and systemic risk 4  
Capital market incentives 5  
Value-at-Risk 5  
Organisational structure 6  
Too much data, too little understanding 6  
The intersection of IT and moral hazard 7  
Yesterday’s spaghetti 7  
So, what should be done? 8  
Conclusion 11  
Bibliography 12  

## Part 1 – Operations and IT risk

2 Operational risk in the insurance industry 17  
*Russell Walker, Kellogg School of Management, Northwestern University*  
Overview 17  
Origin of operational risk 17  
Operational risk – notional definitions 18  
Market risk – drawing a distinction from operational risk 19  
Credit risk – drawing a distinction from operational risk 19  
Operational risk – in detail 20  
Operational risk is hard to measure 20  
Operational risk is driven by complexity 20  
Operational risk is not like other risks 21  
Operational risk – types and drivers 21  
Risk contagion from operational risk 25  
Reputational risk – a by-product of operational risk contagion 26  
Regulatory risk – another by-product of operational risk contagion 27  
Operational risk is hidden in product performance 27  

www.rasabourse.com
Product development and model development are critical processes in managing operational risk 29
Operational risks are often manifested as lawsuits and/or penalties 30
IT capabilities become increasingly critical 31
Principal-agent problems are magnified in insurance 32
Impact and importance of operational risk in insurance 33
Management actions for controlling operational risk in the insurance industry 36
Conclusion 40
Bibliography 40

The future of operational risk management in banking: from deterministics towards heuristics and holistics? 43
Brendon Young, Risk Research Foundation and Institute of Operational Risk
Overview 43
Risk management is failing 43
The banking environment is changing – risk management must adapt 44
Regulation is inappropriate 45
Beware excessive regulation and legislation 46
Basel II and III – simply more of the same 47
Basel III 48
Financial reporting is in need of reform 48
Longer-term risk management requires assessment of longevity and richness 51
Culture determines ethics – ethics guide principles – principles govern actions 52
Concentrating on customer service is not sufficient 54
Where risk management is poorly understood, standards may not be the answer 55
Risk managers prepare for the future; gamblers predict it 58
Qualification has limitations 58
Risk management works well in other industries 60
Aerospace 61
Nuclear 61
NHS 62
High reliability – its relevance in banking 64
Conclusion 66
Bibliography 68
Risk appetite definitions 69

Keeping markets safe in a high speed trading environment 72
Carol L. Clark, Federal Reserve Bank of Chicago
Overview 72
Introduction 72
Impact of technology, regulation and competition on market microstructure 74
Technology glitches 78
Life cycle of a trade 80
Contents

Risk management in ATRADE
  Risk management between strategy and OME layers 114
  Risk management policies 116
  Risk management responsibilities and ATRADE risk layers 117
Quant, IT and policy-making challenges 118
  Algorithmic trading impact 118
Lessons learned 120
  Market impact 121
  Company impact 121
  Regulatory impact 121
Cultures of risk? 124
Paul Willman, London School of Economics and Political Science,
Mark Fenton-O’Creevy, Open University Business School
  Overview 124
  Introduction 124
  Culture in organisations 125
  Risk takers 128
  Emotions and risk 130
  People, risk and environments 134
    Case study: rogue trading and cultures of risk 137
  Conclusion: ‘fixing’ risk cultures 139
  Bibliography 140
Reputation management after 2007 143
Roland Schatz, Media Tenor International AG
  Overview 143
  Introduction 143
  So, what should financial services institutions do? 156
  Bonus category 1: employee satisfaction 158
  Bonus category 2: key markets satisfaction 159
  Bonus category 3: financial markets satisfaction 160
  Bonus category 4: politics satisfaction 160
  Bonus category 5: NGO/NPO satisfaction 160
  Bonus category 6: science satisfaction 160
  Bonus category 7: media satisfaction 161
  Conclusion 161

Part 2 – Regulatory and enterprise risk

Theories of regulatory risk: the surprise theory, the arbitrage theory and the regulatory mistake theory 165
Jonathan R Macey, Yale Law School
  Overview 165
  Introduction 166
Public choice theory and regulatory risk 167
Regulatory causation risk 172
Credit ratings and systemic risk 173
Flawed reliance on VaR models 179
Concluding observations 182

9 Key principles of enterprise risk management for insurance groups 189
David Ingram, Willis Re
Overview 189
Introduction 189
Enterprise risk cycle 192
Diversification of risks 193
Underwriting of risks 197
Controlling risks with a cycle 198
Consideration for risks 202
Provisioning for risks 204
Portfolio of risks 205
Pitfalls of RPM 207
Future risks 208
Risk environment 211
Conclusion 213
Bibliography 214

10 Enterprise risk management in financial intermediation revisited 215
Stuart Greenbaum, Washington University in St Louis
Overview 215
Introduction 215
Risk, risk management and ERM 215
Risk 216
Risk management 217
ERM 218
The framework 218
Core risk 219
Ancillary risks 220
Plasticity of risks 222
Best practices in formalising the ERM process 223
ERM protocols 223
Organisational design for ERM 225
External versus internal information flows 225
Conclusion 226

Part 3 – Future

11 Re-engineering risks and the future of finance 233
Charles S Tapiero, NYU-Polytechnic Institute
Contents

Overview 233
Introduction 234
Finance and theories 235
  Big-data and the future of finance 237
  Global future finance challenges 241
  Alternative financial risks models 242
  A multi-agents’ strategic pricing model 244
  Finance and future risk externalities 248
Conclusion 248
Bibliography 250

12 The future of financial regulation 255
Patricia Jackson and Stefan Walter, EY
Overview 255
Introduction 255
Key drivers of the financial reform agenda 256
  Phase 1: market value and liquidity 257
  Phase 2: solvency and counterparty exposures 258
  Phase 3: sovereign risk and bank resolvability 259
Capital regulation 260
  Agreed changes 260
  Future direction 263
Liquidity regulation 266
Risk governance and stress-testing 268
Strengthening recovery and resolution frameworks 270
  Structural reforms 271
OTC derivatives markets – regulatory reform 273
  Background to the changes 273
  Central counterparties 274
  OTC contracts covered by mandatory clearing 276
  Requirements for non-centrally cleared derivatives 276
  Margin requirements 277
Implications 277
  Derivatives trading and reporting 278
  Progress on reforms 278
  Overall implications of the OTC reforms 280
Conclusion: the wider effects of regulatory change 280
Bibliography 282

13 Global banking after the cataclysm 286
Roy C Smith, NYU Stern School of Business
Overview 286
Introduction 286
Investment banks as intermediaries 288
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased competition</td>
<td>288</td>
</tr>
<tr>
<td>Basic business models emerge</td>
<td>288</td>
</tr>
<tr>
<td>Ten banks dominate markets</td>
<td>289</td>
</tr>
<tr>
<td>Cataclysm</td>
<td>291</td>
</tr>
<tr>
<td>Regulatory responses</td>
<td>291</td>
</tr>
<tr>
<td>Basel III</td>
<td>292</td>
</tr>
<tr>
<td>G20 bank standards</td>
<td>292</td>
</tr>
<tr>
<td>Dodd-Frank</td>
<td>292</td>
</tr>
<tr>
<td>The Swiss ‘finish’</td>
<td>294</td>
</tr>
<tr>
<td>Ring fencing</td>
<td>294</td>
</tr>
<tr>
<td>European Union regulations</td>
<td>295</td>
</tr>
<tr>
<td>Changing compensation practices</td>
<td>296</td>
</tr>
<tr>
<td>Market corrections and other responses</td>
<td>297</td>
</tr>
<tr>
<td>Bank stock prices collapse</td>
<td>297</td>
</tr>
<tr>
<td>Management changes</td>
<td>299</td>
</tr>
<tr>
<td>Reduced capital market activity</td>
<td>300</td>
</tr>
<tr>
<td>Increase in bank funding spreads</td>
<td>300</td>
</tr>
<tr>
<td>Downgraded credit ratings</td>
<td>301</td>
</tr>
<tr>
<td>Liquidity squeeze in Europe</td>
<td>302</td>
</tr>
<tr>
<td>Legal settlements and reputation loss</td>
<td>302</td>
</tr>
<tr>
<td>Diagnostics</td>
<td>303</td>
</tr>
<tr>
<td>A protracted downcycle</td>
<td>303</td>
</tr>
<tr>
<td>New structural constraints</td>
<td>304</td>
</tr>
<tr>
<td>The regulatory dilemma</td>
<td>305</td>
</tr>
<tr>
<td>Adapting to the new order</td>
<td>306</td>
</tr>
<tr>
<td>Break-up the bank</td>
<td>307</td>
</tr>
<tr>
<td>Retreat from wholesale banking</td>
<td>308</td>
</tr>
<tr>
<td>Re-engineer the trading model</td>
<td>309</td>
</tr>
<tr>
<td>Strategic mergers</td>
<td>310</td>
</tr>
<tr>
<td>Adapting to survive</td>
<td>311</td>
</tr>
</tbody>
</table>
Foreword

The Great Financial Crisis of 2007–2008 requires policy-makers, regulators and financial institutions themselves to focus – as never before – on the subject of risk, and this book makes a timely and important contribution to that effort. Excessive risk-taking at individual institutions and in the financial system as a whole helped drive the world’s financial system to the brink of collapse. The abyss was avoided only by courageous actions by a small number of individuals, which proved to be wildly unpopular politically, and by a large measure of luck.

Had we plunged into the abyss, the consequences could have extended beyond the financial system, and even our economic system, to political and social turmoil. Our socio-political system is built upon certain basic assumptions about our economic well-being and future promise. If those assumptions are rendered invalid, our political comity could be at risk. The demagogy of the 1930s had many roots, but undoubtedly a principal one was economic desperation.

Even though the most dire systemic consequences of the Great Financial Crisis were escaped, the suffering of millions of individuals was not. The ranks of the unemployed swelled to levels not seen for decades. The American dream of owning one’s home was dashed for millions more individuals.

Even more recent events demonstrate that risk extends beyond poor credit decisions and the resultant losses on loans and investments. Practices that were negligent, and in some cases shoddy, have exposed banks to billions of dollars of losses as a result of government enforcement actions and private litigation. Even more damaging has been the reputational damage. The fines and settlements, as enormous as they have been, will be earned back within a few years. The reputation loss could take far longer to recover.

As the last five years have demonstrated, risk must be addressed comprehensively, holistically and aggressively, and always with an eye to the future. It was not so long ago that all home mortgages, whatever the terms, were regarded as involving such low credit risk that they were assigned a risk-weighting 50% lower than other loans. Just three years ago, if you had polled banks and their regulators as to the greatest risks, I doubt that anyone would have mentioned cyber-security. Today, it would rate as among the top three on everyone’s list.

In addressing the subject of risk at financial institutions, there is a preliminary, but fundamental, question. Should the objective of the government and the financial institutions themselves be to eliminate risk or to control and manage risk? The devastating consequences of the risks incurred by financial institutions in recent years have led a number of observers to conclude that the objective should be risk-elimination. That goal, however, creates its own risk of damage to the bank customer populace and the economy as a whole.

In addressing this question, it is essential to understand that a bank’s fundamental function is to take two basic types of risk in supporting its customers’ needs and the country’s economy.

The first is credit risk. Banks are the principal source of credit in this country, and other countries, and their credit-providing role is, of necessity, even greater where the borrowers, such as small businesses or consumers, lacks access to the public credit markets or other
credit sources. If banks seek to reduce credit risk by tightening credit standards, the inevitable result is less credit availability, particularly for smaller borrowers and those who do not have a demonstrated pristine credit record. To state it differently, if every loan were repaid in full, numerous good loans are not being made.

The second risk is maturity-mismatch risk. Depositors and other funders are seeking to place their money short term, while the credit needs of consumers and businesses are often intermediate and long term. Banks perform the role of converting liquid funds into less liquid assets. If banks seek to match assets and liabilities more closely, the consequence will be to reduce the longer term credit that businesses need to grow and consumers need for home-buying and other long-term expenditures.

There are other risks that banking organisations incur on behalf of their customers and that also benefit the economy. One example is underwriting. The benefits of underwriting to issuers of securities and to the markets are presumably beyond dispute. Our capital markets are still the envy of the world and have enabled our companies to expand, produce and hire. The underwriting system has developed to the point where risks are normally small and well-managed. Even the Volcker Rule recognises that the economic benefits of underwriting outweigh the risks.

Undoubtedly, these key risks should be carefully managed, with robust capital and liquidity requirements, lending limits and strong underwriting and asset/liability risk management. Moreover, it is appropriate for the bank regulators to take an active role in assuring that these risks are appropriately managed. As the Financial Crisis demonstrated, if these risks spin out of control, banking organisations will no longer have the capacity to support their customers and the economy more broadly; indeed, if excessive risk-taking reduces credit availability, the economy can be put into a tailspin. It is precisely because of the centrality of banks to our country’s economy that they must take the precautions to be in a position to perform their role successfully.

This book is devoted to the objective of enhancing the ability of policy-makers, regulators and financial institutions to identify, manage and control risk so that financial institutions can fulfil their role in the economy. It eschews simplistic solutions, which actually threaten to embed risk rather than reduce it, and provides realistic solutions. For example, some observers argue that the banking system was much safer before there was substantial industry consolidation and product and geographic expansion, particularly following the Riegle-Neal Act of 1994 and the Gramm-Leach-Bliley Act of 1999. This idea of a halcyon past is refuted, however, by a clear-eyed view of what actually happened. Between 1982 and 1993, for example, there was an extraordinary wave of failures, as approximately 2,300 depository institutions, including approximately 1,650 banks, failed. Nor were these failures limited to small banks. When Continental Illinois failed in 1984, it was not only the country’s seventh largest bank, but had been widely lauded as one of the country’s most progressive and innovative banks. Other major failures included First Republic and Bank of New England.

Effective risk management is instead, to borrow football terminology, all about blocking and tackling rather than ‘Hail Mary’ passes into the end zone. It is hard work that requires conscientiousness and culture. In addition to the recommendations that are made by the contributing authors, I offer the following nine observations.

First, although both Congress and other regulators have appropriately focused on risk management, superior risk management cannot ultimately be legislated or regulated. It must
be a function of a company’s own culture and commitment. To that effect, a financial institution’s board should be periodically asking the following basic questions:

1. Are my company’s risk parameters appropriate? How do they correlate to reward?
2. Do the company’s risk management personnel have the appropriate skills, expertise and independence to manage risk?
3. Does the board itself have sufficient expertise and information to exercise its oversight role over risk?
4. How can we as a board assure ourselves that the company is complying with the risk parameters we have approved?

Second, board-level expertise in effective risk management at financial institutions obviously requires knowledge of the particular risks encountered by the particular type of financial institution. This can create a dilemma because the persons with the most relevant knowledge are generally employed at other financial institutions, and there are significant restrictions under the Federal Reserve’s Regulation L and the Clayton Act on director and management interlocks between financial institutions. One, at least partial, answer to this dilemma would be to consider a higher maximum age limit for directors in order to take advantage of the expertise of retired financial executives and regulators.

Third, risk management must be enterprise-based rather than business-confined. Because of the numerous ways in which risk can be created and incurred, risk exposure to similar circumstances can exist in multiple business lines. This risk can only be evaluated from the top on an overall basis.

Fourth, the boards and executive management of financial services companies must understand one fundamental principle relating to risk. When a particular business line or unit is experiencing outsized growth in revenues or profits, that should be a clear warning signal and not just a cause for celebration. With all due respect, it is unlikely that one group of bankers is so much smarter than everyone else or has discovered some magic formula that explains their extraordinary performance. In most cases, the simple explanation is that these bankers have gone further out on the risk curve. If an athlete sharply improves his or her performance, there is a suspicion that the individual has taken steroids. Excess risk is the steroids for financial institutions.

Boards should insist on a thorough evaluation of the risks in any unusually high-performing business operation. An inability to understand what is often a complicated business strategy is a reason for more rather than less scrutiny.

Fifth, if the board becomes aware of a likely or actual problem, particularly in the area of compliance, investigate it promptly and thoroughly. It is essential to avoid compounding the risk. The financial institutions that have got into the most trouble have often done so less because they flunked the underlying conduct and more because they flunked the investigation.

Sixth, a financial institution is placed at heightened risk when there is a vacuum of leadership. A board must recognise its obligation to provide for management succession, not only in the normal course but when the CEO is hit by the proverbial bus – which often happens when the CEO steps into a busy street without looking. In several high profile cases, the CEO was dismissed because of a failure to control risk, but the company was placed
in peril when many months elapsed before a successor was appointed. The institution was
adrift during that interim period. Some CEOs have suggested that they have a sealed envelope
dealing with succession locked in their desk drawer. I would suggest that boards also have
that envelope, and the contents should be the result of careful evaluation.

Seventh, in a consolidating industry, the depth of the risk in acquisitions must be recog-nised and managed. This risk is not just a reduced EPS or ROE if the transaction does not
pan out as anticipated. In recent years, a number of banks of all sizes have been sunk or
nearly sunk by misbegotten acquisitions involving serious credit or other problems at the
target – Wachovia, BofA, National City, Lloyds, Huntington and Citizens. More recently, prior
compliance violations at the target that come to light after the acquisition is consummated
are providing no insulation or immunity for the acquirer. The adequacy of due diligence
should trump every other consideration – including time and possible loss of the deal itself.

An eighth risk that must be fully recognised is the implication of the current enforcement
environment. The risk of a violation being determined by the regulators or law enforcement
authorities has sharply increased, the response of the authorities to violations has become much
more aggressive, and the authorities have been imposing monetary penalties that are literally
exponentially greater than ever before existed. The potentially most threatening aspect of this
new enforcement approach is the Department of Justice’s insistence that two banks – UBS and
RBS in connection with Libor settlements – accept a guilty plea to a criminal charge, albeit by
a small affiliate. The consequences of a guilty plea by a bank can be potentially life-threatening.

Financial institutions must understand – and respond to – this enforcement phenomenon,
which shows no signs of abating and, indeed, may well continue to intensify. The regulators
and law enforcement authorities are themselves under siege from legislators and the media.
They are criticised for being insufficiently harsh and insufficiently diligent, and for failure to
bring criminal charges against financial institutions and their executives. In recent Reports of the
Permanent Subcommittee on Investigations on HSBC’s alleged money laundering activities and
JPMorgan’s derivatives trades, sharp criticism was directed at the regulator as well as the banks.

Ninth and last, but of perhaps most importance, the public and private sectors need to
work collaboratively to develop constructive solutions to risk management. There are many
complex problems that require nuanced and balanced responses, and those responses can be
best shaped if multiple perspectives and knowledge pools are brought to bear.

A predicate to this approach is collaboration within the financial industry itself. Furthermore,
industry co-operation could be useful in enhancing the industry’s reputation. For example, the
industry could develop guidelines in such areas as customer products and services and anti-
money laundering. In the long run, the issue should be whether the industry is benefited, not
whether a particular institution’s susceptibility is greater or less than its competitors.

In summary, achieving sufficient management and control over risk to prevent damage to
the financial system, while minimising the damage that this effort reduces credit availability
and other productive bank services, should be of the highest priority. It requires a proactive
approach and recognition that the search should be for solutions rather than villains.

H Rodgin Cohen
Senior Chairman
Sullivan & Cromwell LLP
July 2013

www.rasabourse.com
About the editors

Shahin Shojai is the Chairman and CEO of Applied Thinking FZE, and until recently the Global Head of Strategic Research at Capco, Director of the Capco Institute, and the Founding Editor of the Journal of Financial Transformation. Shahin was also the Founding Director of the Cass-Capco Institute Paper Series on Risk, the SFI(Swiss Finance Institute)-Capco Institute Paper Series on Banking, and the Zicklin-Capco Institute Paper Series in Applied Finance. In addition to his role as the head of the Capco Institute and its many partnerships, Shahin was a senior subject matter expert and a member of Capco’s Private Client/Asset Management practice. Shahin was also a Senior Professor in Finance and Strategic Management at the Institute of Management Technology.

Prior to joining Capco, Shahin was President and CEO of WRA Group, a syndicated research company providing research and advisory to the treasurers and CFOs of Fortune 500 companies. Before that, Shahin was Director of the European Institute, a think tank for CEOs of the 92 largest institutional asset management firms in Europe and North America.

Previously, Shahin was a member of the world of academia, holding visiting or full professorial appointments at renowned academic institutions around the world, including London University, Boston University, Paris Graduate School of Management (Groupe ECSP), Universite Paris-Dauphine, Syracuse University, and the American University in London, where he was the chairman of the finance faculty.

Shahin holds a PhD in Finance and Economics from the University of Nottingham, and an MBA in Finance from City University Business School. Shahin is a member of SSRN Top-50, which represents the 50 most downloaded economic scholars on the Social Science Research Network.

George Feiger is the Executive Dean of the Aston Business School at Aston University in Birmingham. Prior to assuming this position George was the founder and CEO of Contango Capital Advisors, a US wealth management company and also Executive Vice President of Zions Bancorporation, Contango’s principal investor.

George’s other positions in financial services include Global Head of Onshore Private Banking at UBS, Global Head of Investment Banking at SBC Warburg and member of the Group Management Board of both UBS and Swiss Bank Corporation. George was a Director of McKinsey & Co.

In his prior academic life he was Associate Professor of Finance at Stanford’s Graduate School of Business. He holds a PhD in Economics from Harvard University where he was a Junior Fellow of the Society of Fellows.

He has authored or co-authored a variety of academic papers and a text book on International Finance and has been a frequent media contributor on the subjects of economics and capital markets.

www.rasabourse.com
About the contributors

Carol L Clark
Carol L Clark is a senior policy specialist in the financial markets group at the Federal Reserve Bank of Chicago. Her work focuses on public policy issues relating to trading (high speed trading, high frequency trading and algorithmic trading), market microstructure, clearing, settlement and payments. Prior to joining the financial markets group, Carol Clark held a variety of positions at the Bank including national market research manager in the customer relations and support office, payments research manager in the economic research department, examiner in the capital markets unit, research co-ordinator in the national book-entry securities office and analyst in the loans and reserves department. Her research has been published in the Journal of Payment Systems Law and the Federal Reserve Bank of Chicago’s Chicago Fed Letter and Economic Perspectives.

Carol Clark has a BA from St. Xavier College and an MA from the University of Chicago.

Mark Fenton-O’Creevy
Mark Fenton-O’Creevy is Professor of Organisational Behaviour and Associate Dean (International) at the Open University Business School, UK. He has been involved in research into the behaviour and performance of financial traders since 1996 and has published widely on this topic. More recently he has focused on the impact of emotion on traders’ decision-making and approaches to improving the emotion regulation of professional traders and private investors. He has recently carried out a major collaboration with the BBC looking at private citizens’ emotional and psychological relationships with their money. He has also carried out research to develop approaches to enabling traders and investors to improve their management of emotions and gain greater control over the impact of emotions on their decisions. He has been an academic adviser on several BBC documentaries including Can Gerry Robinson Fix the NHS, For the Love of Money and the Money Programme. He acts as a consultant and adviser to a range of organisations in both the public and private sector. His blog on emotions and finance can be found at http://emotionalfinance.net.

Michal Galas
Michal Galas is Chief Programmer at the UK Centre for Financial Computing, based at University College London (UCL). Michal is an experienced technical lead, enterprise-size-systems architect and research-groups manager. He currently leads a number of commercially-oriented research projects for the UK Centre in Financial Computing. Michal specialises in R&D of large-scale financial systems including Financial Platforms, Analytical Engines, Cloud-based Simulation Environments, Algorithmic Trading Models, Connectivity Engines, Order Books, Data Handlers, Data Aggregators, Real-Time Data Processors, Service-Oriented and Event-Driven Architectures, Libraries, Frameworks, Engines and APIs.
About the contributors

Stuart Greenbaum
Stuart Greenbaum is the former Dean and Bank of America Professor of Managerial Leadership at the Olin Business School, Washington University in St Louis. He is also the Jacobs Visiting Professor at the Kellogg School of Management, Northwestern University. Before joining the Olin School in 1995, Greenbaum served for 20 years on the faculty of the Kellogg School of Management where he was the Director of the Banking Research Center and the Strunk Distinguished Professor of Financial Institutions. From 1988 to 1992, he was Kellogg’s Associate Dean for Academic Affairs. Before Northwestern, Greenbaum served as Chairman of the Economics Department at the University of Kentucky, and on the staffs of the Comptroller of the Currency and the Federal Reserve. Professor Greenbaum has on numerous occasions testified before Congressional committees, as well as other legislative bodies. Greenbaum has published two books and more than 75 articles in academic journals and other professional media. He is founding editor of the Journal of Financial Intermediation and has served on the editorial boards of 11 other academic journals.

David Ingram
David Ingram works with Willis Re’s insurance company clients to develop and improve their Enterprise Risk Management practices. He was previously in the Insurance Ratings Group of Standard and Poor’s where he led their initiative to incorporate ERM into insurance ratings. Dave has also held executive positions within insurance companies. Dave is a frequent writer and speaker on ERM. He is currently the Chair of the International Actuarial Association’s Enterprise and Financial Risk Committee.

Patricia Jackson
Patricia Jackson, based in London, is a Partner in Ernst & Young’s Financial Services Risk Management Group and leads the EMEIA practice covering the Prudential Regulatory agenda in banks including stress testing, liquidity risk, risk appetite and the different Basel Accords. She joined Ernst & Young in 2004 and was instrumental in building the banking risk practice. Before that she was a Senior Official at the Bank of England, Head of the Financial Industry and Regulation Division, and a member of the Basel Committee on Banking Supervision. She has published a range of research papers on bank capital, credit risk and market risk modelling, securities fraud and shadow banking. She is a Trustee of the CEPR, a Council Member of SUERF and an Adjunct Professor at Imperial College. She chaired an Institute for International Finance group looking at the causes of the credit crunch and the failure of the originate to distribute model.

Jonathan Macey
Jonathan Macey is the Sam Harris Professor of Corporate Law, Corporate Finance and Securities Law at Yale University, and Professor in the Yale School of Management. He is a member of the Board of Directors of the Yale Law School Center for the Study of Corporate Governance and a member of the Faculty Advisory Group of Yale’s Millstein Center for Corporate Governance and Performance as well as Chairman of the Yale University Advisory Committee on Investor Responsibility.

www.rasabourse.com
Professor Macey has published over 200 articles in leading legal and economics journals and nine books, including: *The Death of Reputation: How Integrity Has Been Destroyed on Wall Street*; *Corporate Governance: Promises Kept, Promises Broken*; *Macey on Corporation Law*; *Corporations: Including Partnerships and Limited Liability Companies*; *Banking Law and Regulation*; and ‘The Corporate Governance of Banks’ (with Maureen O’Hara).

**Roland Schatz**

Roland Schatz is Founder and CEO of Media Tenor Ltd, the Research Institute of InnoVatio Verlags AG. With over 120 employees and offices in Beirut, Boston, Hanoi, London, Melbourne, New York, Pretoria, St. Petersburg, Tianjin, Windhoek and Zürich, Media Tenor is the leading provider of ongoing international media content analysis, including in-depth analysis of new and traditional global media content.

Mr Schatz has a Masters in philosophy, economics, history and political science from the University of Fribourg and Bonn. Aside from a background in journalism and numerous entrepreneurial ventures, Mr Schatz has served as a trustee for the Education Africa Foundation in Johannesburg, the Innovation Institute in Pretoria and the Board of E-Standards in New York. In 2008, the UN High Advisor President Sampraio appointed Roland Schatz as Global Media Expert. Together with Prince Ghazi of Jordan he founded the C1WorldDialogue Foundation in 2009. In 2010, Mr Schatz launched the Global Media Impact Center in Boston, enabling PhD students to write their thesis based on the more than 100 Mio datasets of ongoing Media Analysis.

In 2013, Allianz Insurance Group together with Media Tenor launched the first Reputation Protect Product based on the Media Tenor data. Mr Schatz has also been teaching strategic communication management at universities in Augsburg, Atlanta, Berlin, Bonn, Lugano and Prague, since 1990. He publishes regularly on reputational risk, financial sentiment and media impact, recently in *Washington Post, Harvard Business Review* and *Business Day*.

**Roy C Smith**

Roy C Smith has been on the faculty of the New York University Leonard N Stern School of Business since September 1987. Prior to assuming this appointment he was a General Partner of Goldman, Sachs & Co, which he joined in 1966, specialising in international investment banking and corporate finance. During his career at Goldman Sachs he served as President of Goldman Sachs International Corp while resident in the company’s London office in the 1980s. Professor Smith received his BS degree from the US Naval Academy in 1960 and his MBA degree from Harvard University. In addition to various articles in professional journals and op-ed pieces, he is the author of *The Global Bankers, The Money Wars, Comeback: The Restoration of American Banking Power in the New World Economy, The Wealth Creators, Adam Smith and the Origins of American Enterprise*, and *Paper Fortunes – The Modern Wall Street*. He is also co-author with Ingo Walter of several books including *Street Smarts, High Finance in the Euro Zone, Global Banking and Governing the Modern Corporation*.

**Charles S Tapiero**

Charles S Tapiero is the Topfer Chair Distinguished Professor of Financial Engineering and Technology Management at the Polytechnic Institute of New York University. He is also
founder and department head of the Risk and Financial Engineering Department, and serves as the director of its Advanced Degrees Programs. Professor Tapiero has earned a worldwide reputation as a researcher and consultant, and has sat on the boards of large companies. Professor Tapiero is currently the co-editor in chief of Risk and Decision Analysis. His fields of interests span financial engineering, risk assessment and analysis, actuarial and insurance science, computational finance, infrastructure finance, networks and supply chain risk. Professor Tapiero has contributed more than 350 papers in academic refereed journals and 14 books. His research spans risk insurance and finance, operations risk and quality, supply chain risk, stochastic and dynamic systems, range processes and R/S statistics, logistics and industrial management, operations research and decisions analysis. Professor Tapiero has held numerous positions of responsibility at the highest levels of an industrial conglomerate (Koor Industries, 1994–2000), quasi-government and government agencies (1978–1982) and professorial positions in the United States, Europe, Israel and Asia. His books include Risk Finance and Assets Pricing, Applied Stochastic Models and Control in Finance and Insurance, The Management and the Control of Quality, Risk and Financial Management: Mathematical and Computational Methods and Supply Chain Risks Games. He received his doctoral degree in Operation Research and Management from New York University’s Graduate school of Business Administration, and held University positions at Columbia University, the University of Washington, Case Western Reserve University, the Hebrew University of Jerusalem, the Institute of Financial Mathematics in Montreal and ESSEC (France) before joining NYU-Poly.

**Philip Treleaven**

Philip Treleaven is Director of the UK Centre for Financial Computing and Professor of Computing at University College London. For the past eight years Professor Treleaven’s research group has developed algorithmic trading systems with many of the leading investment banks and funds, and for the past three years they have worked on HFT trading risk and systemic risk. The UK Centre is a collaboration of UCL, London School of Economics, London Business School and the major financial institutions and commercial organisations. The Centre undertakes analytics research in finance, retail, healthcare, services and sport.

The UK Centre has over 70 PhD students working on finance and business analytics, and is unique in placing them in banks, funds and companies to develop advanced analytics and software.

**Russell Walker**

Russell Walker, PhD is the Associate Director of the Zell Center for Risk Research and Clinical Associate Professor at the Kellogg School of Management of Northwestern University. Russell has developed and taught executive programs on Enterprise Risk, Operational Risk, Corporate Governance, and Global Leadership.

Russell leads the Kellogg PRMIA Complete Course in Executive Education for Risk Management. He founded and teaches the Analytical Consulting Lab, an innovative experiential learning class at the Kellogg School of Management, which brings Kellogg MBA students together with corporate sponsors seeking analytical assistance. He also teaches
About the contributors

courses in risk management, analytics, and on strategies in globalisation. He was awarded the Kellogg Impact award by Kellogg MBA students for excellence and impact in teaching Enterprise Risk Management in 2011.

His is the author of *Winning with Risk Management*, which examines the principles and practice of risk management through business case studies. He has also authored many business cases and published multiple Kellogg case studies. These cases include examinations of the deployment of analytics in media, best practices in global supply chains, optimal portfolio investment strategies, and corporate governance.

He is often quoted in the *Financial Times*, the *International Herald Tribune*, the *Washington Post* and CNN among other news media, and has been invited to share his perspective internationally through seminars. He is a frequent lecturer at the IESE Business School in Spain, the Sasin Graduate Institute of Business Administration in Thailand, and the Indian School of Business in India.

Russell began his career with Capital One Financial, Inc., where he served as a corporate strategist specialising in the advancement of analytics in the enterprise for the purposes of improved marketing and risk management. He founded and led multiple centres of excellence in analytical fields while at Capital One. His work also included international market entry evaluation and technology review.

He received his PhD from Cornell University, where he studied catastrophic risk analysis. He also holds an MS from Cornell University, an MBA from the Kellogg School of Management and a BS from the University of South Florida. Russell speaks Spanish fluently.

Russell can be reached at russell@walkerbernardo.com and www.russellwalkerphd.com.

**Stefan Walter**

Stefan Walter is a Principal in Ernst & Young’s Global Banking and Capital Markets Center and is based in New York. He has more than 20 years of experience in global bank supervision and regulation. As Ernst & Young’s Global Bank Supervisory and Regulatory Policy Leader, he works with clients to help them understand the evolving financial and regulatory landscape and the implications for companies’ governance, risk management and internal controls.

Prior to joining Ernst & Young, Stefan was Secretary General of the Basel Committee on Banking Supervision (BCBS). During his five year tenure at the BCBS, Stefan oversaw the most fundamental transformation of global bank regulation in decades, including the Basel III capital and liquidity reforms. He chaired the BCBS’ Policy Development Group, which is the key policy-making body of the Basel Committee. Stefan also was a member of the Financial Stability Board and the Group of Governors and Heads of Supervision.

Before joining the Basel Committee, Stefan was Senior Vice President at the Federal Reserve Bank of New York (FRBNY) and led domestic and international policy, as well as the Bank-wide financial stability efforts. During his tenure at the FRBNY, he ran the Market, Liquidity and Counterparty Credit Risk Department and co-ordinated the development of international guidance in the areas of counterparty credit risk, operational risk, market risk and economic capital planning. Stefan also managed the FRBNY’s off-site Bank Surveillance Department.
Stefan is a regular speaker in the US and abroad on topics related to global regulation, supervision, risk management and financial reform. He holds an MA degree in International Banking and Finance from Columbia University and a BA degree from the University of California, Berkeley.

Paul Willman
Paul Willman is Professor of Management at the London School of Economics and Director of the Private Wealth Institute at Oxford University. He has held Chairs at London Business School and Oxford. Paul has published widely on trading and risk in financial markets and he consults to a range of financial institutions.

Brendon Young
Brendon Young is recognised internationally as a leading expert in the field of risk management within the financial services sector. Professor Young is chairman of the Risk Research Foundation (ORRF); founding president of the Institute of Operational Risk; and currently, Non-Executive Director of George Eliot Hospital NHS Trust where he is actively engaged in its performance improvement turnaround. He has been an adviser to prominent financial institutions including Moody’s and the Financial Reporting Council’s Board of Actuarial Standards. He has published papers and lectured widely, giving presentations at the FSA, the Bank of England, BaFin, the Dutch National Bank, the OCC, and the New York State Banking Department. Previously, he was director of Arthur Andersen’s risk research centre. In academia he was business school associate dean, at Birmingham City University, responsible for risk research and business development. His early career was in consultancy with Spicer & Pegler Deloitte and later in venture capital with Lazard-WMEB. Initially, he trained in industry with Rolls-Royce aero-engines and Jaguar Cars, qualifying both as a chartered engineer and a chartered management accountant. His work has received international recognition for research excellence. Professor Young can be contacted via LinkedIn.
Chapter 1

Managing the risks in financial institutions

Shahin Shojai
Applied Thinking FZE

George Feiger
Aston Business School

Overview
This chapter introduces the reader to each of the topics covered in subsequent chapters of this book. It also explains what might, at first blush, seem like a strange omission on our part, namely the absence of exposition of quantitative modelling of market and credit risks. We have left this out not because we believe it to be unimportant but rather because we are convinced that today’s approaches are as yet unsuited for practical applications in real financial businesses. We firmly believe that qualitative models should play a greater role in risk assessment and risk management strategies. We were fortunate to have a number of the most respected experts in these fields join us and contribute their insights. We believe that the main contributions of this book are first, its focus on risks that are truly manageable and second, that we provide such analyses for both banks and insurers. There is a lot that these two types of institutions can learn from each other within the risk arena and we hope that this book encourages this.

Introduction
The financial services industry has been at the receiving end of a huge amount of criticism in recent years. Politicians and those who elect them have placed the bulk of the blame for the Great Recession squarely at the doors of the major global banks and brokers. As a consequence, both to punish the guilty and to ward off future crimes, a myriad of new regulations has been introduced across the globe, regulatory bodies have been recreated or reshaped, and steps are being taken to ring-fence, or even break up large banking conglomerates. In addition, despite their evidently limited value in preventing the crisis, many more people with degrees in quantitative and mathematical finance have been hired to develop methodologies to predict future crises and to improve the process of risk management within financial institutions.

We have assembled this collection of essays in the effort to provide an objective overview, from internationally recognised experts, on how the major risks within banks and insurers are measured and managed and how these processes might be improved.
What you will not find is a lot of quantitative modelling of the sort favoured in today’s finance PhD programs. Our focus is on management issues and processes. Our contributors examine how operational risk can be properly measured and managed, especially within a world in which high-frequency trading means decisions can, and need to, be made within milliseconds. They look at the damage to reputation that the financial services industry has endured and discuss what can be done to regain the clients’ trust. There is a discussion of how enterprise-wide risk management methodologies have been in many cases misunderstood and misapplied, and of how a good enterprise risk management (ERM) mentality might help in the future. Of course, no book on the major risks within financial services would be complete without an investigation into the implications of the new regulatory landscape. The book concludes with a peek into the future of risk, both financial and regulatory, and what financial institutions need to take into account in order to not only survive but possibly thrive in the future.

Before we get to a fuller explanation of the contributions of each chapter, we would like to explain our justification for ignoring quantitative modelling of credit and market risks. We are not against research-based modelling, we just do not think that its output is yet ready for practical application except in a modest way, as an aide to informed judgment. Today’s quant models are outgrowths of so-called modern portfolio theory (MPT). This theory does not work to explain asset values and the risk management tools developed from it do not, as a result, manage risk. Much of the explanation that follows is based on two of our past papers.1,2

**MPT models do not work in practice**

MPT is a theory taught in business schools all over the world and used by risk managers in financial institutions and by the regulators that attempt to police the risks that these institutions take. It has evolved from the pricing of assets directly to the pricing of derivatives on those assets, such as options. The predictive statement in MPT is this: if you know the statistical distribution of future returns of assets, you can work back to the prices that a rational investor should place on these assets today. There have been thousands of attempts to apply these models to actual markets. The conclusion of this research is unambiguous. No version of MPT explains more than a small portion of the observed behaviour of the prices of assets or their derivatives.

Indeed, the theory’s failure is so extreme that an objective observer of the quasi-mathematical and a priori theorising that is used in these models is inevitably drawn to the old joke about the man who looks for his dropped keys under the street light because that is where the light is rather than where he dropped the keys. Financial economists have stood conventional scientific methodology, which develops theories to explain facts and tests them by their ability to predict, on its head.

**Because of irrational investors?**

A popular attempt to explain why the theories do not work hits on the concept of the ‘rational investor’. The school of behavioural economics argues that investors (and other
people too) do not make rational, calculated economic decisions. Behavioural economists
have shown, definitively in our opinion, that decision making by virtually everyone, even
economists, is nowhere near as rational as MPT or any other economic theory supposes.
However, before we rush out to teach Economics 101 to the population at large, we should
consider whether this is the main issue. In our view, understanding of market dynamics is
still so limited that the key problem is not that market participants do not make rational
decisions; it is that they cannot.

Every variant of quantitative models used in risk management needs to assume that out
there, there is a probability distribution of future asset values which can be used to inform
risk management tools. But as we have recently been reminded, we operate in a global web
of financial connections, structures and information flows that we do not even know in detail,
let alone understand in dynamic behaviour. Gorton\(^3\) captured this very well. In financial
markets, we have a modelling problem that is more similar to weather forecasting or earth-
quake prediction than predicting where the moon will be in its orbit at 4 pm tomorrow.
This is not all bad. Look how our ability to predict weather has improved as our ability to
model large, complex systems has evolved. We can now predict weather pretty well several
days ahead. But weather scientists achieved this through decades of incremental effort. We
need to learn more about how beliefs propagate, how incentives affect financial behaviour,
how political scenarios create economic scenarios, the global linkages between bank solvency
and the derivative markets and much more.

Let us look at some concrete examples of these issues, not captured in the Gaussian
Copula of the distribution of future asset prices.

**Securitisation, mortgages and CDOs**

Collateralised debt obligations (CDOs), and especially CDOs of securitised mortgages, have
been seen as the heart of the financial crisis. The rating agencies stand accused of wilfully
misestimating the risks in these structures and so lulling investors into buying them. Without
entering into issues of intent, we believe that it was rather difficult to provide any effective
valuation or rating of these instruments. The rating agency analysts used historical default
data to make their assessments. Consider first what was in the structures and then how they
were likely to behave.

Securitisation fundamentally changed the risk structure of the mortgage pool, rendering
historical quality/default data much less useful than one might have imagined. Prior to the
development of individual risk rating models,\(^4,5\) banks used to base their lending decisions
on the information they had gathered about their borrowing clients over a number of years
of a ‘banking relationship’. When banks began treating a mortgage as a stand-alone product
rather than as part of a relationship, they had to resort to using the credit report information
provided by personal risk/credit ranking/rating organisations. This information was already
a step-down in value\(^6\) but, as long as the loans stayed on the bank’s balance sheet it had
strong incentive to monitor the origination activities of its employees or third-party brokers.
Securitisation, the packaging of mortgages for sale to third-party investors, transferred out
the loss risk. If the sold packages were credit-enhanced, the investors did not even know
or care who the originator was. The banks were transformed from risk-bearing lenders
to commission-receiving sales machines. Non-banks, like independent mortgage origination companies in the US, perfected the mortgage production line based on volume rather than credit quality. The Wall Street packagers and underwriters thought of mortgages as interchangeable product ingredients and, at the end of the line, so did their quantitative risk analysts recently minted from physics departments who had never met a sub-prime borrower. Neither they nor the credit enhancer nor the rating agency had any idea about the underlying quality of the borrowers. Consequently, all that was needed to bring the house of cards down was a correction in house prices, which is exactly what happened.

From a modelling perspective, what was needed was a theory of how changing incentives for originators and the degree of actual knowledge transfer about borrowers along the securitisation value chain would, in quantitative terms (!) affect likelihood of default, by characteristic of borrower. We, the economics/social science profession, do not know how to do this yet.

**Structured products, liquidity and systemic risk**

Structured products were houses of cards but let us look at the structure of the houses. Specifically, consider the example of the collateralised loan obligation (CLO) market in 2008. Several hundred billion of corporate loans were held in CLOs, on the largely invisible books of off-shore hedge funds and underwriting institutions (via off-balance-sheet vehicles), in structures leveraged 20 times and more. The leverage came from major banks and insurance companies which, we should remember, devoted the bulk of their other business activities to making loans to entities like businesses and real estate developers and to each other. They in turn raised their liabilities by issuing bonds and commercial paper.

Some of the loans in the CLOs started falling in price in secondary market trading (the market making in which, by the way, was provided by the same banks which were providing the leverage to the CLO holders). This precipitated margin calls on the holders that they could not all meet. With leverage of 20 times, the fund equity could quickly disappear so the only recourse was to dump loans and deleverage as quickly as possible. Blocks of hundreds of millions or billions of loans were thrown out for whatever bid they could get. Seeing this, would you buy, even if you thought that almost all of the loans were ultimately money-good? Likely not, because in the panic it was more likely than not that prices would fall further, which they did. Indeed, at some points apparently-money-good loans were selling at around 50 cents on the dollar.

Normally the banks as market makers would buy such bargains but they were, simultaneously, in their role as prime brokers, providing the leverage in these trading funds and they were holding their own trading inventories that were tumbling in price. So they withdrew leverage from the funds and reduced their own inventories and so forced prices down further. This spiral created substantial losses on their own balance sheets so that the prime broker banks were in trouble themselves. Fear of a defaulting counterparty dried up the interbank lending market, essential for liquidity in the world trading system, and their commercial paper appeared risky and fell in price, damaging the money market fund industry which held a large part of liquid assets in the US. Similar things happened with mortgage-backed CDOs and other structured instruments.
We need not elaborate on the history but you can see why the variance/covariance matrix in the nice quantitative risk management models did not work out too well. These are fascinating system dynamics for a social science research agenda but they are not tools for actual day-to-day management of anyone’s risk taking business. Or for their regulators.

**Capital market incentives**

Financial market participants are actual businesses. All the significant ones are public companies subject to market scrutiny of quarterly performance. Moreover, for good or ill most shareholders are no Warren Buffett with patience to wait out economic cycles. Most of the readers of this book will be aware of the infamous statement of the Chief Executive of a major bank that it was impossible for his company to stop originating mortgage securities: ‘as long as the music is playing we have to dance’. This statement should not be interpreted as a criticism of this gentleman. Suppose, for example, he had ordered his company to get out of the mortgage business in, say, 2006, when the house price explosion was very evident as was the deterioration in the quality of underwriting. The bottom did not fall out for another two years. Who doubts that, well before the end, he would probably have lost a lot of his key staff and would, perhaps, have been forced out as CEO by investors focused entirely on the growth of quarterly earnings?

**Value-at-Risk**

We do not need to elaborate further but we cannot resist. Value-at-Risk (VaR) is the most reported aggregate summary of the ‘risk taking’ of large financial institutions. It is not hard to understand in concept. A financial institution is a ‘portfolio of gambles’. Its management can choose how many gambles to take and how much to bet. Assume that the probability distribution of, say daily, outcomes of the portfolio of gambles is Normally distributed and known. Obviously it would vary with the choices made by the management. Assume that the management can assess the impact on the distribution of each decision that they take. Then, the 5% VaR is the upper bound of daily loss that happens 95% of the time. If the 5% VaR is $100 million, then 95% of the time the daily loss will be less than $100 million. Or, put another way, a loss of this magnitude or greater ought to happen only one trading day in 20. If the management do not like this then they can change the portfolio and lower it to $50 million or $5 million.

Sadly for us all, the distribution of outcomes is not Normally distributed nor is it known nor do managements know the impact of their decisions on the distribution nor is the distribution independent of the actions of other managements in other institutions, of the actions of the central banks of major economies, and of the mood of the investing public.

So why use it all? Are we back to the joke about the lost keys and the street light? Not entirely. ‘Most of the time’ tools like this work satisfactorily. Only sometimes, in times of crisis, they do not work at all. This is the sense in which quantitative models are aides to judgment but not replacements for it. The management needs to decide the tools for the times. Today there is no quantitative model that can substitute for analysis of the fundamentals and making a call that, say, the housing market is grossly over-valued or the universe of tech
companies is grossly over-valued. Then, and this is harder, you need to have the patience to wait. This waiting may indeed mean staying out of the business altogether. Many shrewd investors who shorted the on-line funeral business or mortgage securities ran out of money before they were vindicated. As Keynes said, the market can remain irrational longer than you can remain solvent. These are huge communication challenges in the short-term-oriented world in which we find ourselves.

Let us now turn to other issues that risk managers must grapple with.

**Organisational structure**

Even the most simplistic modelling process requires as its foundation the ability to aggregate risks being taken within an enterprise. As a practical matter, this is a difficult task. Many businesses are organised into ‘silos’. Financial institutions are notorious for this.

Different operating divisions have different management styles, compensation structures and risk appetites. In order to determine the company’s overall risk appetite we need to be able to firstly place each of the risks within their correct buckets, such as credit, market, operational and so on. Most institutions have major difficulties in clearly reporting the risks of their credit and market instruments, let alone their operational or liquidity risks.

Most institutions have only recently started to come to grips with the fact that operational risk is a major risk and that it must be managed. Learning how to quantify it will take years, if not decades. Once that is done, we need to be able to compute the operational risks of each division. Many institutions still permit different businesses to operate with different accounting structures and concepts, hence it is literally impossible to quantify operational risk for the group and determine what diversification benefits could be derived. For example, many institutions combine credit and FX instruments within the same divisions, others keep them separate. Some combine complex instruments with the fixed income division, others with equities. In some companies FX, credit and equities each have their own quants teams, whose risk concepts differ in poorly understood ways. As a result, companies often have a view on the risk of each silo, but will not be able to aggregate them in any useful way for the group. Similarly, since they are sitting in different parts of the business it is difficult to accurately determine the correlation benefits that the company might experience.

**Too much data, too little understanding**

Even if there is effective structural integration in concept, there is just too much data to deal with. Risk management teams face a hosepipe of data which they have to decipher, and its contents change with every trade and even by hour of the day.

Even if so much data could be effectively analysed the next tough task is to present them in a useful way to the management, since it is they who determine the company’s overall risk appetite and not the risk management division. Once the management sets the parameters then it is the risk management team’s job to ensure that everyone operates within the set boundaries.
These kinds of complex reports are very hard for the management to understand. And the simpler they are made the greater sacrifices have to be made about specificities of risk. The result is that either the company does not allocate adequate capital to the risks it faces or it allocates too much, hence limiting some of the company’s profit potential.

**The intersection of IT and moral hazard**

While academic literature does account for personal greed (politely referred to as the ‘agency problem’) it is usually overlooked when it comes to subjects that deal with anything other than corporate finance. For some reason, there is an assumption that just as investors all behave rationally, employees are also mostly truthful about their work. A risk management system is no more reliable than the data it receives, and financial executives are inclined to reduce the reported allocation of risk to their activities insofar as possible so that they can take on greater risks. They have an option on the bank or insurance company in that they share in the gains but not the losses.

Given the complexity of many of the instruments that financial institutions are dealing with it is only natural that they will be collecting underestimated risk profiles from the traders and engineers. Furthermore, because of the silo structures many companies will find that they have assets that are highly correlated with each other but sitting in different silos. As a result, during a downturn the company might find itself significantly more exposed than it thought and not as capitalised as necessary.

The desire to underestimate risk is not limited to the traders or engineers sitting on the quants desks. To the extent that the company’s own management are compensated on their annual performance, if not quarterly, they would also like to be able to take more risks than they should be permitted to. And, none of these people are stupid. The traders know that they will at worst lose their jobs, but if their bets pay off they will be able to retire with savings that even their grandchildren cannot spend. And, the management know that if the bank goes close to failure it will be bailed out by the government if it is important enough to the local economy or the globe. The moral hazard problem, as some had predicted, has become severely worse since banks are now even more confident of taking big risks. After all yesterday’s large banks and insurance companies are now even bigger, with more certainty that they will be bailed out in the future.

**Yesterday’s spaghetti**

In addition to the human aspects of risk management, most companies are unable to get a holistic view of their risk due to the fact that their IT systems are simply unable to provide that kind of data. Most financial institutions are unable to determine the actual cost of the instruments they develop let alone their risk, and how it might change over time.

In today’s world very few financial institutions have not undergone some sort of a merger or acquisition, with the pace increasing in the past couple of years. The result is a spaghetti-like infrastructure of systems that are simply unable to communicate with one another. Incompatible systems make the development of a viable enterprise-wide risk management
close to impossible, since without accurate and timely data the company will be ill prepared to respond to changes in asset compositions and prices.

And, sadly that situation will remain for many years to come. Despite the huge annual investments made by major financial institutions different systems supporting different instruments are simply unable to share information in any meaningful way. This means that no matter how remarkable the models used are, the data that they will be dealing with are incomplete.

Consequently, despite the best of intentions and ambitions, the dream of having a reliable and effective enterprise-wide risk management shall remain just that, a dream. And as long as companies remain incapable of determining their own exposures, irrespective of which of the reasons mentioned above is the main culprit, they will continue to face enormous risks at times when markets do correct themselves rapidly and sadly the precedent set in the recent crisis does nothing but put flame to the fire. Financial institutions are now even more confident of taking risks than they were two years ago, which means that the next crisis can only be greater than the one we just lived through.

**So, what should be done?**

The chapters in this book provide useful guidance.

We hope that we have provided a convincing argument against over-relying on so-called scientific models for evaluating and managing credit and market risk, in fact all types of risks. Given that we cannot, or should not, rely on these models, what can banks and insurance companies do to be safer in the future?

Well, that is the question we aim to address in this book. Our objective, and by that we mean all the contributors to this book, is to move away from the heavy reliance on the quantitative models of risk measurement and management and to focus instead on what can be measured and managed from a qualitative perspective.

The chapters that you will read within this book all aim to help us understand, or maybe even remember, what qualitative tools the management have available to them and how they should apply them within their enterprises. In so far as possible we have tried to provide the analysis for both banks and insurance companies, since while they have many similarities they are very different businesses after all and we believe that one of the important contributions of this book is that it allows the readers to find out about the major risks that both types of enterprises face and how they each go about managing them.

It is also important that we allow senior executives from within both insurance companies and banks to see how their peers within the other type of organisation are managing different types of risks so that perhaps they can apply them within their own businesses too. This is especially important since the lines between many of the products and services that are offered by insurance companies and banks are blurring. And, as these institutions continue to grapple with falling margins they are much more likely to start, or even continue, developing products that would have historically belonged to the other type of business.

In Chapters 2 and 3, Russell Walker of the Kellogg School of Management of Northwestern University and Brendon Young of the Institute of Operational Risk provide insightful analyses of the types of operational risk that insurance and banking institutions,
Managing the risks in financial institutions

respectively, face and provide prescriptive solutions of how they can best be managed. Operational risk has grown in importance in recent years as more and more financial services institutions have learned that managing market and credit risk is not enough to ensure that their businesses are compliant with the myriad of regulations and that they are avoiding the kinds of pitfalls that has caused many of their peers to face problems and fail. Of course, what is important is to recognise that simply meeting the operational risk guidelines set by regulators is not enough to ensure that these institutions are safe. Sadly, the types of operational risks that both types of institutions face are many, and regulators have typically focused on only a few of them.

What is for certain is that most institutions are just coming to grips with what operational risk really is, and moving away from what they used to do, which was to bucket all risks that are not credit or market into operational risk. Over time, having a good understanding of recognising what operational risks an institution faces and learning how to manage them becomes essential. We are very fortunate to have two such world renowned experts provide us with such holistic perspectives of operational risks within banks and insurance companies so that our readers from the banking and insurance institutions can in one place see whether what their peers are doing in other banks and insurers in order to identify and manage operational risk.

An example of operational risk that is growing in importance is the growing proliferation of high-frequency trading (HFT). Financial institutions are now operating in an environment where they need to service some clients – and their own proprietary trading desks, for those that still have them – that undertake multiple trading activities within a few milliseconds. It does not take a good understanding of the subject to realise just how risky such transactions can be, and how large the damage should a problem occur with the models used by the institutions that undertake these types of trades. HFT is a new area to finance, facilitated by the growing power of computer processing powers and models, and we felt it is essential that we look at what they are and the types of risks that they introduce into the financial system. After all, a flash crash, which can be caused by a faulty trading model used in HFT, impacts all of the participants in the market, and a number of examples are mentioned in the book. In specific, flash crashes impact investors and those that provide brokerage services to their clients.

In order to get a better understanding of HFT and the risks that they generate we felt that we needed to dedicate two chapters to the topic.

The first was to find out what HFT is, the kinds of participants that take part in HFT, including the infrastructure providers, and how financial institutions try to avoid the types of mistakes that could cause flash crashes. In Chapter 4, Carol Clark of the Federal Reserve Bank of Chicago discusses the results of a number of studies that she and her colleagues had conducted in order to find out about the controls that are in place to manage the risks that are associated with such activities. Carol Clark looks at what constitutes HFT and provides a review of a number of well-known flash crashes and their implications.

In addition to finding out about what HFT is and the kinds of risks it introduces into the system, we also wanted to get a better understanding of the types of technology used by HFT traders and what types of glitches, and where, could cause problems. This analysis has been eloquently provided in Chapter 5 by Philip Treleaven and Michal Galas of the
University College London. The authors look at the types of technologies used by HFT traders, the architecture of IT systems used and the kinds of risk management infrastructures that have been in place to avoid the kinds of flash crashes that caused a 1,000 point drop in the Dow Industrial Average, and its reversal, within 20 minutes.

HFT is a new area in financial services, and many are just coming to grips with what it is and how it will impact our industry. We felt that it is imperative that an in-depth overview of the subject is provided in this book.

Of course, one of the greatest operational risks that financial institutions face is that of individuals who undertake transactions that they should not and find ways to hide them. In most situations the losses are recovered to some extent and no one hears about them. But, in certain situations they can cause huge damages to the institution concerned and even the larger markets. In Chapter 6, Paul Willman of the London School of Economics and Political Science and Mark Fenton-O’Creavey of the Open University Business School, look at the impact of culture on risk, and find that ‘management of risk taking behaviour in financial organisations is a complex managerial task, not something subject to a quick regulatory fix’. In other words, when a financial enterprise faces a situation of rogue trading it is usually the result of internal actions within the company rather than their inability to meet regulatory guidelines. This is a fascinating chapter in that it gives us a window into how personal actions and group dynamics within these organisations can impact risk behaviour and risk taking.

Chapter 7 looks at one of the most important operational risks that financial institutions face, namely reputational risk. Most readers are very familiar with the damage that the recent financial crisis has inflicted on major global financial services companies. Most are also fully aware that financial institutions do not spend as much time and effort as their peers in other industries in managing their reputations. Roland Schatz of Media Tenor International looks at how the media has viewed and represented financial services companies during the recent crisis and just how bad the situation has got. Most readers would be quite surprised to see just how badly the reputations of financial services companies have been damaged in recent years. Roland Schatz then goes on to suggest some ways in which financial services companies can better manage their reputations and the steps they need to take in order to be better prepared in the future.

Chapters 8, 9 and 10 look at what regulatory risk is and how financial services companies can best manage risk from an enterprise-wide perspective.

In Chapter 8, Jonathan Macey of Yale Law School describes what regulatory risk is and suggests that the regulatory environment in the US in specific, ensures that the institutions that are most impacted by new regulations are the ones that have the greatest say in how they are structured. Instead of looking at the types of risks that financial institutions face, which are covered in Chapter 12, he explains why some regulations in themselves create the environment for future crises. For example, he cites the introduction of new regulations in the US by the SEC that increased the power of rating agencies and the regulators’ fascination with VaR models and their push for them to be used by financial institutions. In both cases, financial institutions became more engulfed in implementing and applying them than critically determining whether either was any good at measuring and managing risk. The rating agencies’ ability to predict failure at the levels of granularity they suggest is highly unlikely and VaR models are highly questionable in their efficacy.
Chapters 9 and 10 look at how insurance and banking institutions, respectively, should look at risk from an enterprise perspective and how ERM tools and methodologies can best be applied by these institutions. In Chapter 9, David Ingram of Willis Re presents the seven key principles of enterprise risk management and discusses how insurers apply and misapply them. In Chapter 10, Stuart Greenbaum of the Olin Business School of Washington University in St Louis starts off by firstly deconstructing concepts of risk, risk management and ERM, and then looks at risks taken to make profit and those that need to be minimised. Having explained ERM, he goes on to explain the processes and best practices for instituting ERM within financial institutions.

Chapters 11, 12 and 13 try to look into the future to see what types of risk financial institutions could face going forward. In Chapter 11, Charles Tapiero of NYU-Polytechnic Institute discusses the kinds of risks financial institutions face today and how pricing of risk might change in the future. He takes a more philosophical perspective of the major risks financial institutions will face, and these include the impact of new technologies, regulations and globalisations. In Chapter 12, Patricia Jackson and Stefan Walter of Ernst & Young take a peek into the future of financial regulation. Both being veterans of financial regulation, they try to decipher the messages coming from global regulatory bodies to discuss what the future regulatory landscape might look like for financial service industries.

Finally, in Chapter 13, Roy Smith of the NYU Stern School of Business looks at the recent history of global banking institutions and tries to decipher how they might look in the future. Roy Smith highlights the growing concentration of power within the hands of a few giant banking institutions, yet explains that the growing pressures on these businesses might result in business models that look very different to today’s and more similar to the past. He cites the regulatory and competitive pressures that global banking giants currently face and suggests that unlike what many believe, some of these giants might be forced to exit some businesses that are viewed as too important today.

Conclusion

It is our objective with this book to provide an in-depth analysis of the major risks that financial institutions face; and to do so for both banks and insurance companies. We look at the current environment and try to see how the world of financial services might look like in the future, given the risks that the institutions operating within it are facing, and will face.

In order to achieve this objective, we brought together truly world-renowned experts in each field to share with you their views on the specific risks that they cover, ranging from operational and regulatory to reputational and enterprise-wide.

We are grateful to the contributors of this book. They have helped create a unique analysis of the risks that banking and insurance companies face and provided useful advice on how they can best be managed, in the hope that the executives within these companies can apply the best thinking from across our ecosystem within their enterprises.

Our hope is that you, the readers of this book, have as much enjoyment in reading these chapters as we had in bringing them together to create a coherent message for the financial executives and students who will buy this book.
Bibliography


Reinhart, CM and Rogoff, K, This time is different: eight centuries of financial folly, 2009, Princeton University Press.


The views expressed in this chapter reflect only those of the authors and are in no way representative of the views of any organisation that they have been, or are, associated with.

8 Shojai, S, ‘Why the rush to repay TARP?’, Riskcenter, 12 May 2009.
Part 1

Operations and IT risk
Chapter 2

Operational risk in the insurance industry

Russell Walker
Kellogg School of Management, Northwestern University

Overview
Operational risk has increased in importance in the insurance industry over recent years. The actuarial process for product formulation, the long duration of many insurance products, and the need to match assets and liabilities in time mean that deviations in product performance, stemming from operational risk, have large impacts on the insurers’ profits. Operational risk is embedded in the functions of an insurer. Detecting operational risk is challenging, as it is often hidden in product performance as unexpected losses and experienced years after products are sold. Research findings show that operational risk in the insurance industry is increasing in size over time, occurring most frequently and severely in the handling of customer interactions. The product complexity and customer interaction that is inherent in the insurance business leads to an operational risk profile that is different to other financial institutions, and one that appears to be more expensive in terms of cost of capital. Operational risk is, at its core, an error in a decision or process that impacts the policyholder, and can, through contagion, result in regulatory fines, poor product performance and persistent reputational harm. Managing operational risk in the insurance industry ultimately requires promoting a culture that is data-driven and a management philosophy that invests in understanding complexity and developing process excellence. Setting a standard for preventing errors and seeking explanations for errors is critical to managing operational risk and preventing the spread of further, more damaging risks to the insurer.

Origin of operational risk
The origins and notions behind the word ‘risk’ have a heritage in the perils and uncertainty of sea commerce, emerging several centuries ago. Risk first appeared in the English language as part of insurance and shipping terms: shippers and receivers had to bear the ‘risk’ of lost cargo in transit. Previous to that, the concept of risk and the word risk are not seen in any language. Risk thus developed as a concept in commerce that could account for acts of God, force majeure, and the general danger and peril related to the loss of goods in transit. This is perhaps best preserved in the current Spanish phrase, ‘por su cuenta y riesgo’, which is literally ‘by one’s cost and risk’, or more familiarly, ‘at your own risk’. The aversion of risk by merchants gave rise to a highly profitable industry: shipping insurance. Lloyds of
London was one of the first insurers to gain industry prominence, a mighty accomplishment for a 17th century coffee house.

These early notions of risk focused on loss that could not be separated from the act of conducting some commerce. So risk owes it origin to, and is embedded in, the operations of commerce. Frank Knight\(^2\) pointed out that risk is an economic concept at work in business decisions, owing to the inherent existence of known and unknown factors. Buyers and traders of goods were unwilling to bear the risk of the unknown factors, which gave rise to a specific part of the insurance industry. So, it seems appropriate that we examine operational risk and its impact on the insurance industry.

**Operational risk – notional definitions**

Operational risk is driven by the company’s operations, not the company’s investments. That said, operational risk comes with all investments, but for a company, the risk itself is embedded in the function of making investments and running the company. Operational risk is also embedded in other activities undertaken by the enterprise, but company leaders do not explicitly take on operational risk as they would market risk or credit risk.

<table>
<thead>
<tr>
<th><strong>Risk definitions for common risks</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regulatory risk</strong></td>
</tr>
<tr>
<td><strong>Reputational risk</strong></td>
</tr>
<tr>
<td><strong>Operational risk</strong></td>
</tr>
<tr>
<td><strong>Credit risk</strong></td>
</tr>
<tr>
<td><strong>Market risk</strong></td>
</tr>
</tbody>
</table>


In the banking and financial services industry, the Basel Accords define operational risk as that risk coming from ‘inadequate or failed internal processes, people and systems, or from external events’. This definition is broad and does not focus on the causes of operational risk.\(^3\) To be clear, operational risk in this industry is the risk that comes from operating a financial services company, not the risks associated with credit or market risks. Credit risk occurs when a loan results in less than full payment, and market risk exists when asset values fall. These two risks are paramount to the financial services
industry, thus making it difficult to separate such risks from the business of financial services. Regulatory risk and reputational risk, byproducts of operational risk, are similarly important risk types that matter a great deal in financial services and will be discussed in detail later in the chapter.

**Market risk – drawing a distinction from operational risk**

The purchase of any asset involves an active investment. This essentially involves the transfer of cash into another market instrument or asset (stock, gold, security, and so on). The asset is measured over time and carries with it market risk, or the risk that its value may fall.

Market risk has a direct upside and an opportunity to experience a gain. The downside of market risk losses is that it is rarely fully-experienced, except, of course, when an asset is completely worthless.

Notice that market risk involves investment of a fixed amount, whereby the loss is more or less confined to the investment amount. The volatility in the price of the asset enables a marketplace of buyers and sellers with different risk goals, capital needs and outlooks on the future price of the asset. In most cases, players in such markets can leverage rich data to create robust models that accurately quantify risk. Market risk, however, can be most cumbersome when dealing with illiquid or unique assets, like real estate or collectibles.

In contrast to market risk, operational risk gives rise to losses that are definite and involve no direct upside. Additionally, for the most severe forms of operational risk, there is no counterparty with whom to trade the risk, requiring management to continually focus on addressing operational risk and self-insure the ensuing loss.

**Credit risk – drawing a distinction from operational risk**

For banks, credit risk also involves an investment. It involves transferring cash into a future receivable. The selection of the loan and its terms are therefore paramount to the banking business. The amount that can be lost on a loan (other than opportunity cost) is capped by the amount lent. Credit risk management relies on a holistic consideration of the borrower’s current and past financial situation, as well as his/her future economic situation. There are robust mathematical models for credit risk modelling, although assumptions in those models can often prove problematic.

Credit risk, like market risk, involves the potential for upside and gain. Lenders can renegotiate terms in order to reduce losses. Terms might even be modified for the borrower’s benefit to reduce risk. And even if a borrower goes bankrupt, the creditor can expect some financial recovery (which is often material and high). Historical data, credit surveillance information from rating agencies and economic models can guide a bank in the selection of credit risk.

In contrast to credit risk, operational risk leaves no opportunity to renegotiate terms and little opportunity to reduce losses once the damages are realised. Surveillance of operational risk is also challenging – managing operational risk cannot be derived from the financial performance of a counterparty. While historical operational risk data may prove useful in modelling levels of loss, it does not prove an explanation for the operational risk.
Operational risk – in detail

From the Basel Accords, operational risk includes those risks embedded in processes, systems, people decisions and hazards. It might be finite in exposure, but as we will see from examples, operational risk often is persistent and can cause contagion. Contagion can subsequently lead to liquidity risk, lawsuits, regulatory scrutiny and even reputational harm.

In many ways, excellence in managing operational risk requires revealing the risks embedded in business decisions, as operational risk is often an ancillary risk not specifically desired, but implicitly accepted as part of a business function. As mentioned before, unlike credit and market risk, operational risk is not measured by looking at a borrower or a market, or by relying on economic assumptions. In most companies, the measurement of the means remains less important than the measurement of the ends. This, therefore, relegates operational risk to a poorly understood form of risk, even though it is largely embedded in every business decision and possesses a liability that can impact the prudential of the company.

Operational risk is hard to measure

Operational risk, as measured and calculated by financial services companies under Basel regulations, is most certainly the least evolved of our risk methodologies. In general, operational risk captures the ‘mistakes, errors and hazards’ inherent in doing business. But how does one measure something that is not supposed to happen and perhaps has never happened before? The risk methodology used by financial services companies looks at data on losses from previous loss events and through some statistical scaling and distributional assignment, providing guidance on operational risk levels that are typical in the industry. Although many risk managers question the validity and trustworthiness of credit risk and market risk models in the aftermath of the Great Recession, the reality is that our most advanced tools for risk measurement are in credit and market risk models. Beyond that, the tools and data for operational risk management are weaker, owing generally to the poor understanding of risks implicit in business decisions. Operational risks have gained a great deal of attention in the recent decades (and rightfully so), and great advances have been made to measure and model operational risk. Financial service companies employ Basel-endorsed techniques that ultimately map operational risks to techniques (like Value-at-Risk) used in market and credit risk management. However, the explanation for the operational risk might not be considered thoroughly, leaving the path to the error undefined.

Operational risk is driven by complexity

Operational risk is inherent in the company’s selection of a client, country, technology, or system. Operational risk increases as systems and processes become more complex and less understood. Modelling the complexity of a system requires an understanding of how it fails. Since operational risk is often the result of an unpredictable or a not-fully understood failure mechanism, predicting operational risk remains a challenge.

Managing operational risk requires an almost engineering-like understanding of systems, processes and failure mechanisms. Sadly, most businesses do not invest enough to understand
processes and systems in such a granular level, especially when the systems and processes are secondary to the business models or goals. As such we see weak operational risk data and generally nascent models for considering operational risk in most companies.

The goal of the business manager is, therefore, to recognise the type of risk facing the organisation and to make the investment to understand it further. Doing so will go a long way to reducing persistence and removing implicit operational risk as much as possible.

**Operational risk is not like other risks**

Operational risk does not involve a direct capital deployment with the opportunity of an investment gain. Losses from operational risk are not due to bad loans or to holding volatile assets. The risk occurs because something undesirable results in an expense or other loss to the company. Furthermore, there is no market or counterparty in operational risk, the loss cannot be renegotiated, and payment cannot be prolonged. In fact, operational risks are often detected long after issues emerge, and such issues are frequently intertwined and correlated, meaning that one form of operational risk increases the exposure to more risk.

Operational risk is at its core, a mistake, error, or hazard. Generally, company hazards, like damages from wind, rain, fire and so on, can be insurable with regards to physical property. However, it is nearly impossible to fully insure against the missed business from a lost or damaged piece of property. Thus, the presence of hazard does materially impact business beyond insurance.

Other forms of operational risk are embedded in how the enterprise functions, and are often driven by people and IT systems that do produce errors. Operational risk is not strictly the product of an internal decision process. But rather it is manifested through the complex web of employees, products, clients, systems, legal judgments, regulation and fines. Operational risk is never really anticipated, but companies must be prepared for it.

Perhaps the most challenging feature of operational risk is that it is not necessarily easy to identify once it has happened. Decisions that involve the implicit acceptance of operational risk may not clearly expose the operational risk involved.

Lastly and perhaps the most important reason to focus on operational risk is that it remains a material cost to the enterprise. It directly impacts profitability and needed capital. For insurers, research indicates that marginal capital is more expensive to insurers than to banks, making the impact of marginal operational risk more dangerous for insurers than for banks.

**Operational risk – types and drivers**

When evaluating operational risk in the insurance industry, the Basel committee framework on operational risk provides a valuable set of operational risk types. The Basel framework outlines seven types of operational risk (Exhibit 2.2).

*Internal fraud:* this fraud typically involves the misappropriation of assets by employees and is generally seen as malice against the company. Fraud is not restricted to just stealing; it can be seen in accounting misstatements and other actions that deceive shareholders.
Operations and IT risk

Exhibit 2.2

Types of operational risk in the Basel framework

<table>
<thead>
<tr>
<th>Types of operational risk</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal fraud</td>
<td>Unauthorised transaction resulting in monetary loss.</td>
</tr>
<tr>
<td></td>
<td>Embezzlement of funds.</td>
</tr>
<tr>
<td>External fraud</td>
<td>Branch robbery.</td>
</tr>
<tr>
<td></td>
<td>Hacking damage (systems security).</td>
</tr>
<tr>
<td>Employment practices and workplace safety</td>
<td>Employee discrimination issues.</td>
</tr>
<tr>
<td></td>
<td>Inadequate employee health or safety rules.</td>
</tr>
<tr>
<td>Clients, products and business practices</td>
<td>Money laundering.</td>
</tr>
<tr>
<td></td>
<td>Lender liability from disclosure violations or aggressive sales.</td>
</tr>
<tr>
<td>Damage to physical assets</td>
<td>Natural disasters (for example, earthquakes).</td>
</tr>
<tr>
<td></td>
<td>Terrorist activities.</td>
</tr>
<tr>
<td>Business disruption and system failures</td>
<td>Utility outage (for example, blackout).</td>
</tr>
<tr>
<td>Execution, delivery and process management</td>
<td>Data entry error.</td>
</tr>
<tr>
<td></td>
<td>Incomplete or missing legal documents.</td>
</tr>
<tr>
<td></td>
<td>Disputes with vendors/outsourcing.</td>
</tr>
</tbody>
</table>


research by the US Federal Reserve System, internal fraud is shown to negatively impact company stock price, especially when shareholder involvement is high. In insurance companies, fraud also includes management actions that purposefully obfuscate or disguise information that is material in claim judgments. Such actions, when discovered by regulators, often result in punitive penalties, meaning that the impact of fraud is ultimately adjudicated by investors and regulators.

External fraud: this is essentially theft against a customer or the company. It is robbery from outside the company walls and may involve organised crime outfits and sophisticated techniques. As opposed to internal fraud, there is often a customer who is negatively impacted, either through their accounts or identity theft, meaning that the company must also resolve the impact to the customer.

Employment practices and workplace safety: as insurance companies employ many people, the disagreements and dangers resulting from employment are a form of operational risk. In the US, owing to our litigious nature for resolving conflict, this is frequently manifested by employees suing the company for specific management actions. This might be due to preferences in hiring, awarding of bonuses, and so on. In addition to the costs and settlements of employment lawsuits, it is important to keep in mind that employment lawsuits often involve the flight of talent and can make attracting new talent more difficult. Often these more persistent impacts are not measured in this form of operational risk.
Clients, products and business practices: in the operations of an insurance company, there are many opportunities for product misuse, client requests and business practices to pose risks. Moreover, the interpretation of product terms, the enforcement of product terms and the action by insurers against policyholders are all grounds for disagreement and litigation. In total, these risks are typically seen through lawsuits against the company or may be enforced by regulatory action, too. Product policies can be viewed after issuance, with a new lens on what is appropriate to the customer, causing new risks to the company. Clients, too, may pose unique risks, such as conducting money laundering or transfers that put the company at risk of regulatory or accounting inquiry. For insurers, as research shows, this is the predominant form of operational risk.

Damage to physical assets: this risk is the most visual, involving the loss of physical assets from natural hazard. Although dramatic as an operational risk, physical damage rarely results in the failure of a company. In fact, it is the insurance industry that normally provides risk transfer vehicles for the management of hazard risk. It is important to point out though, that while the loss or damage to physical property can be insured, it is unlikely that the lost business opportunity can be protected or recovered.

Business disruption and system failures: in an insurance operation, the rhythm of decisions may not be as instantaneous as that of a trading floor in a bank, where a few seconds of disruption can result in millions of dollars lost. For insurers, momentary disruptions from mission critical systems are not likely to result in significant losses as they would at banks. However with the continued reliance on critical internet and IT systems, the impact to customers and agents can have large downside, potentially also resulting in reputational damages. It is worth noting that systems that operate investment platforms of insurance companies do pose some operational risk. However, as few insurance companies take on highly leveraged or aggressive investment strategies, the impact of down systems, even on the investment side, is less than would be expected at banks.

Execution, delivery and process management: this category deserves much attention in insurance and should probably be divided in multiple categories. In essence, this category involves all forms of operational risk stemming from internal errors or failures. It suggests that although management intent might be correct, the actual execution can fail and result in losses. In the insurance industry, the focus and importance on underwriting cannot be overstated, and there is a general reliance on centralised, electronic systems to process underwriting. This has increased the possibility that an error can touch many accounts and lead to great operational risk.

In research by Cummins et al., the manifestation of operational risk in US insurers is almost entirely in the ‘clients and products’ category. In Exhibit 2.3, actual operational loss data from US insurers was examined. The category of ‘clients and products’ dominates in both frequency and severity, indicating that for insurers, operational risk is highly dependent on interfacing with customers.

The concentration of the occurrence of operational risk differs between insurers and banks. For banks, ‘clients and products’ also produce the largest form of operational risk, however, ‘internal fraud’ generates a much higher risk in banks than in insurers. Specifically, operational risk stemming from ‘clients and products’ are about 60% in frequency and severity, and risk from ‘internal fraud’ is approximately 25% in frequency and severity at
banks.\textsuperscript{11} For insurers, risk from ‘internal fraud’ is less than 10\% in severity and frequency (Exhibit 2.3). These differences between banks and insurers are because of the more active trading role found in banks (especially investment banks) and the ability for bank employees to trade in large sums. Many risk managers would point to the fraud cases of Barings Bank and Société Générale as prime examples of this point.\textsuperscript{12,13}

According to a research study of European insurers, in the insurance industry the largest perceived operational risks were internal fraud and deliberate employee misconduct.\textsuperscript{14} The implication is important in that risk managers are presumably focused on looking for internal fraud within insurance companies more so than looking at the risks that arise from ‘clients and products’. This reminds us of the biases and judgment lapses experienced with perceived large and severe risks.

On the whole, operational risk for insurers has increased greatly and continues to rise.\textsuperscript{15} There are many explanations on why it has increased. First, it is being separately detected and tracked as opposed to being lumped into other risk or cost categories.
Operational risk in the insurance industry

Second, the regulatory environment for insurance has grown, meaning that more errors that impact customers are tracked and even acted upon by regulators. Third, insurers have grown in size and complexity, now relying on large centralised IT systems and independent agent sales models, both of which lead to increased operational risk. Additionally, most major insurers have absorbed businesses through mergers, accepting the operational risks in those businesses.

Risk contagion from operational risk

Many risks in business are intertwined and linked; often one risk leads to another. Indeed, many risks driven by economic shocks happen simultaneously and build on each other. That phenomena, although troublesome to the business, is a form of contagion. In operational risk especially, we see that risks, when manifested, give rise to new or additional risks. This is especially true when the risk in question is left unattended.
In some ways, the risk contagion might be expected. Specifically, a shock to an enterprise might weaken its capital position and expose it to additional risks in funding or an inability to sell new products to customers; this is typical. In the example of a weakened company experiencing new operating or funding risks, the impacts of such shocks are most likely to exist over a prolonged period of time. The additional risks are implicit in that the companies did not specifically elect to accept them.

Another unfortunate feature of risk contagion is that it often activates a risk that is driven by external forces. In the insurance industry, the inherent disagreements that arise between the insurer and policyholders often lead to reputational and regulatory risks. The danger of this contagion is that operational risk (to the extent that it is driven by internal processes and systems) is controlled largely by the actions of the enterprise. Once the risks give rise to another risk driven externally through contagion, the control mechanism is lost. The fate of the company is, to some regard, put in the hands of customers, regulators, investors, and so on.

Preventing risk contagion is a key benefit of effective operational risk management in the insurance industry. As the research illustrates, most operational risk is due to ‘clients and products’ and is realised as legal or regulatory fines.16

Reputational risk – a by-product of operational risk contagion

When mentioning reputational risk, most companies worry over the potential loss of customer trust which has been built over many years. This is because insurance companies rely on reputation and trust as the hallmark of their relationship with customers. The storied example of Lloyds of London paying in full all claims stemming from the 1906 San Francisco earthquake (and that it is remembered today) is testament to the importance of reputation in the insurance industry.17 Companies that have trust and promise as their major asset are especially sensitive to reputational harm. This explains why the accusations against the audit and accounting company, Arthur Andersen, proved so tumultuous and why actions by various insurers in the 1980s and 1990s to deny and reduce asbestos claims harmed insurance-customer relations for many years.18

Fulfilling the explicit and implicit promises is critical in the insurance industry. For example, insurers are expected to have sufficient reserves needed to provide the protection that customers purchased. The initial market entrants in the long-term care insurance business exercised some especially questionable decisions in denying coverage to ill and dying customers.19 This reputation continues for the industry, even though it involved few companies and impacted a small number of customers; the entire industry suffered reputational harm from the actions of a few companies. Reputational risk is challenging to enumerate, because it is unclear which constituents will react negatively to the company and to what degree they will do so. Consider the following definition of reputational risk: ‘Reputational risk is the potential that negative publicity regarding an institution’s business practices, whether true or not, will cause a decline in the customer base, costly litigation, or revenue reductions.’20 Investment to overcome reputational harm is generally expensive and may indeed be impossible. Reputational risks linger over a long period of time and are implicit in the business operations of the company.
Operational risk in the insurance industry

Regulatory risk – another by-product of operational risk contagion

Regulatory and reputational risks share an important and unfortunate similarity: they are often perpetuated by an offended or motivated agent (regulator or customer) seeking a judgment or penalty against the company. Regulatory risks generally reference or involve customer harm. The economic impact of these risks means that companies must manage reputation and regulation through a series of on-going activities. These risks, when manifested, cannot be directly controlled by the company through internal operations.21

The recent movement by US and European governments to consider new regulation in financial services and other consumer-facing industries has drawn new attention to the reality of regulatory risk. In many ways, regulatory risk can be viewed as a game changer in the industry in that it can help some market participants and harm others. Specifically, regulations may not be objectively applied and may even be directed at specific market participants for political or economic reasons. Reversing regulation, while not impossible, is difficult and expensive as it often becomes ingrained in US code. Regulatory risk may arise for many reasons, but it is more often caused by customer concerns about practices by a specific company or an entire industry.

The insurance industry is subject to more regulation than almost any other industry, except perhaps the pharmaceutical industry. Regulation varies by state and product, and governs even capital holdings. Regulatory bodies often look to customers for direction on concerns and thus, the experience of customers can and does directly drive additional regulatory scrutiny.

The nature of insurance contracts draws special scrutiny from courts and regulators and provides some inherent empathy to policyholders. Insurance products sold to consumers are generally contracts of adhesion. A contract of adhesion is a legally binding agreement between two parties to do a certain thing, in which one side has the majority of the bargaining power and uses it to write the contract primarily to his or her advantage. Since insurers draft policy language and the consumer cannot negotiate the terms, but can only accept or decline the terms drafted by the insurer’s attorneys, a great deal of ambiguities are interpreted by the courts in favour of the policyholder. This means that errors by the insurer in policy design or terms can result in operational risk, either through lawsuits or through regulatory review once a disagreement with a policyholder arises.

The reputation of a company can and does impact the way that regulators view the company and whether punitive treatments are in order. As we will see, operational risk is often the basis for how customers become harmed in the insurance industry, suggesting that treatment of operational risk will help with the management of not only reputational risk, but also regulatory risk.

Operational risk is hidden in product performance

For credit and market risk, detection and measurement is straightforward. In market risk management, one monitors asset prices. In credit risk management, one monitors credit markets or the financial health of the borrower. In either case, the outcome of risk is made obvious through an evaluation of an asset (either a loan or other asset). Market risk and
credit risk signals are clear: ‘Prices are down! Loan payments are not being made!’ The risk manager can easily survey the market and all its agents. In addition to dynamic updates, credit and market risks generally allow for some realisation in a rather short period of time.

Consider a bank that issues mortgages and forgets to confirm borrower income during the loan application. This form of operational risk occurred during the real estate boom leading up to the Great Recession. The missing information meant that credit information was subject to error and misrepresentation. The error was not due to a credit policy but rather to an operational failure to confirm data. This is a process failure and is clearly an operational risk: the bank becomes adversely exposed to possible loan default. The loss and impact from this error is not manifested as an operational cost. Rather, it is manifested as higher delinquencies in loans than the bank likely predicted. Late loan payments or no payments at all may be an early-warning sign of default, but this does not fix the problem of the delinquent loan and the number of other likely delinquent loans. Fixing the problem in this mortgage example does not come from changing the credit policy, but rather by changing the process of confirming borrowers’ income.

In the insurance industry, accepting a customer’s liability places great importance on the processes for selecting and accepting risk. The processes of product development and policy underwriting are generally in place to provide the insurance company with the desired risk at (presumably) the right price. The importance of product development cannot be overstated. For this discussion, we will consider ‘product risk’ to be those negative outcomes arising from improper product structure or features, such as unfavourable pricing or ambiguity in terms. In the process of developing a product, insurance companies rely on actuarial data to provide statistical measures of the product’s performance and confidence in the purchasing of the liability. Indeed, since it is part of a statistical measure and involves the behaviour of customers, there is variation in projection performance and the possibility that customer assumptions were wrong. The risks arising from the variation and imprecision resulting from predictive data models, we will term ‘actuarial risk’.

While the impact of operational risks in a bank is manifested in the performance of its loans or assets, in insurance, we find that the impact of operational risk is manifested in the performance of its products. However, the duration of insurance products and the nature of the liability may make the detection of operational risk harder in insurance than in a bank. Since liabilities in an insurance company are not realised until many years or decades into the future, after product issuance, deviation in payments by the customer may not serve as an early-warning sign. Small errors, owing to the long duration of the product, can drastically alter the financial performance of the product too, meaning that there is high sensitivity in the insurance industry to errors in the model development process.

The insurance industry operates successfully when the projected timing and magnitude of the liability for a book of policies are aligned with assets. Loss performance relative to projections and profitability goals are typical benchmarks. Presumably, building up some level of assets happens concurrently with a build-up of liabilities, so the insurance company is sufficiently capitalised for its liability projections. However, when liabilities are manifested earlier than expected, it can be because the actuarial assumptions were wrong, the product was improperly priced or designed, or the model development process was flawed – essentially, product and actuarial risks are realised, which include the underlining operational risk.
Post analysis of these factors years later can be challenging, if not impossible, in determining which of these three explanations was the cause.

This confluence of risks makes operational risk detection even more difficult in the insurance industry. Consider a data processing error of an individual’s health condition at the time an insurer is underwriting a life insurance policy (a not so unimaginable operational risk). The products are underwritten and booked, resulting in adverse consumer selection by the insurance company. The customer makes premium payments (maybe because he did not expect coverage or even such cheap coverage). As a result of the data entry error, policies are written that take on a higher mortality risk than expected. The higher loss than expected might be interpreted as product or actuarial risk, but its real cause is an operational risk – failure in a data collection process. The damages to the insurance company are not seen in operations but rather as higher liabilities, or sooner realised liabilities than expected. Since life insurance companies trade premium payments (assets) for long-term liabilities, the deviation in the timing and size of liabilities observed is of great concern. The impact of such an operational risk is seen as negative performance, relative to pre-issuance expectations. Fixing this error does not come from changing the insurance policy, but rather in changing the process for data confirmation. More dangerously, the impact of the error may not be seen for years or decades, making detection challenging, if not impossible, to separate from the actuarial risks of the product.

**Product development and model development are critical processes in managing operational risk**

Life insurance companies have a lot of data and perhaps some of the best actuarial models in the insurance industry, owing to their need to predict life expectancies across time and space and take into account the advancements in health care that extend longevity. Such depth of information and robust customer risk profiles may not be present in other insurance products. The development of product and risk models outside life insurance often rely on critical and novel assumptions, especially as related to customer behaviour that ensure risk identification.

The perils of long-term care are now known to the insurance industry, but early market entrants in long-term care assumed incorrect lapse rates, and did not properly account for actuarial risks such as the increase in longevity and associated diseases (for example, Alzheimer’s). We now know too that customers that selected long-term care behaved differently than those who did not select long-term care at the time. The result was much higher long-term care liabilities than previously expected by the industry. Was that model error or product error? Yes, there was error, perhaps in both the model and the product, but the process for developing a product and its corresponding model possesses operational risk.

For many companies, identification of critical model assumptions is problematic, because many models were developed and implemented years prior. Moreover, blocks of business might have been acquired from other insurers over the years and even the exact policy terms may have varied widely. Examining the block of business for critical assumptions requires mapping modes to products and tracking critical customer assumptions and observed
behaviours. This is operationally intensive and complex. Specifically, many organisations use spreadsheets as the analytical tool of choice and critical data may reside in hundreds if not thousands of spreadsheets. So the actual process of developing and managing models and products is wrought with complexity. As complexity increases, we see operational risk increase. Although this may be due to errors or unintended mistakes in implementation, it is also due to the omissions by the organisation.

Understanding the actual path and reason for an operational risk is especially challenging. In an examination of operational risk in insurance, Acharyya\textsuperscript{24} notes that knowledge of the cause of operational risk is an overarching problem in financial services and especially problematic in insurance. Acharyya further notes that many insurers fail to separate operational risk from underwriting risk, meaning that its true impact is not fully known to insurers, and approaches to how to manage it remain highly ineffective.

**Operational risks are often manifested as lawsuits and/or penalties**

The insurance industry is wrought with regulatory minefields. As previously mentioned, the rules governing policies and policy payments vary from state to state, making adherence to each and all regulations nearly impossible. As many insurers can testify, the actual execution of policies is challenging. In addition to a highly charged regulatory environment, there exists a climate for customer empathy. Inevitably, insurance companies are faced with paying a claim or denying a claim, and such a black and white outcome is problematic when the trigger events for the payment are not clear or are subject to multiple interpretations. The long-term care insurance business, as an example, is full of such dilemmas.\textsuperscript{25} When an insurance company declines a payment, the customer, depending on their state of residence, can appeal or bring the matter to the attention of regulators. Customers can also, of course, bring lawsuits to challenge the decision of the insurance company in denying a claim. This often triggers an investigation, or results in a court order for the company to pay the claim and a punitive charge. Any error or perceived error by the insurer in this process can result in an operational risk, often amplified through a penalty or lawsuit.

Countless examples exist of a customer being denied a claim because of an error in the treatment of some critical piece of information. Payments can also be delayed, presumably because of an insurer’s slow or otherwise overtaxed process for claim evaluation. In each case, the operational errors impact the customer and demonstrate that processes drive the operational risk and the customer experience. This is critical to point out. Equally important is to consider the emotional state of the customer making a claim. Such customers have suffered a devastating loss of life or property. Through a legal lens, a victim versus an insurer with a failed operational process that denied a legitimate claim leads to a tough situation for the insurer. The contrast is dramatic, making it nearly impossible for the insurer to challenge the customer or their intent.

As demonstrated in Exhibit 2.5, operational risk to insurers is mainly manifested through legal and political costs, namely lawsuits and regulatory fines.\textsuperscript{26} Additional operational loss comes from actuarial and unintended human errors. Theft and money laundering, although noteworthy when occurring, are the least likely types of operational risk to US insurers.
Exhibit 2.5

Operational losses are most heavily manifested as lawsuits and regulatory fines

![Operational Losses Chart]

- Loss in US$ million per year per risk type (magnitude × frequency) according to US study


The data supports the points made earlier that processes to develop products and models that manage product risk and profitability are critical in the insurance industry and require in-depth analysis. Additionally, preventing or managing lawsuits stemming from client and product risk are important.

**IT capabilities become increasingly critical**

Exhibit 2.5 illustrates the reporting and accessibility of critical information to insurers, which puts at the forefront the state of critical IT systems at insurers. Of course, given the duration of products in life insurance, many insurers have policy information from legacy systems primarily due to acquiring blocks of business through mergers.  

IT systems at insurance companies must successfully collect, track and standardise myriad information. First, the tracking of critical documents, internally and from the policyholder,
is important not only for regulation but also for internal governance. Second, companies need data easily accessible to respond to queries accurately. Third, insurers are expected to provide all information promptly, since customer’s expectations for instant access grow with changing technological norms for immediate access. Relatedly, as many insurers rely on Internet and mobile applications, the risks from these IT platforms will surely result in increased operational risks. The danger with such IT platforms is that these involve customer interactions and the possibility that many accounts can be compromised. Insurers will continue to rely on these IT systems more to reduce operating expenses. In sum, the tradeoff of quick response versus controlling multiple platforms is an increased exposure to operational risk, meaning that insurers must continually maintain, improve and monitor their IT systems.

**Principal-agent problems are magnified in insurance**

The fierce competition in the insurance industry that has resulted in a focus on the cost of operations, and the importance of IT systems to control cost also has given rise to the independent broker sales model. For many insurers, especially those in life insurance, there has been a movement to work with independent insurance brokers or agents. The practice is widely adopted and offers unique and attractive benefits to both parties. For the insurance company, talented and successful agents can be sourced without large up-front training expenses. However, embedded in the independent agent model is a great deal of operational risk and selection bias. For the independent agent, the goal is to maximise commissions and thus maximise the acceptance of the policy by the insurer. The insurer stands to accept a great deal of liability, whereas the agent is poised to sell policies. The capital at risk is much different. Even in the absence of nefarious intent, the independent agent and insurer have a misalignment of goals and how risk is realised. This is always the case between employers and employees. Moreover, the independent agent might influence the customer to take on a specific product or insurer based on his or her previous experiences with insurers. The independent agent can and does influence the selection of customers and might even advise the customer on how to answer specific questions.
An asymmetry of information develops as the principal and agent each strive to maximise un-aligned goals.


The challenges posed by the principal-agent problem can presumably be addressed by a thorough and rigorous underwriting process: insurers use collected information on the applicant to assess the risk and the price needed to accept the risk. However, the independent agent is the one marketing the product and can materially impact the type of customers recruited, thus changing the holistic performance of the insurer. In sum, careful evaluation of agents (independent and employees) is important, since the principal-agent problem, as manifested in the independent agent model, can exacerbate operational risk to the insurer.

Impact and importance of operational risk in insurance

Risk management in financial services companies has always involved a combination of credit, market, product and operational risks. However, companies are operating under increased regulatory scrutiny with greater complexity and scale and now place more focus on operational risks.

The Economist Intelligence Unit surveyed 180 chief risk officers (CROs) of banks and insurers on their outlook on risks and the impact of those risks on their enterprise. The results are summarised in Exhibit 2.7.29
Most interesting is the CRO’s ranking of their concerns and each risk type’s potential impact. Reputational and regulatory risks are ranked first and second, but operational risk types are amongst the very largest risks of concern by the CROs. As mentioned above, reputational and regulatory harm are often the product of lawsuits and client/product disagreements, which at their core, are operational risks. Moreover, the outlook by these risk professionals is telling in many ways in that companies that excel in managing both operational risk and the complexity of reputational and regulatory risk should enjoy success in executing their respective strategies. Less impactful forms of operational risk, which are unlikely to harm the company’s performance, though more visual to the public, may be a distraction to the company in executing its strategy.

Research from Cummins et al.\textsuperscript{30} also suggests that operational risk is more expensive and persistent for insurers than banks. Exhibit 2.8 shows data that tracks the share price drop for banks and insurers after an operational loss announcement.

\textbf{Source:} EIU, ‘Regulatory risk and trends for the CRO’, \textit{Economic Intelligence Unit}, July 2005
For many insurers, the operational loss announcement comes in conjunction with a regulatory penalty or legal judgment, affecting the company’s reputation and stock price. There are many other impacted parties, whose concern can create further risk contagion:

- policyholders might question the solvency of the insurer;
- independent agents might not recommend new clients to the insurer;
- regulators might take a more negative view of a company and exercise greater review;
- employees might question the safety of and future of their careers; and
- investors, as the data in Exhibit 2.8 suggests, expect less profit from the company.

Exhibit 2.8

Share price drop for banks and insurers after operational loss events

Excursus: cumulative abnormal return for banks and insurance companies before and after an operational loss event.

Management actions for controlling operational risk in the insurance industry

Management must actively perform a deep dive into the causes of operational risk. Operational risk is often hidden to the company and hard to detect from actuarial error, but is embedded in product design and actuarial processes. Operational risk often results in early liability claims or unexpected costs through lawsuits and regulatory penalties. Understanding operational risk requires a deep examination of the path to the errors. The following are a list of major steps that insurance companies (and presumably other financial services companies) can make to better manage operational risk.

1. **Focus on critical information, especially in product development and underwriting**: in today’s business operations, where algorithms process information quickly, an error in a process can impact thousands of policyholders. It is important to understand which data are most critical, especially in the development of a new product, and how that critical information can be verified in underwriting.
   - Deploy a process to verify the most important information needed in underwriting. Rarely does the underwriting fail as a whole, but rather critical information is poorly sourced or extended beyond its appropriateness in actuarial models. This will bring focus to the importance of how information is gathered and actuarial models verified.
   - Verify and ensure the integrity of any information pertinent to risk while also working to identify new information. This can be done via internal audits or redundant checks. In an information-rich world, the integrity of the information has never been more important.

2. **Examine critical processes in underwriting and model development**: in any business, there are critical processes that impact the customer experience and the profitability of the company. In insurance, the importance of the underwriting process cannot be overstated. Additionally, processes that touch the customer are important.
   - Examine how information flows to the customer through marketing channels, independent agents, policy binding, account management and through a claim. Customers may interact with brokers, independent agents, websites and company employees, each providing different perspectives on a policy.
   - Examine how internal processes can and do break down. Operational risk arises from failed processes. The processes that impact customers are most likely the ones that generate operational risk. For instance, investigate delays and errors in processing through root cause analysis. An error in one policy can help prevent many more.
   - Develop capabilities that can contain operational failures. Focus intensely on those processes that touch all customers and transactions that pose the highest leverage risk. The movement to large-scale platforms has brought great economic efficiencies, but the downside is that small disruptions can result in large-scale impacts.
   - Although risk transfer is never perfect, leverage best-in-class technology, best practices and expert partners, to handle operations that are outside of core competencies. However, measure and track the risk events of partners and have in place key risk indicators to manage partners.
   - Invest in operations to protect the customer experience and brand promise before a risk event. This is far less expensive before the event. Companies often fail to link
investments to protecting the company’s brand and reputation with operational risk. Damages to the brand are expensive if not nearly impossible to overcome.

3 Look for explanations for ‘unexpected loss’: when product claims materialise and the losses are higher or realised sooner than expected, it is tempting to assign the errors to the product design or to critical actuarial assumptions. Often these are the case. However, it could be that these errors stem from customer selection. The process of customer selection is the result of marketing, underwriting and management. It is important to determine how an unexpected loss comes to the company. Is it because of the execution versus the policy? If so, it is operational risk and not product or actuarial risk.

- Determine if operations may be taxed or insufficiently examined and if human systems and suppliers are operating outside of norms. Rapid growth can often disguise and make difficult the reporting of unusual results. Extra effort and focus is required to overcome these forces.
- Examine losses that are small, too. It is tempting to disregard small, unexpected losses as immaterial. However, some sampling of the losses for in-depth analysis is important to draw out meaningful lessons. This will also set a cultural tone within the organisation that losses are to be examined and managed through explanation.

4 Measure the losses – bring focus to the errors: measuring losses is necessary so that improvements can be made. Otherwise, operational risk will remain hidden in product risk unless explicitly examined. It is thus critical to develop internal metrics that look at the impacted number of accounts, the impact of operational errors on customer selection, and the role of such errors in increasing loss claims. This places company-wide focus on operational risk, separate from just examining overall losses or claims.

In other industries where operations are critical, there has been tremendous focus in the use of metrics to drive errors down. The manufacturing industry and many aspects of call centres have embraced Six-Sigma analysis and LEAN analysis to bring focus to the errors. For many IT and service-oriented companies, the development of service-level agreements (SLAs) are used to examine how performance can and should vary during stressful periods of operation.

In addition to tracking errors and setting team goals to reduce errors, it is important to seek information on how and why errors occur in the first place.

- Development of operational risk metrics is critical. Create metrics that focus on how operational data links to the profit function in your company. Measure the operational loss events in dollars lost. Operational risk metrics must ultimately link back to the profit function of the company.
- Use risk information to identify scenarios that impact future profits. Investments early on to prevent such losses are less expensive than investments later on that attempt to recover lost market share.

5 Seek explanation for errors: it is challenging to unravel how and why a mistake ever happens in business. Why did the data load fail? Despite great technological advancement, mistakes happen, and often. It is important that examination of the operational errors occur so that the erroneous processes or practices can be stopped and prevented from going forward.
• Develop a team that audits performance and investigates discrepancies for additional learning. Often companies look at the cost of the error to examine it through some lens of materiality. Although reasonable in terms of short-term financial expenditure, that outlook fails to resolve growing operational risks in the company.

• Understand the holistic operational environment, in great detail. The time spent understanding processes leads to opportunities for improvements too. In a world where operations are complex and scale across millions of accounts, focusing on the seemingly small details in operations is critical and will reduce operational risk.

• Encourage the investigation of the unknown, misunderstood, perplexing and complex. Insights gained from that enquiry become valuable company information and ultimately lead to competitive advantages. This is a goal motivated by management. Disregarding errors, especially small ones, is counterproductive to developing a culture that can address errors.

6 Create a culture for seeking improvements and for using a data-driven approach to making improvements: when investigating errors and losses, it is important to realise that some team members may feel pressured or uncomfortable if the investigation looks unfavourably upon them or their team. Overcoming this natural human reaction is only possible if a culture is in place that values continuous learning and all team members see it as their duty to improve and assist in the discovery of errors. Setting this culture and tone is best done by senior management and reinforced with actions that legitimately support and recognise process improvement, audit investigations, and overall risk management as critical to the existence of the organisation. Elevating these functions by authority, title, recognition, resources and pay will go a long way in developing a culture for seeking improvements.

• Determine if there are risk aversions or risk-seeking tendencies in historical company behaviour. These behaviours are built over time and may lead to faulty risk decision-making. Recognise the history and make a clean break if the history is problematic.

• Overcome institutional bias and confirmation bias. Drive decisions from a thoughtful consideration of data. Develop and examine counterarguments. Understand how the counterargument fails. If the counterargument does not fail, examine your assumptions and treatment of data.

• Develop a work culture that rewards investigation and examination. Perfect surveillance is never possible, so develop a culture that welcomes contradicting information and challenges the status quo.

7 Develop an operational loss database: successfully learning from past mistakes means they are not repeated. With respect to operational risk, this requires understanding the details of the risk and its origin. The financial services industry has moved from collecting and storing information about operational loss events, to using historical data in the development of statistical models for the prediction of operational risk. Although the temptation to fit statistical distributions to historical data is present, a key benefit of operational loss databases is that the organisation can examine the nature of a loss. How did it happen? How was it discovered? How were customers, if at all, impacted? How was it remedied? Answering such questions helps in identifying processes that are critical and prone (due to complexity, say) to increased operational risk. While the mathematically-inclined see the operational loss database as a source of critical information for modelling, such statistical
models fail to explain the operational risk. This loses sight of the most critical information in an operational risk database, which is: how and why is there operational risk? The lack of information in the path or the cause of operational risk, when modelling the operational risk with statistical models, remains a great and most appropriate criticism.

- Risks realised are not simply losses, but they are opportunities for the company to learn and improve. Operational losses are forgone profits, too! Open and honest discussions about what went wrong and how the company failed, in the wake of failing to recognise risk, encourages personnel to examine their work and even bring new information forward.

- When examining any operational risk information, examine what the information says about future profits. Protection of existing profits is critical and often more important than seeking additional profits in the short term. Remember that risk loss events ultimately erode profits in the form of asset transfer or market share transfers to competitors.

8 Embrace and practice customer centricity: customer centricity is more critical than ever in managing reputation and accompanying regulatory risks. With the advent of e-mail and social media, it has never been easier for a customer to spread his or her displeasure, discontent and experience with other customers and to bring their experience to the attention of regulators. Additionally, the speed at which this can happen has never been greater.

Management of the customer and his or her experience is in many ways a different function from claim processing. Customer centricity should proactively be holistic in the treatment of the experience, considering the communication and flow of information, as well as setting expectations for customers’ enquiries. Should a customer have a legitimate complaint, there is an opportunity for engagement (albeit reactive) that can hopefully diffuse a legally dangerous and expensive conflict. It is important to develop both proactive and reactive processes that focus on the customer experience and use metrics to determine if complaints are resolved or remedied. Having such processes in place can also expose internal decisions that were incorrectly made so that the claims process can proceed without additional conflict.

The benefits of a holistic approach towards customers can also ameliorate the language and accusations that may be used in a legal challenge too. Insurance companies that show a history of callousness or customer misdeeds are often subject to great scrutiny and fines by regulators.

- Recognise that processes, which impact customers negatively, can activate victim status in the customer and pose a unique risk to the company. Focus on making these operationally excellent.

- In business, it is seemingly easy for the customer to become a victim or perceived victim, given the failed processes of the company (or worse its negligence). Impacts to the customer experience and brand are hard to erase. Customers have an inherent bias to only remember key and highly negative customer experiences (they remember negative experiences four times longer than positive experiences). The brand and message are always at risk, and damages to them are inherently felt over a long period of time. Work vigorously to protect the brand and customer promise at all times. Proactively measure risk information that suggests trends in the health of your brand and validity of your customer promise.
- The ultimate judge and jury for any company is the customer. As companies move to a greater number of digital footprints, the impact to customers’ privacy becomes greater. Remember that in the absence of laws and legal direction, company actions, especially with customer data and information, will be tried in the court of public opinion. The same holds true for other socially-sensitive issues.

**Conclusion**

Operational risk in the insurance industry is increasing in importance, owing to its ability to impact the prudential of the company and an insurer’s relationship with its many constituents (shareholders, policyholder and regulators). Operational risk losses to insurers are increasing in frequency and magnitude, confirming that more operational risk is being realised. There are distinctions in the way operational risk is manifested in the insurance industry that need to be recognised. For insurers, the largest operational risk type is ‘clients and products’, reminding us of the importance of the insurer-policyholder contract. Operational risk also appears to have a greater impact on an insurance company’s share price than it does for a bank. Operational risk is also more expensive (in terms of marginal dollars) to insurers than to banks. In addition, operational risk in insurance companies remains hard to detect until deviations in product performance emerge. For insurers, operational risk is frequently realised through court or regulatory imposed fines. Understanding and reducing operational risk requires examination of complexity and errors, as well as a data-driven culture that is open to diving into the causes of the unexpected losses. Since operational risk does not offer an upside like credit or market risk, successful operational risk management can directly reduce losses to the company, protecting profits. Excellence in processes and a great attention to errors are important in managing operational risk. Given the critical processes of underwriting and product formulation, excellence in these processes should be a priority for insurance companies.

**Bibliography**


Ernst & Young, ‘Operational risk for insurers’, 2012.


2 Knight, F, Risk, Uncertainty and Profit, 1921, Cornell University Library.
6 Cruz, M, Modelling, Measuring and Hedging Operational Risk, 2002, Wiley.
Operations and IT risk

Chapter 3

The future of operational risk management in banking: from deterministics towards heuristics and holistics?

Brendon Young

Risk Research Foundation and Institute of Operational Risk

What we know is little, and what we are ignorant of is immense.

Pierre Simon Laplace

Overview

Operational risk management in the banking sector is expected to undergo fundamental change over the next few years. At the time of writing, it is said to be at least 10 years behind other industry sectors. Regrettably, operational risk management remains poorly understood, representing a serious barrier to its development. The failure of Basel II to prevent the recent financial crisis, despite significant investment in regulatory compliance by the banks, is seen as clear evidence of the need for change. Unfortunately, Basel III appears to be lacking intellectually and fails to advance the discipline of risk management. The Bank of England has identified the need for a move from deterministics towards heuristics and holistics. However, the need of regulators is fundamentally different to the need of the banking industry itself. Recent actions by banks appear to be more a reaction to immediate short-term socio-economic problems rather than recognition of the need for fundamental change and adoption of a longer-term perspective.

Risk management is failing

The raison d’être of the banking and financial services sector is ‘to transfer risk to those more able to accept and manage it’, however, it has been said that risk management in the sector itself is poorly understood and is at least a decade behind peers in other industries. Indeed, some believe that risk management in banking has effectively failed; the financial crisis, which plunged the world into the longest recession in history and threatened social instability in many European countries, being clear evidence of this. We are now faced with a financial system that has destroyed a very considerable amount of global wealth and created a level of debt that some believe may be impossible to fully repay. Certainly, Basel II, which despite being 10 years in gestation, proved of little value in averting or limiting the crisis; Basel III has added little, intellectually, to improvements in risk management.
As Keynes stated: ‘The world is a deeply unpredictable place, and uncertainty an inescapable part of economic life.’ Keynes believed that markets were not self-correcting, that economies could get stuck or sometimes sink. In such a situation it was up to the government to create the necessary economic stimulus. Hayek considered a different solution, believing that the fundamental requirement was to balance supply and demand, even if that meant removing excess capacity (including jobs) in a recession. Marx also believed that markets were inherently unstable and that a laissez faire approach, as originally espoused by M. Le Gendre in around 1680, would lead to inequality and inefficiency. Whilst these economists may strongly disagree on emphasis, their fundamental ideas are not entirely incompatible or mutually exclusive. In summary, markets are unstable and the ‘efficient market hypothesis’ is questionable. Indeed, the idea of rationality was also rejected by existentialists, like Kierkegaard and Sartre, who believed that ‘human reason has boundaries’, that reason is insufficient. Socrates said that ‘the only thing certain is change’. This is supported, scientifically, by the second Law of Thermodynamics, which states that in an unstable environment, change is inevitable and irreversible (that is, entropy increases). This gives rise to ‘the arrow of time’, that is, what is right today may not be right tomorrow; it is not possible to go back to how things were. Complexity and contingency theories suggest that the future is unpredictable and that actions may lead to unintended consequences.

With regard to uncertainty, George Soros stated: ‘My financial success stands in stark contrast with my ability to forecast events. In this context, we must distinguish between events in financial markets and events in the real world – even in predicting financial markets my record is less than impressive. With regard to events in the real world, my record is downright dismal.’ It is in the assessment of the future that uncertainty arises. Risk results from the impact of this uncertainty on chosen objectives (ISO definition). In the financial services sector it is the interaction of a myriad socio-economic factors that leads to uncertainty and risk.

To attempt to predict the future is born of naivety; to prepare for the future is born of necessity. The challenge is in determining the appropriate actions to be taken and controls to be applied, for whatever possibility comes to pass. Darwin recognised the resultant need for responsiveness, stating: ‘It is not the strongest of the species that survives nor the most intelligent but the one most responsive to change.’ It is against this background that we must judge the true relevance and value of risk management. By requiring a more forensic assessment, risk management enables complexity to be unravelled, facilitating responsiveness and transparency; transparency being essential for the market to work efficiently and effectively (if not perfectly).

The banking environment is changing – risk management must adapt

The banking environment has changed considerably in recent years, with resultant changes in risk management and regulation. Prior to the financial crisis, leading banks achieved an average return on equity of around 25% during the period 2000 to 2007; however, this has been reduced significantly, with single digit returns now being the norm. Future returns are less predictable, given the effective disappearance of certain lines of business (now considered
socially and politically unacceptable) and the possible movement of the centre of gravity of investment banking towards Asia. As a consequence, the banking industry in the UK, Europe and the US is faced with the task of increasing profitability through simplifying excessively complex infrastructures whilst also implementing new multifarious regulations that have far reaching consequences. It is an often quoted apophthegm in the financial services industry that ‘the seeds of destruction are sown in times of plenty’. Prior to the financial crisis, a myriad of mergers and acquisitions led to an increase in the size and complexity of organisations. High cost structures and inflexibility have made it difficult for complex organisations to scale back their activities in leaner times, to keep pace with the rate of decline. Although headcount levels may be unsustainable, cutting staffing is unlikely to prove sufficient. Risk management will require the removal of layers of complexity and inefficiency, in favour of more scalable, modular, integrated and responsive infrastructures. In addition, the significance of regulatory changes will require that organisations in the banking and financial services sector consider the potential impact upon their strategic business models. Regulatory compliance is no longer sufficient; the new environment requires consideration of the impact upon competitive advantage and profitability, now and into the future. Structural change requires new ways of working.

In April 2013, the Joint Committee of the European Supervisory Authorities called for action on cross-sectoral risks facing the EU financial system. Gabriel Bernardino, Chairman of the Joint Committee, said: ‘Today’s first report by the three European Supervisory Authorities has identified some key cross-sectoral risks facing the EU’s financial system including, in particular, the risk of fragmentation, over-reliance on collateral, concerns with regard to the quality of financial institutions’ assets, and the loss of confidence in institutions and benchmarks. This cross-sectoral work provides EU policy-makers and regulators with an overall view of the risks they face and moves us away from a narrow sectoral approach, and the inherent risk of failing to see the big picture.’

Regulation is inappropriate

Whilst banking has invested very considerably in risk management, most of this appears to have been for regulatory or compliance purposes and has done little to advance and improve the industry itself. Certainly, the Bank of England has questioned the performance of banking and the commercial relevance of some banking activities. With regard to regulation, the Bank of England has indicated that this is currently inappropriate: ‘The type of complex regulation developed over recent decades might not just be costly and cumbersome but sub-optimal for crisis control. In financial regulation, less may be more.’ Similarly, the former chairman of the now superseded FSA (UK Financial Services Authority), Lord Turner, also acknowledged that risk-regulation is lacking intellectually.

Whilst in principle, the need for regulation in the banking sector may have long been accepted, the appropriate level and type of regulation remains an area of some considerable debate. The IMF’s 2008 Global Stability Report sharply criticised banks and other financial institutions, stating that there was a failure of banks’ risk management systems leading to ‘excessive risk-taking’. It also stated that the situation was exacerbated through lax regulation by governments. Poor supervision of banks allowed the situation to develop further.
Indeed, financial sector supervision and regulation ‘lagged behind the rapid innovation and shifts in business models, leaving scope for excessive risk-taking’. The IMF concluded that more fundamental changes were needed in the medium term.

Similarly, in 2009, John McFall, MP, whilst Chairman of the House of Commons Treasury Select Committee, said that the FSA had demonstrated a ‘substantial failure of regulation’ and that light touch regulation had failed. He stated that: ‘Whilst there has been much debate over the appropriateness of a “light-touch” regulatory framework, it is clear now that, in the past, we have experienced not light-touch, but “soft-touch” regulation. Whatever the regulatory system will look like in the future, the days of soft touch regulation are over... The Northern Rock debacle, as well as problems at other major UK banks including RBS and HBOS, can be seen as part of a wider problem... The FSA has told the Treasury Committee that it did not consider it to be within its remit to analyse companies’ business models, question their appointments of directors, and so on; there was an assumption that the banks knew best in these areas. In light of the failure of this approach, the FSA has told the Committee that it will adopt a radically different attitude in the future.’

In 2011, the Bank of England confirmed its intention to move away from prescriptive rules, which it believes increase risk in the system, to a fresh approach that is more judgement based, relying on expertise and broad heuristic indicators.

**Beware excessive regulation and legislation**

As a consequence of failures, the regulatory and legislative framework has become more voluminous and complex, giving rise to the potential for unintended consequences. In terms of volume, Basel II amounted to 347 pages, which itself was an order of magnitude greater than the original Basel Accord (around 30 pages). However, Basel III is approximately twice the length, with related legislation possibly adding some thousands of pages more in both the UK and the US. Similarly, in February 2012, *The Economist* claimed the US Dodd-Frank Act was ‘partly unintelligible and partly unknowable’, and as a consequence there is ‘the risk that the Dodd-Frank apparatus will smother financial institutions in so much red tape that innovation is stifled and America’s economy suffers’.

Previously, Charles Haddon-Cave QC, stated, in 2010: ‘Pumping out complex, elaborate, prolix volumes of regulations may give the writers, the promulgators and the powers-that-be a warm feeling, but these are not helpful to those on the frontline who have to read, comprehend and act on legislation.’ There is a risk that risk frameworks kill mindfulness, by overly producing checklists, procedures and other ways of making the organisation operate. This points to the possibility of concealing dangers and lulling people into a false sense of security.

An over-reliance on prescriptive rules drives out intellectual passion and professional judgement, leading instead to protectionism of both organisation and self. Critics claim that regulation distorts the free market mechanism, thereby preserving and subsidising weak competitors at the expense of the more able, thus destroying overall economic wealth.6 ‘Regulation should emphasise issues of structure and incentives rather than control of behaviour.’7 Indeed, outside factors are considered poor at governing inside behaviour.
**The future of operational risk management in banking**

**Basel II and III – simply more of the same**

Unfortunately, the underlying principles of Basel II appear to have been forgotten in its application. Basel II was based upon the interaction of three pillars of control and management, these being: capital adequacy, supervisory review and market oversight. Basel III also follows this approach. The fundamental question to be asked is why this methodology appears to fail?

**Capital adequacy:** unfortunately, there has been an undue reliance upon capital adequacy by the regulators. Determination of the level of capital based upon models has proved to be inadequate. Models have tended to ignore correlation, which can vary over time, and have relied upon extrapolation (that is, an attempt to predict the future) in order to reach the 99.9% confidence level set by Basel. Models are merely simplifications of the real world. They, together with the data used, have a number of limitations, which need to be taken into account, not least of which is the fact that the past may not be a good indicator of the future.

A regulatory confidence level of ‘one event in a thousand years’ has been widely criticised, since the banking and financial services sector is a globally interconnected industry where contagion and illiquidity are real and ever-present threats. Coleman has suggested a simpler but more robust modelling approach based upon a chosen multiple of a lower but more robust confidence level (that is, 4 times the 97.5% confidence level, say) using a standardised model for all banks (to be selected by the regulators). It is important to recognise that capital is a vital protection against risk-events; however, no amount of capital can protect a fundamentally flawed bank, as illustrated by the collapse of Northern Rock.

**Supervisory review:** a fundamental difficulty in assessing a bank is that it is possible for an organisation to do all the right things but still to fail (for example, playing all the right musical notes but not necessarily in the right order). Equally, the presence of errors does not necessarily indicate failure at some future time. A primary consideration, which is well recognised in the venture capital industry, is the quality of the senior management team. It is the culture and ethics they engender, together with the quality of their strategic business plans (which takes into account responsiveness and resilience) that determines the level of success. Unfortunately, the FSA chose to ignore these issues, instead concentrating on the minutia, through application of its ARROW framework. In comparison, ‘Moody’s analytical framework for operational risk management of banks’, adopts a more strategic, integrated, approach.

**Market oversight:** market oversight is said to be more responsive than fundamental analysis as carried out by the regulators and rating agencies. However, it should be remembered that the market is not entirely rational. It exhibits herd behaviour, with players following one another. Although the existence of an economic bubble may be overt, players will continue to participate, in the hope of making money in what for them is effectively a zero-sum game. The resultant market correction can be excessive and lead to economic recession. Unfortunately, the reluctance of the regulators to provide information, which they considered to be too sensitive to release to the market, has served to undermine Basel’s market oversight pillar.
Baseline III

The new Basel Accord (Basel III) was introduced in 2009, by the Basel Committee on Banking Supervision, in light of failure of its previous risk-based capital accord. Basel III is an enhanced global regulatory standard on bank capital adequacy and liquidity, which is to be phased in between 2013 and 2019. It is claimed to be a comprehensive set of micro and macro-prudential reforms, designed to improve resilience by strengthening regulation, supervision and risk management. The aims of these new measures are:

• to improve the banking sector’s ability to absorb shocks arising from financial and economic stress, whatever the source, by enhancing capital adequacy, liquidity and leverage;
• to improve risk management and governance; and
• to strengthen transparency and disclosures.

The new Basel Accord, which includes detailed measures, calls for an increase in regulatory capital to cover previously ignored risks. However, it should be recognised that no realistic amount of regulatory capital could have provided adequate protection against the level of losses incurred as a result of the Financial Crisis. Interestingly, it should be noted that Dexia Bank failed more recently despite approval by regulators of its increased capital level.

In the UK, in 2010, the Independent Banking Commission (IBC) was set-up by the Chancellor of the Exchequer to consider structural and related non-structural reforms to the UK banking sector (that is, to include issues outside normal regulatory scope) in order to promote: (i) financial stability; and (ii) competition. With regard to financial stability, the Commission acknowledged that ‘shocks cannot be eliminated, and the UK will always be subject to global events’. IBC also addressed the ‘too big to fail/save’ issue, stating that: ‘Banks ought to face market discipline without any prospect of taxpayer support...’ Such support ‘gives the banks concerned an unwarranted competitive advantage over other institutions, and will encourage too much risk taking once the market normalises’. Similarly, the FSA stated: ‘Although the idea, that regulation should seek to eliminate all failures, may look superficially appealing, in practice this would impose prohibitive costs on the industry and consumers. Consumers benefit from healthy, competitive markets, where different companies try to meet their needs...’ Banks are always going to fail since that is the nature of competition. However, moral hazard exists if consumers do not face the risk of loss resulting from their own decisions and actions. The need for more enlightened regulatory changes has been accepted, as evidenced through the recent establishment of a European-wide bank regulator and the diminution and replacement of the UK FSA with the transfer of much of its former powers to the Bank of England.

Financial reporting is in need of reform

Unfortunately, financial reporting, which is a key component of market oversight and transparency, is considered to be failing in its purpose of properly informing stakeholders of the true risk-reward relationship. This represents a serious threat to the efficacy of the overall
The future of operational risk management in banking

financial system. Transparency is seen as essential for the market to work properly, allocating resources efficiently and effectively, as well as providing protection against fraud, corruption and financial crime.

A fundamental aim of accounting and financial reporting is ‘to provide a true and fair view’. However, ‘prudence’, long considered one of the fundamental concepts in accounting, appears to have been abandoned by recent developments in Generally Accepted Accounting Principles. Similarly, ‘consistency’, appears to have been relegated to a ‘desirable’ quality of financial information rather than a fundamental concept. This has led to widespread criticism of the international standard-setting body, IASB, both by academics and professionals alike. In a 2011 report on the financial crisis, the House of Lords bemoaned the demotion of prudence as a governing principle of accounting and auditing. In June 2012, the Daily Telegraph reported: ‘Banks are sitting on undeclared losses of £40 billion but are delaying de-risking and de-levering balance sheets due to accounting practices.’ The report went on to say: ‘PIRC, the shareholder advisory group, has analysed the 2011 accounts of the UK’s top five banks to calculate how much they expect to write off as bad debt in the coming years but have yet to take against profits. Royal Bank of Scotland (RBS) was in the worst condition, PIRC found, with £18 billion of undeclared losses that would wipe out more than a third of its capital buffer and potentially force the 82% state-owned lender back to the taxpayer for another rescue. HSBC had ($16 billion) £10 billion in undeclared losses, Barclays £6.7 billion, Standard Chartered $3.6 billion (£2.3 billion) and Lloyds Banking Group £2.9 billion.’

In December 2012, Boris Johnson, the Mayor of London, said ‘you can’t blame finance directors for doing their job and playing by the rules – that is what they are paid to do. However, it does raise the question of whether the rules are right.’ This issue highlights a further threat to the credibility of the overall financial system, as well as to the accounting profession. As stated previously, the fundamental issue relates to transparency and confidence. Indeed, it has been suggested that a lack of confidence is one of the primary underlying reason for the current on-going financial and economic problems in the UK.

Unfortunately, financial reporting currently attempts ‘to be all things to all men’. Hans Hoogervorst, Chairman of the IASB (International Accounting Standards Board), has acknowledged the problem concerning the volume of disclosure: ‘The last “grand theme” we are likely to address is the voluminous nature of disclosures. The plain fact is that businesses have become more complex. It is the job of financial reporting to describe this complexity, not to mask it. However, it is also true that there is the potential to rationalise disclosure requirements.’

Unfortunately, describing complexity does not serve to unravel and simplify it. Increased financial reporting tends to produce diminishing returns and leads to information asymmetry. Data is not information – information is not knowledge – knowledge is not understanding – understanding is not wisdom – and wisdom is not ‘action’. This raises the fundamental question, ‘who is this information for’ and indeed ‘do the published accounts give the right information’. Certainly, different stakeholders may have different requirements. Professor John Kay stated, in 2012, ‘Relevance to investors should be the principal criterion in determining reporting obligations’. Burgeoning information is typically a sign of distrust; no amount of information will solve such a problem.
The UK Government and the financial services sector have recognised the need for improvements and in June 2011, Lord Sharman’s Inquiry,\textsuperscript{13} commissioned by the Financial Reporting Council (FRC), produced its recommendations. In summary, it suggested that the ‘true and fair view’ principle should be enhanced by refocusing the ‘going concern’ assessment and integrating it into narrative reports on the business model, strategy and principal risks. The primary aims of the recommendations were to reinforce responsible behaviour and to demonstrate effective stewardship of the going concern status, through making directors more aware of and responsive to early indicators of distress, whilst also ensuring investors are better informed about those risks. The inquiry panel concluded that whilst banks’ funding models are inherently unstable, a special regime was not necessary. It considered the question: ‘should transparency be sacrificed for financial stability’, on the grounds that making the market aware that a bank was receiving special support risked undermining that support. However, the panel concluded that appropriate market transparency should be maintained, but not so far as to thwart the whole purpose of the permanent and special liquidity support made available to maintain financial stability; support does not constitute a material uncertainty if the bank is judged to be sound.

At a meeting of the Risk Research Foundation (ORRF), in October 2012, which was held to discuss ways of overcoming some of the current problems with financial reporting, it was suggested that published accounts should be restructured to provide a clearer, more transparent, view of a company, its performance and its future prospects. In particular, it was proposed that published accounts should be the subject of significant external assessment and assurance. Externality could be enhanced if published accounts were comprised of the following, although not necessarily exclusive, sections.

- **Chairman’s Statement:** the ‘Chairman’s Statement’ should be the main report to which all others relate. It should be comprised of three parts: (i) giving a concise but clear and unambiguous introduction and background to the company (that is, stating what its objectives and competitive advantages are and the reasons for any previous changes to these); (ii) detailing its current performance and overall competitive standing; and (iii) setting out its expected future performance (stating the assumptions upon which these assertions are made, identifying the main risks and challenges faced).
- **An Assurance Report:** it is important that an adequate level of assurance is provided in order to ensure credibility and confidence. The ‘Assurance Report’ should be prepared by the non-executive Chairman of the Independent Audit Committee and should serve to underpin the ‘Chairman’s Statement’. In the UK’s NHS (National Health Service), assurance is an essential consideration and, as such, a comprehensive framework has been developed. Non-executive directors (NEDs), whose primary role it is to ensure adequate assurance, exist, in part, as representatives of shareholders and the public. They are assisted in their task by various sub-committees and other bodies, including the internal auditors (who, as part of their responsibilities, should ensure the existence of an adequate reporting and whistle-blowing facility).
- **Statement of Adequacy:** a ‘Statement of Adequacy’ should be prepared by a separate company of external accountants or consultants, independent from the existing company of external auditors. The statement should comment on the adequacy and completeness of
the ‘Chairman’s Statement’ and the ‘Assurance Report’, as well as highlighting the main issues in the ‘audited accounts’.

- **Audited accounts and notes to the accounts**: summary statements of profits and loss, cashflow and balance sheets should be included in the published accounts. A detailed set of ‘audited accounts and notes to the accounts’ should be prepared in the normal way, but should only form part of the appendices. These documents could be made available via the company’s website and do not need to be part of the published report. However, this should not be interpreted as diminishing their importance.

- **Assessment of Standing**: an ‘Assessment of Standing’ should be prepared by a risk rating agency, such as Moody’s, Standard & Poors, or Fitch. This report should give a comprehensive risk assessment of the company and indicate its competitive standing. In particular, the report should identify the factors determining the company’s rating, those risks that could change the rating, and the probability of a rating change over the next five years, say. In addition, the report could consider those enhanced analytical factors (EAFs), which may significantly change the longer-term structure of the industry (for example, environmental, social and governance issues). It should be noted that this would require the rating agencies to have the same professional standing as independent auditors.

- **Additional Information**: whilst there may be demand for further information and, indeed, the company may wish to provide more information, this could be made readily available via the company website. A website allows various levels of sub-reports to be easily linked to the main report, thus facilitating analysis.

**Longer-term risk management requires assessment of longevity and richness**

The need for published reports and accounts to give a more holistic, longer-term perspective is certainly well recognised in the fund management industry. Short-termism and short-term reporting leads to volatility and a legalistic, moribund, state, as indicated in the Shell Trilemma diagram, thus reducing confidence and thereby restricting investment and growth. The Sharman Inquiry recognised the need for a more holistic view, incorporating multiple lenses of perception, with greater integration of accounts and risk management, in order to properly ascertain the ability of a business to continue to trade into the future.

In September 2012, Hans Hoogervorst stated: ‘In the last 10 years we have seen nothing short of a revolution in financial reporting... So, after the revolution, what comes next? What is the future of corporate reporting? There are various initiatives trying to address this question. One of the most important is the work to take a more holistic view of corporate reporting, known as integrated reporting. The goal of integrated reporting is to bring together many reporting requirements including sustainability, the environment, social issues as well as of course financial reporting. These topics are becoming more inter-dependent, and many investors want to understand the interplay between them. I also think for investors to properly understand financial statements, they are in need of non-financial key performance indicators.’

Assessment of a company’s ability to withstand the future involves consideration of such factors as its responsiveness, resilience and mindfulness. These factors are difficult to quantify. Instead, it may be necessary to determine, heuristically, the company’s richness. Rich companies are those that are rich in culture and leadership, rich in knowledge and
communication, rich from a sustainable perspective, and rich when considering factors such as contagion and liquidity, as well as performance. This requires identification of key characteristics and their determinants (that is, ‘Performance shaping factors’). Richness requires both an absolute and a comparative assessment against near competitors as well as against the market leader. Company characteristics can be divided into quartiles, with the top quartile being rich, the lower quartile being poor, the second quartile being impoverished in some way, and the third quartile being under-developed.

**Culture determines ethics – ethics guide principles – principles govern actions**

Leadership and culture are key determinants of success; however, socio-economic forces can bring about cultural change. For example, the raison d'être of banks appears to have changed considerably following deregulation (epitomised by Big Bang in the City of London in October 1986), which did away with unlimited liability partnerships and the principle of ‘my word is my bond’, instead ushering in a new level of virtually unbridled competition. As a result, banking moved from providing services for customers, towards a modus operandi approximating to trawling, with a view to maximising bonuses, profits and shareholder value, irrespective of customer perception.

Culture is a major factor influencing risk and the level of stability within the banking sector. Indeed, it may be more important than the reform of external regulatory structures. However, the difficulty in changing culture should not be underestimated, particularly in countries where free market forces have tended towards the creation of plutonomies. Professor John Kay emphasised the need for moral values; morals being something regulation cannot provide. In general, the response to cultural and behavioural issues has tended to be the introduction of further ‘box-ticking’ audit-based processes rather than addressing underlying root causes. Indeed, Lord Turner, whilst Chairman of the FSA, stated: ‘We simply do not know if we have the tools to change the banking culture.’ Despite potential difficulties, in 2010, the FSA affirmed its intention to improve the regulatory oversight of culture within the banking and financial services sector.

There is a widely held belief in the financial services industry, particularly in investment banking, that winning is everything. This has the effect of generating a culture of fear and greed. The personification of this greed culture was described in an article concerning Goldman Sachs, published in the *New York Times* on 14 March 2012. The article was written by Greg Smith, who at the time was executive director and head of the company’s equity derivatives business in Europe, the Middle East and Africa. In the article, which was entitled ‘Why I am leaving Goldman Sachs’, Smith states: ‘To put the problem in the simplest terms, the interests of the client continue to be side-lined in the way the company operates and thinks about making money... I truly believe that this decline in the company’s moral fibre represents the single most serious threat to its long-run survival... Goldman Sachs is one of the world’s largest and most important investment banks and it is too integral to global finance to continue to act this way.’

Speaking to the UK Parliamentary Commission on Banking Standards, in January 2013, Andrea Orcel, Chief Executive of UBS’ investment bank, told MPs that the banking industry must change. In an effort to recover its former honour and standing, UBS is in the process
of improving its corporate structure. The bank has faced a number of fines in recent times, including a US$1.5 billion (£940 million) fine for rigging Libor rates. The bank also paid US$780 million to US authorities for tax evasion and has been accused of losing very significant amounts on derivatives trading in the financial crisis, including US$2.3 billion lost by rogue trader, Kweku Adoboli. In an effort to reform its business, the bank has closed down its fixed income desk and cut 10,000 jobs. Orcel stated: ‘We all got probably too arrogant, too self-convinced that things were correct the way they were. I think the industry has to change.’

In an effort to change behaviour, if not the entire culture, the US Attorney, Preet Bharara, said, in January 2013, the ruling relating to offshore accounts, which resulted in the closure of Switzerland’s oldest private bank, Wegelin, (founded in 1741), represented ‘a watershed moment in our efforts to hold to account both the individuals and the banks – wherever they may be in the world – who are engaging in unlawful conduct that deprives the US Treasury of billions of dollars of tax revenue’. In February 2013, the new chief executive of Barclays Bank, Antony Jenkins, announced closure of its controversial tax avoidance unit, in a bid to repair the bank’s reputation. The tax planning operation is one of several areas being reviewed, with a view to assessing whether the bank’s businesses are ethical as well as profitable. The tax planning unit of the structured capital markets (SCM) division, which generated significant profits for the bank, has been accused of orchestrating tax avoidance on an ‘industrial scale’.

In the wake of numerous scandals, the ethical standing of bankers has fallen considerably. Their reactions to these scandals indicate cultural problems of a fundamental nature. In December 2012, the British Prime Minister, David Cameron, stated: ‘It is morally unacceptable to engage in actions for personal gain, on the pretext that they are legal.’ This raises the question of when does negligence or failure by a bank move from being a moral operational risk management issue, to one of regulatory concern and possibly something requiring criminal investigation. In December 2012, HSBC was fined a record US$1.9 billion in settlement of US money-laundering accusations. The US Senate said the bank had operated a ‘pervasively polluted’ culture that had lasted for years. Assistant Attorney General, Lanny Bank, said HSBC was guilty of a ‘blatant failure’ to implement money-laundering controls and that it had wilfully flouted sanctions: ‘HSBC is being held accountable for stunning failures of oversight – and worse.’ The bank’s chief executive, Stuart Gulliver, said he was ‘profoundly sorry’ for the failures. Gulliver stressed that the bank had co-operated with the US authorities: ‘Over the last two years, under new senior leadership, we have been taking concrete steps to put right what went wrong and to participate actively with government authorities in bringing to light and addressing these matters... We are committed to protecting the integrity of the global financial system... To this end we will continue to work closely with governments and regulators around the world.’ HSBC said it would be finalising an ‘undertaking’ with the UK regulator, ‘shortly’. However, some have expressed surprise at the lack of legal action by the UK regulator or by the Serious Fraud Office, given that the bank had been warned of the problem several years previously and had chosen not to take corrective action.

It should be recognised that the culture of an organisation is not fixed but continuously changes over time, in response to market forces, technological advances, regulation and
legislation, as well as external pressures from investors and other stakeholders. Indeed, in large dispersed organisations, culture may not be homogeneous throughout, as evidenced in global financial services groups having both investment banking and retail banking operations. Some have suggested that banking (with the obvious exception of casino banking) is a fundamental public service and as such should be subject to something akin to the ‘seven principles of public life’, these being:

1. selflessness;
2. integrity;
3. objectivity;
4. accountability;
5. openness;
6. honesty; and
7. leadership.

Holders of public office should promote and support these principles by example.

**Concentrating on customer service is not sufficient**

Stephen Hester, Chief Executive Officer of RBS stated, in October 2012: ‘Banks must undergo a wholesale change in their culture and refocus their behaviour on meeting the needs of customers to restore trust in the industry.’

Whilst a customer focus is essential in all industries, it is unlikely to be sufficient in banking. Changing the way banks do business may be difficult, if not impossible, since to do so would require a wider cultural change, focused around not only ‘doing what is right for customers, but also for shareholders, the community and employees’. With regard to casino banks, such as Goldman Sachs, the management team would need to become the last to be rewarded (not the first, as at present), with incentives being based upon longer-term risk-adjusted performance metrics. The company would need to change from ‘principle’ (gambling in its own right), to ‘agent’ (acting for and in the best interests of the client). Clearly, to act as both principle and agent engenders a potential conflict of interest. In April 2013, The UK Parliamentary Commission on Banking Standards published a letter from its Chairman, Andrew Tyrie MP, to the Chief Executive of the new Prudential Regulation Authority (PRA), Andrew Bailey, asking how the PRA intends to give effect to the Commission’s recommendations on proprietary trading, contained in its Third Report. Andrew Tyrie said: ‘Proprietary Trading carries large risks, both prudential and for banking standards. It is not a suitable activity for UK-headquartered banks but a blanket ban now is not the most effective approach to mitigating these risks... The PRA should be given the opportunity to try to bear down on proprietary trading. A Volcker-style ban may subsequently be needed.’ It should be recognised that even in a morally ethical and risk aware institution, like the UK’s NHS, a concentration on customer service is not in itself sufficient. The NHS has a culture that is predicated on the principle of ‘do no harm’, with patient-care being paramount. Even so, serious incidents do still happen as evidenced in the report that emanated from the recent Francis Inquiry.
Like banking, the NHS is faced with a very considerable challenge. Put simply, the NHS has been tasked with ‘doing more with less’; becoming more responsive in order to meet changing customer demands. Fundamentally, this change is being brought about by medical advances (including improvements in drugs and treatments), together with population changes (that is, an increasing and ageing population that is making poor lifestyle choices, such as, smoking, drink and drugs, obesity and lack of exercise).

Professor Bryan Stoten, Chair of the UK Public Health Register, has said: ‘It is no longer acceptable to simply drag people out of the river, downstream from where they fell in. It is necessary to better educate them and stop them falling in in the first place.’ In addition, people will need to accept greater personal responsibility and do more themselves, as is happening in the airline industry.

A silo-based approach is not appropriate. A local acute hospital cannot simply act as a stand-alone facility, offering short-term care (that is, emergency/accident care and medical operations). Instead, an integrated service (including social care) is required across the whole community. Whilst acute hospitals have never really been totally stand-alone, there is now a need for far greater integration, both upstream and downstream. People need to be seen earlier and to receive appropriate treatment, rather than being left languishing in acute hospitals due to failure in other elements of the care sector. Putting General Practitioners (GPs) in ambulances and treating people at home are potentially important developments, which increase customer focus and overall responsiveness.

Unfortunately, the NHS does not appear to have the level of skills and resources currently necessary to carry out the very considerable change-management task with which it has been charged; effectively, the game is getting far harder. Consequently, there is a need for greater integration between local NHS acute hospitals and private sector providers. It is likely the private sector will increasingly be required to add professionalism with regard to customer service, financial control and management skills, in addition to subsidising NHS services, in exchange for access to a sound business base upon which to build. Currently, the private sector tends to provide lucrative ex-ante ‘feel-good’ services such as fitness, weight-control and beauty services, whilst NHS acute hospitals provide essential post-ante services (which tend to be loss-making, due to both inefficiencies and the government’s current funding regime for the NHS).

Additionally, there is need for high-level specialist facilities placed strategically around the country, with ‘centres of excellence’ being more regionally based. A possible scenario is for university teaching hospitals to be linked nationally and possibly internationally, as well as servicing local acute hospitals.

As the NHS has demonstrated, the need is for a holistic customer focused vision. Similarly, banking should consider adopting an enterprise-wide view, centred upon acting in the capacity of agent, educating and providing value-added services to customers. The NHS exists to protect people’s health; banking should exist to protect their wealth.

Where risk management is poorly understood, standards may not be the answer

John McFall said: ‘During its first inquiry into the financial crisis in late 2007, the (UK) Treasury Committee was astounded to discover that managers and directors did
not understand the risks their banks were taking, nor the products in which they were investing.\textsuperscript{21}

In an address to the US Senate Financial Crisis Inquiry Committee, Warren Spector, co-COO of Bear Stearns, stated that in his opinion the company was well managed and that its risk management practices were exemplary, being robust and effective, and with the risk management team being highly trained and very experienced. He was, therefore, surprised when Bear Stearns failed.

Risk management remains a discipline that is poorly understood. Some say that despite the significant advances that have been made in risk management, it is still impossible to predict the future. Certainly, this is correct and will always remain so. However, the statement demonstrates a fundamental misunderstanding. Risk management is not about predicting the future but being prepared for it. It is a popular misconception that risk management enables the future to be predicted and that better models give better extrapolated predictions. This misinformed belief has been reinforced, in part, by regulators requiring banks to hold capital against ‘once in a thousand-years events’. Extrapolation is merely guesswork.

Risk management recognises that there is a range of options for the future, all of which are valid, though with differing chances of occurrence. The fundamental aim of risk management is simply to increase the chance of achieving predetermined objectives – however that may be achieved.

The following questions, which are frequently raised, show some areas of misunderstanding.

- \textit{Is risk real or perceived?} Risk implies a range of possible future outcomes, all having different probabilities of occurrence. Like Schrödinger’s cat, reality collapses to a singular outcome upon observation. However, similar future events may have different outcomes. Hence, simplistically, the past may not be a good indicator of the future.
- \textit{Can risk be positive or must it always be negative?} This question correctly recognises that there is a range of possible outcomes. However, the extent to which an outcome is positive or negative depends upon the value chosen for the objective, against which the outcome is measured.
- \textit{Is risk an event, a cause, or an effect?} Risk is the probability of not achieving the required objective. Consequently, it is neither an event, nor a cause, nor an effect. However, it is related to these factors. An event can have multiple causes and give rise to multiple effects, which can impact upon set objectives.
- \textit{Is it always necessary to quantify risk?} Not all risks can be quantified. Where quantification is possible it can increase understanding, although the final answer may be less important than the analysis itself.
- \textit{Is it necessary to link risk to objectives?} Risk does not exist if there is no threat to pre-determined objectives (that is, a probability, however small, that the objective may not be met). This is recognised by ISO 31000, whose definition is widely, though not universally, accepted.
- \textit{Is risk another term for uncertainty?} Complexity generates uncertainty, not risk. Uncertainty requires complex problems to be unbundled, that is, simplified. Clearly, trying to fully define all future states of the world, and probability-weight them, is beyond cognitive limits. A key factor determining uncertainty is the size of the sample from which a
The future of operational risk management in banking

model is determined. A model seeking to infer behaviour from the past, based on too short a sample, may experience ‘over-fitting’. Similarly, probabilistic weights from the past may be a poor guide to the future. Consequently, noise could be mistaken for a real signal. A model which is ‘over-fitted’ may be unstable and yield rather fragile outputs. In general, where model uncertainty is greater, the performance of simple heuristics tends to be better because simple rules are more robust to a lack of information. Clearly, when carrying out data analysis, if part of the distribution is ignored, distortion will occur. In contrast to uncertainty, risk management must consider all outcomes that could potentially occur from a particular distribution, since risk management is about being prepared for the future.

• What is meant by ‘operational risk?’ Operational risk is the name given to a category of risks fundamentally concerned with management. The term tends to be used mainly in the banking and financial services sector and adds little to the understanding of the discipline of risk management. Originally, operational risk was said to be all risks other than market risk and credit risk (that is, what it is not, rather than what it is). Basel provided the now generally accepted positive definition, on the basis that ‘what cannot be defined cannot be measured, but what gets measured gets managed’. Whatever the definition chosen for operational risk, it is important to remember that the business must deal with all risks and should adopt a holistic view.

• What is meant by ‘risk appetite?’ There is no agreed definition of risk appetite and indeed it may mean different things to different people. However, it is often taken to mean the amount or type of risk an organisation (or individual) is willing to take (or accept), given its particular risk management skills and expertise, in order to achieve its set objectives. A simple example is: ‘should we be in this business, taking into account the possible threats and volatility it brings, given our level of knowledge, skills and resources? In other words, are the rewards sufficient in our particular case?’ Risk appetite is concerned with strategy, that is, determining the risk appetite for each of the various business units comprising the organisation, with a view to fully embedding the strategy. A definition of risk appetite was given by the Basel Committee in its consultative document: ‘Risk management data aggregation and risk reporting.’ COSO has also given a definition, which differs slightly from Basel.

There are an infinite number of risks, which theoretically, can be considered to emanate from three fundamental types: (i) people and management risks; (ii) systems and processes risks; and (iii) external source risks. It is the interaction of these that leads to complexity and the possibility of misunderstanding.

• President Mikhail Gorbachev blamed managers and employees for failures, whereas in fact the system was at fault.

• Edwards-Deming (one of the founding fathers of Total Quality Management) stated that in the vast majority of cases 90% or more problems were down to the systems used rather than the workers; simply exhorting people to do better was misguided.

• Professor Jim Reason recognised that a rigid concentration on inherent systems weaknesses may obscure important active errors.

www.rasabourse.com
Operations and IT risk

- External risks lead to economic drift. As a consequence, what is considered ‘best practice’ at the moment will decay with time.
- Research into ‘High reliability organisations’ has shown that team work in a mindful learning environment can result in significant improvements in performance. This is supported by the early Hawthorne Experiments.

The ISO 31000 Risk Management standard could be regarded as current best practice. To regard it as anything more may prove dangerous. ISO 31000 is in fact a set of voluntary guidelines whose purpose is to assist in the management of risk (with a view to encouraging performance improvement and flexibility) rather than a rigid compliance reference standard orientated towards certification. It is a cause of concern that standards orientated towards certification may eventually (if not immediately) lead to ossification, thus preventing intellectual development of the discipline of risk management. Unfortunately, standards tend to promote singularity with regard to the means through which outcomes sought by defined objectives are secured. Stacey emphasised the importance of adaptability and flexibility ‘in order to roll with the punches’. Ansoff and McDonnell stated that hardwiring of business senses and processes may jeopardise an organisation’s flexibility and adaptability. In promoting standards, such as ISO 31000, consideration should be given to the possibility of unintended consequences.

Risk managers prepare for the future; gamblers predict it

Risk management is concerned with preparing for all possibilities because in the fullness of time (that is, as T tends to infinity) all possibilities will occur. In contrast, gambling is concerned with guessing the order in which these events will occur. It is important that risk managers appreciate the difference and properly understand the principles behind gambling. A simple mathematical model can show that not only are the chances of winning a game (over a sufficiently long period) influenced by the probability for each event but also by the size of each bet and by the initial capital sum available, as well as by any inflow of funds from sources such as profitability. Put simply, although the probability of winning may seem favourable, it is still possible to lose the game (that is, win the battle but lose the war). This goes to the heart of the difficulty in properly interpreting the ratings awarded by risk rating agencies. As a very simple example, consider the possibility of tossing a coin many times. The chance of guessing correctly every time can be determined from Pascal’s Triangle. Deming demonstrated the problem with misunderstanding chance, in a simple experiment in which he praised those who guessed correctly and demoted those who guessed incorrectly. In such a system, which is often found in practice, it is easy to promote someone beyond their true capability, or award a company a rating which is not justified by its true ability to understand and manage risk.

Qualification has limitations

In 2009, the FSA's Turner review entitled ‘a regulatory response to the global banking crisis’, specifically identified two serious faults in risk management: (i) an inappropriate
focus on detailed risk calculations; and (ii) the misguided application of unwarranted faith in statistical risk models.

Risk management is born of a desire to understand and control complexity; to turn complexity and uncertainty into order. Quantification is an expression of this drive for order. Statistics is realisation of imprecision and variability. Quantification has the lure of reality; the measure of something real. However, mathematical techniques such as ‘capital adequacy modelling’ and ‘value-at-risk’ have long been criticised by statisticians for being fundamentally flawed in their application and interpretation.

A person’s view of risk and its management will be influenced by where they are at a point in time; by their past experience; and by their bounded views of the future. An ambiguity of beliefs can exist within an individual and between individuals. The young are beguiled by the beauty of rationality. The more senior are beguiled by the beauty of irrationality. Sir Mervyn King commented that while in his early career he believed strongly in his models. As a more seasoned banker he became less sure of the future and of his ability to predict and control it.

In 1814, Laplace published ‘the first articulation of causal or scientific determinism’, more commonly referred to as Laplace’s demon – a problem of such complexity as to be impossible to solve: ‘We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.’ Clearly, the impact of irrational human endeavour makes this impossible task ever more complex.

In 2009, John McFall, talking about the financial crisis, stated: ‘The crisis has brought to light many more examples of flaws in our systems of managing risk. For example, the Treasury Committee has questioned credit rating agencies on the conflicts of interest inherent in their business model, and their overarching reliance on quantitative models – even relying on models devised by the banks themselves.’

It should be recognised that management and regulators may need to use quantification differently. In addition to taking a strategic, enterprise-wide view of risk, management must also take a more detailed forensic view, possibly involving root cause analysis and reliability engineering techniques. Effectively, the overall loss distribution curve needs to be broken down into its constituent elements and methods for improving the probability distribution for each element considered. Fundamentally, management is concerned with causality. It is less concerned with determining capital adequacy through improving the fit of the overall loss distribution curve; this is more a current regulatory concern.

Regulators are beginning to question the appropriateness of the 99.9% confidence level for capital adequacy. The Bank of England has spoken about ‘decision-making under uncertainty’. This is beginning to attract increasing interest because of its ability to yield powerful insights in comparison to more traditional detailed quantification techniques, which tend to give a somewhat blurred picture. Rather than taking a detailed
forensic view of risk, the Bank of England has suggested that, based on the work of researchers, a few simple well-chosen heuristics may work better than risk-based approaches, in a complex environment. In 2012, Haldane and Madouros stated: ‘Take decision-making in a complex environment. With risk and rational expectations, the optimal response to complexity is typically a fully state-contingent rule. Under risk, policy should respond to every raindrop; it is fine-tuned. Under uncertainty, that logic is reversed. Complex environments often instead call for simple decision rules. That is because these rules are more robust to ignorance. Under uncertainty, policy may only respond to every thunderstorm; it is coarse-tuned… Other things being equal, the smaller the sample, the greater the model uncertainty and the better the performance of simple, heuristic strategies.’ The danger lies in assuming through a process of induction, that what is right for one bank is right for another, thus ignoring the risk profile and performance shaping factors of an individual bank. It should also be recognised that high volumes of short-term volatile data can mask the overall longer-term trend. Young and Coleman have suggested that fundamental analysis requires triangulation through multiple lenses of perception; whilst weaknesses may be identified, time to failure may prove more difficult to determine (that is, the rating agency dilemma).

### Risk management works well in other industries

Whilst risk management in the banking sector has advanced considerably in recent years, it remains some way behind other sectors, such as aerospace, the nuclear industry and the health service. Risk management in these sectors is fully integrated and effectively forms part of their structural DNA; such a concentration on risk management has enabled them to achieve high reliability. Weick and Sutcliffe said: ‘The signature of a high reliability organisation (HRO) is not that it is error-free, but that errors do not disable it. To learn from errors (as opposed to avoiding errors altogether) and to implement that learning through fast negative feedback, which dampens oscillations, are at the forefront of operating resiliently.’

It is interesting to contrast high reliability organisational behaviour with that of financial services organisations employing the more ubiquitous silo-based command and control approach. In ‘silo’ organisations, employees are required to adhere strictly to pre-determined rules and regulations. As such, they operate within their own limited sphere of responsibility, without taking into account the broader impact of their actions (that is, they lack ‘mindfulness’). For example, within banking, conflict is often found to be present between the risk management function and the compliance function. This can result in exacerbation of problems and, indeed, Hopkins has identified the existence of a silo-based culture as being associated with many organisational accidents. In contrast, within the aviation industry there is a legal duty (at least in certain countries) for a subordinate to challenge decisions made by a superior. In banking, subordinates fear the possibility of losing their job if they challenge their superiors.

Whilst HROs are recognised as exemplars for others to follow, this does not mean that they adopt a formulaic standardised approach. Instead, their approach is attuned continuously to support their objectives.
Aerospace

NASA is often cited as an example of an organisation that, after having experienced disaster, was transformed into a high reliability organisation. The US Presidential Enquiry into the 2003 Space Shuttle Columbia catastrophe, focused attention on the attitude of NASA management. The Columbia Accident Investigation Board (CAIB) concluded that NASA had failed to learn many of the lessons of the 1986 Challenger disaster and, indeed, the catastrophe resulted from the same ‘flawed decision making process’. Although significant changes were made after Challenger, unfortunately, the changes in organisational culture and management structure were neither sufficiently deep nor long-lasting. ‘NASA’s culture of bureaucratic accountability emphasised chain of command procedure, following the rules and going by the book. While rules and procedures were essential for co-ordination, they had an unintended but negative effect. Allegiance to hierarchy and procedure had replaced deference to NASA engineers’ technical expertise... Over time, a pattern of ineffective communication has resulted, leaving risks improperly defined, problems unreported and concerns unexpressed... Leaders create culture – it is their responsibility to change it.’

Many organisations adopt a culture which acts to suppress communication and action. Certainly, this was the case in the early days of NASA and is currently evident in the banking and financial services sector.

Nuclear

It is instructive to consider the US Three Mile Island (TMI) nuclear power station event, which occurred in 1979 and effectively forced all US nuclear power stations to become HROs. Numerous warnings from (front-line) engineers over a sustained period were ignored until the risks that had been identified, eventually occurred (something probability theory states will happen, given sufficient time). Since failures in the nuclear industry can be cataclysmic, safety and reliability are taken very seriously. However, disasters can and do still happen, as indicated by other events such as Chernobyl in 1986 and Fukushima in 2011 (not to mention the potential legacy threat from the gradually decaying and abandoned former Soviet nuclear submarine fleet). Clearly, complacency must be avoided. From a technical perspective, the design of nuclear systems incorporates fail-safe principles and parallel-systems efficiencies, so that the reliability of the whole system is greater than the component parts (that is, system integrity does not rely upon the weakest link). In addition, the system is supported by: (i) planned maintenance; and (ii) unscheduled essential maintenance. Although, technically, the system may be sound, it is at the socio-technical interface that risks are greatest. A nuclear reactor is most vulnerable when moving from one state to another, particularly when human interaction or intervention is required. People are not naturally diligent 100% of the time, particularly, over long periods of time. Indeed, it has been suggested that human failure is nature’s way of continuously ensuring change and responsiveness.

Probabilistic risk analysis (PRA), underpinned by human reliability analysis (HRA), is used to assess the likelihood of failure for the various event/fault tree pathways. Human reliability analysis techniques, such as SPAR-H, can be used to quantify categories of human errors, taking account of performance shaping factors (PSFs) and domain expert opinions.
Importance analysis (IA) techniques, such as Fussell-Vesely, can be applied in order to identify the most significant human actions.

Typically, quantification involves finding the probability of action failure, then modifying that probability by a set of PSFs. Expert judgement may be necessary, initially, to obtain probability estimates. In this methodology, the target event is treated in isolation. Hollnagel\textsuperscript{42} suggested an alternative approach, beginning with consideration of the context and the common performance conditions. Each target event is then analysed taking account of the context. Consequently, the list of possible causes of failure contains those that are likely, given the conditions prevailing.

In general, it is considered that quantification is inappropriate without determination of a substantive qualitative base, which identifies possible dependencies between actions. Whilst quantitative performance prediction is the sine qua non for HRA in PRA, some behavioural scientists consider quantification to be impossible and question its validity, particularly, when critical tasks can be identified and failure modes determined.

\textbf{NHS}

The UK NHS is highly regarded and considered an exemplar institution. Indeed, it enjoys exalted status, being regarded as a source of national pride, providing universal care, free of charge, throughout life. However, this halcyon perception may have led to complacency and a lack of responsiveness throughout. The Francis Report,\textsuperscript{43} published in February 2013, strongly criticised the NHS at all levels and put forward approximately 290 recommendations for improvement, with a view to changing culture and putting patients first. The core message of the Francis Report was that within parts of the NHS, its culture had degenerated to such an extent that there was a lack of care, compassion, dignity and humanity shown towards patients.

Fundamentally, the public inquiry into failings at Mid-Staffordshire NHS Foundation Trust sought to address three issues.

1. How could such poor care have been allowed to happen by local doctors, nurses and managers?
2. Why were the voices of patients and families ignored?
3. Why did it take so long for regional and national NHS bodies to understand the hospital’s failure and take necessary action?

Francis said that criticisms lay at all levels. Whilst government ministers determine overall policy, the Treasury has ultimate covert control, placing financial constraints on the Department of Health. Within the NHS, this resulted in a focus on achieving central targets, a preoccupation with financial performance, and a drive to attain foundation trust (FT) status. Francis said that whilst the Department of Health was not culpable of any direct responsibility for the failings at Mid-Staffordshire NHS Foundation Trust, the evidence he saw revealed ‘remoteness’ from the reality of front-line staff and patient experience. Department of Health directives to hospitals and the emphasis on policy implementation had sometimes been perceived as ‘bullying or applied oppressively... The
The future of operational risk management in banking

Department of Health did not ensure that Ministers were given the full picture.’ The independent regulator of NHS Foundation Trusts, Monitor, ‘focused on corporate governance and financial control’ without properly considering patient safety and care. Standards were inappropriate. ‘The Healthcare Commission was required to assess Trusts against standards, which did not adequately assess the quality of care being given to patients... There was a reluctance to intervene urgently by those who had the powers to do so... The Trust Board was weak. It did not listen sufficiently, to patients and staff or ensure correction of deficiencies brought to its attention. It did not tackle the tolerance of poor standards or the disengagement of senior clinical staff from managerial and leadership responsibility.’ Francis also commented that ‘agencies frequently failed to share their knowledge with each other. Assumptions were made that important functions were being performed satisfactorily by others... The danger of loss of corporate memory from major reorganisations was inadequately addressed.’ The sheer complexity of organisational and supervisory arrangements in the NHS made it nigh on impossible to know who was responsible, how patients were supposed to complain and who should have taken action when failings came to light.

In summary, there are six themes, underlying Francis’s recommendations, which he considers are necessary to bring about cultural change of the sort he envisions.

1 Standards: ‘A structure of clearly understood fundamental standards’ is required.
2 Transparency: there is need for ‘openness, transparency and candour throughout the system... A duty of candour should be imposed and underpinned by statute and the deliberate obstruction of this duty should be made a criminal offence.’
3 Education and training: ‘Training standards need to be created to ensure that qualified nurses are competent to deliver compassionate care.’
4 Leadership: ‘Strong and patient-centred healthcare leadership’ is required. Leadership in the NHS is ‘under challenge’ but he warned the culture he wished to instil would ‘only flourish if leaders reinforce it every day.’ Accountability and a ‘zero tolerance’ are essential. He recommended an ‘NHS leadership staff college’, and a ‘common code of ethics and conduct’.
5 Information and communication: ‘Accurate, useful and relevant information’ is essential. ‘Patients, the public, employers, commissioners and regulators need access to accurate, comparable and timely information’, in order to facilitate benchmarking and ranking with a view to seeing if standards were improving or falling. Francis said it should be a criminal offence to give false statements regarding compliance with standards.
6 Compliance: non-compliance ‘cannot be tolerated,’ he said, ‘any organisation unable consistently to comply should be prevented from continuing’.

Research by Milgram has shown that in highly institutionalised organisations, as found in the NHS (and in banking), people tend to do what they are told by their superiors, with normal moral behaviour being suspended. Similarly, where employees are instructed to look at particular indicators, they often become ignorant of other more important risk factors (that is, the invisible gorilla in the room). Consequently, taking account of the these factors, together with those identified by Francis (that is, a lack of responsiveness to change, poor
leadership and communications, a lack of mindfulness, inaction and a lack of customer focus), failure at some point in time was to be expected. Informed action must be taken to reduce the frequency and severity of failures.

**High reliability – its relevance in banking**

Currently, within the financial services industry, a rigid hierarchical command and control structure is often encountered. Weick and Sutcliffe stated:47 ‘people who refuse to speak up out of fear enact a system that knows less than it needs to know to remain effective. People in HROs know that you cannot develop a big picture of operations if the symptoms of those operations are withheld.’ The collapse of Barings Bank in 1995 and the Allied Irish Bank loss-event in 2002, were prime examples of senior management, who were remote and had lost touch, becoming unaware of the risks their banks were actually running.

A major tenet of the high reliability concept is that high levels of responsiveness require a learning culture and environment in order to sustain good communications and ensure full commitment by all. HROs deliberately seek to establish a situational-based big-picture approach in order to avoid surprises that come with simplification. In contrast to organisations that adopt a traditional cost-minimising approach, HROs employ more people to explore and evaluate complexity. Lawson defined ‘the pool of resources in an organisation that is in excess of the minimum necessary to produce a given level of organisational output’ as ‘organisational slack ... Learning organisations require slack in the form of time to develop – and time for learning must be part of the organisation’s design’.48

Put simply, responsiveness improves resilience. Responsiveness requires that, when appropriate, decisions migrate to those with the greatest knowledge and expertise, in direct situational contact; a return to the normal locus of hierarchy occurring only after criticality. Hence, employees must be competent and continuously engaged if they are to remain situationally aware and responsive. However, some have criticised the idea of empowerment on the basis that it can lead to unintended consequences through an inability of those on the front-line to see the bigger picture. Marias et al. stated:49 ‘The type of bottom-up decentralised decision making advocated for HROs can lead to major accidents in complex socio-technical systems.’ When incorrect actions are taken that increase risk but do not immediately lead to loss or failure, these actions can become accepted as the new norm. Whilst such criticisms are certainly worthy of consideration, they merely emphasise the need to properly address issues concerning levels of responsibility and authority in stress situations (as with business continuity management (BCM)) and the need to ensure an appropriate level of professional competence.

Research into HROs began in the 1980s and concentrated on factors of success. Those organisations that operated in hazardous conditions but which despite this, had operated for long periods without signs of failure were examined. Prior to this, research had concentrated on failure, that is, organisations that had experienced disaster were examined and the reasons for the disasters sought. Clearly, both avenues of research are necessary, however, in practice, it should be remembered that there are an infinite number of ways in which an organisation can fail; success has fewer determinants.
Weick and Sutcliffe\textsuperscript{50} concluded that a state of ‘mindfulness’ was needed to fully achieve high reliability. A collective state of mindfulness is said to exist where the following five characteristics are present.

1. A preoccupation with responsiveness and failure prevention.
2. A commitment to resilience.
3. A reluctance to simplify.
4. Sensitivity to operations (that is, situational awareness).
5. Deference to expertise.

According to Weick and Sutcliffe:\textsuperscript{51} ‘The key difference between HROs and other organisations in managing the unexpected often occurs at the earliest stages, when the unexpected may emit only weak signals of trouble. Mindfulness preserves the capability to see the significant meaning of weak signals and to give strong responses to weak signals.’ Discarding information in pursuit of simplification is regarded as inherently dangerous. ‘Simplification increases the likelihood of eventual surprise.’\textsuperscript{52}

The importance of an accumulation of errors and weaknesses, including weak signals is also recognised in the accident-causation model (commonly referred to as the ‘Swiss Cheese’ model). This has effectively become the de-facto standard model for accident (failure) investigation, having been endorsed by organisations such as the International Civil Aviation Organisation and accepted by many other industries where high reliability is considered essential, including the chemical and energy industries. A rigid concentration on the identification of latent conditions can obscure the importance of currently active errors, which may prove to be dominant factors in their own right and not merely symptoms of inherent systems weaknesses. In recognition of this, Reason stated:\textsuperscript{53} ‘the pendulum may have swung too far in our present attempts to track down possible errors and accident contributions that are widely separated in both time and place from the events themselves.’ Young and Coleman stated:\textsuperscript{54} ‘A more balanced approach involves consideration of both the current active errors and the latent organisational conditions (which are always present but remain dormant until triggered). It should be recognised that active errors might be the result of a defective system; however, active errors may also be independent of and more important than inherent latent factors. Often, the causal relationship between active errors and distant latent conditions can prove tenuous, with many latent conditions appearing inconsequential in a pre-event scenario. From a management perspective, some latent conditions may be very difficult to control and may require considerable time and expense to ameliorate.’ Young and Coleman also recognised the importance of random events (black swans)\textsuperscript{55} and other external events outside the control of management, which require the application of (BCM) techniques.

Perrow’s normal accident theory (NAT)\textsuperscript{56} propounds that accidents are inevitable in complex, tightly-coupled technological systems, since complexity inevitably yields unexpected interactions. Similarly, Deming\textsuperscript{57} stated that: ‘Improvement in performance comes about by informed change to the system. Errors are a function of the system; simply exhorting people to do better can be mis-guided.’ However, the concept of ‘high reliability’, although recognising systems limitations, also recognises that in socio-technical systems, output is variable.
and that the actions of people can significantly influence performance; indeed, mindfulness can reduce failure.

HROs are stable organisations with a low risk of failure, capable of continuing to operate effectively in times of stress (that is, they are ‘mindful’ organisations distinguished by high levels of resilience and responsiveness). They are characterised by exemplary leadership and culture, clear customer centric objectives and recognition of factors influencing the future. Certainly, banking has some way to go before achieving such standards of risk management.

Conclusion
This chapter provides a discussion of the key factors influencing the future of operational risk management in the banking sector. Whilst the need for change may be irrefutable, complexity determines that there is no one simple solution. In addition, Machiavellian vested interests will inevitably influence the type, rate and effectiveness of any future reforms.

The key findings of this chapter are as follows.

• **Risk is poorly understood; regulators have an important role to play in enhancing knowledge:** it is said that risk management in the banking sector is failing. However, it is wrong to criticise risk management for its inappropriate application. It is also a mistake to confuse the tools of risk management (their quality and capability) with the ability of users. Similarly, it is incorrect to criticise risk management for failures in regulation, although, it is valid to criticise intellectually weak regulation and legislation. To simply quote the regulatory mantra, ‘lessons have been learned’, in the aftermath of risk events, is also wrong; merely looking at great works of art will not make the observer a great artist.

  Regulators have an important role to play in continuously driving forward the boundaries of knowledge and in ensuring that this knowledge is properly embedded. Whilst risk experts and behavioural scientists are already well aware of the lessons to be learned from examples such as the Francis Inquiry, the challenge remains to properly embed this knowledge, given ever-changing priorities and the inevitable bounded skill-levels of those more practically engaged in service delivery. Where an organisation fails to achieve its set objectives and adhere to its risk appetite, it is necessary to evaluate the risk lines of defence (management, control and assurance) and to ask the basic question: ‘is risk management failing or is it just poorly understood’.

• **The restructuring of banking is essential:** the apparent failure of risk management in the banking and financial services sector has led to a general lack of confidence and a concentration on short-termism. This has resulted in a somewhat legalistic, moribund, system of regulation. The Independent Commission on Banking and the Kay Review have identified the need for restructuring in order to enhance efficiency and effectiveness, as well as to improve stability and competition. However, restructuring requires a clear and shared vision of the future of banking and its fundamental purpose, something that remains questionable.

• **Banking must clearly serve a public good:** it should be recognised that the first priority of a bank must be to run its business correctly, properly serving its customers and acknowledging that what was acceptable in the past may not be so in the future. Only
then should banks seek to deal with outstanding issues pertaining to regulatory compliance. The regulators need to appreciate the importance of this prioritisation (and possibly reconsider their own objectives), with a view to improving the performance of banking and its contribution to society. However, this will require a move on the part of banks to act morally and ethically; something that external regulation cannot enforce easily, if at all. As stated previously, the NHS protects people’s health; banking should protect their wealth.

• Cultural change will impact risk: it is widely accepted that cultural change is essential. Indeed, it has been suggested that cultural change is more important than change to the regulatory and legislative structure. However, determining the longer-term relative ‘richness’ of culture may prove challenging. It has been suggested that since banking should serve a public good, adoption of the seven principles of public life could be appropriate, providing a sound base upon which to build.

• Distrust is bringing about fundamental and irreversible structural changes in banking: the frequency and size of failures in the financial services sector has resulted in profound distrust. This is beginning to bring about structural changes, with concomitant risks. People are becoming more reluctant to save (for example, for pensions or emergencies, and so on). Indeed, this has, in part, given rise to the emergence of neo-usury pay-day loan companies. Change is also evident through the emergence of ‘peer-to-peer’ banking, which poses a growing threat to the traditional banking model, since it does not require any third-party banking intermediary. Instead, borrowers and lenders trade directly, having only the portal provider and rating agency fees to pay. Similarly, decentralised virtual digital crypto-currencies (for example, Bitcoin and Namecoin), which are not accountable to any ‘central bank’ issuing authority, are beginning to gain credence, boosted partly by the Cypriot crisis that posed a threat to money as a ‘store of value’.

• Banks must learn from other industries in order to advance and achieve ‘high reliability’ status: banks have the opportunity of advancing and becoming ‘high-reliability’ organisations, through learning from other industries, such as the aerospace sector, the nuclear industry and the health service, as well as from new competitors in their own sector. However, to do so will require an open-minded approach. To quote Leo Tolstoy: ‘The most difficult subjects can be explained to the most slow-witted man if he has not formed any idea of them already; but the simplest thing cannot be made clear to the most intelligent man if he is firmly persuaded that he knows already, without a shadow of doubt, what is laid before him.’

• High reliability will not be achieved through quantification and modelling: the focus of risk management in the banking sector, particularly with regard to regulation, will need to change, from reliance upon detailed models and quantification of capital adequacy, important as these are, towards assurance centred upon risk appetite and the attainment of ‘high reliability’. Whilst a move from risk minutia to broad heuristic indicators may be appropriate for regulators, bank management need to take a more holistic approach, considering detailed causality and correlation whilst also taking account of performance shaping factors.

• Inappropriate and voluminous annual reporting masks clarity and understanding: with regard to public reporting, there is a need to place emphasis on assurance through externality. The imprimatur of the rating agencies on annual published accounts would
(eventually, if not immediately) add greatly to relevance and confidence. Reporting needs to take a longer-term, holistic, perspective. However, in order to reduce the current voluminous nature of reporting and to aid analysis and understanding, reports and sub-reports should be linked through greater use of web-reporting.

• **An international perspective needs to be adopted towards standards:** in the past, banking (both individually and collectively) has used its powers to limit the restrictions imposed by regulators; with the regulators frequently lacking support from politicians keen on actively promoting UK plc. Change will require sustained moral indignation and firm political will together with appropriately co-ordinated regulatory and legislative action. John McFall stated, in 2009: ‘The financial crisis has demonstrated that regulation needs to be co-ordinated at the international level. Recently, we have seen authorities in some countries taking a unilateral approach which has undermined stability in the UK. We have also realised that the UK may be host to large branches of foreign banks, which are not regulated by British authorities but which have a very large impact on our financial stability. There is a clear need for an international approach to many of these issues.’ The establishment of a European-wide regulator is an important step in this direction, providing it does not take on a life of its own and result in bureaucratic ossification.

**Bibliography**

*Economist*, ‘Over-regulated America, the home of laissez-faire is being suffocated by excessive and badly written regulation’, 18 February 2012.
Hester, S, ‘Banks must undergo a wholesale change in their culture and refocus their behaviour on meeting the needs of customers to restore trust in the industry’, Reuters, 1 October 2012,
HSBC, ‘How Colombian drug traffickers used HSBC to launder money’, Reuters, 1 January 2013.

Risk appetite definitions

Basel: www.bis.org/publ/bcbs222.pdf.

1 Young, B and Coleman, R, Operational Risk Assessment: the commercial imperative of a more forensic and transparent approach, ch 1, 2009, John Wiley & Sons.
2 Accenture, The point: banking 2012.
10 Johnson, B, ‘You can’t blame finance directors for doing their job’, comment made on Andrew Marr Show (BBC1), 16 December 2012.
12 Stoll, C and Schubert, G, actual quote is: ‘Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom’, in Keeler, MR, Nothing to Hide, 2006, p. 112, iUniverse.
19 Hester, S, ‘Banks must undergo a wholesale change in their culture and refocus their behaviour on meeting the needs of customers to restore trust in the industry’, Reuters, Monday 1 October 2012 9:59 p.m. BST; report by Scuffham, M, edited by Merriman, J and d’Afflilso, T. Available at http://uk.reuters.com/article/2012/10/01/uk-rbs-hester-idUKBRE89013I20121001.
40 Columbia Accident Investigation Report, August 2003, chapters 7 and 8.
41 Boring, RL, ‘How many performance shaping factors are necessary for human reliability analysis?’ Idaho National Laboratory, June 2010.
The future of operational risk management in banking

45 Milgram, S, ‘Experiment on obedience to authority figures’, 1962, video available at: www.youtube.com/watch?v=W147ybOdgpE.
Chapter 4

Keeping markets safe in a high speed trading environment

Carol L Clark

Federal Reserve Bank of Chicago

Overview

This chapter discusses a study the Federal Reserve Bank of Chicago conducted in 2010 and 2011 to better understand the controls that are in place to manage the risks of high speed trading (HST). The many factors that have profoundly altered the microstructure of US markets over the past few decades and hastened the adoption of HST are examined, including a number of technology glitches that have brought this style of trading to the forefront of media attention. Losses arising from HST have the potential to impact the financial conditions of trading companies and, depending on the magnitude, clearing broker-dealers (BDs) and futures commission merchants (FCMs) and possibly clearing organisations. Determining and applying the appropriate balance of financial and operational controls is, therefore, crucial. The study found that at every level of the trade life cycle, risk controls vary and tensions related to responsibilities for setting and administering these controls exist. Of critical importance to risk management is the time frame in which trading venues provide post-trade information to trading companies, broker-dealers and futures commission merchants and clearing organisations, as well as the speed at which those entities incorporate these data into their risk monitoring systems. One surprising finding from the study is that market participants would like regulators to work with the industry to establish best practices for risk management.

Introduction

Many factors have contributed to dramatic changes in the microstructure of US financial markets in recent years that have altered trading behaviours around the globe. Some of these influences include competition, regulatory initiatives and technological and mathematical advances. Taken together, these developments have hastened the migration from floor-based trading to screen-based (point-and-click) trading to HST.

HST includes high frequency trading (HFT) where companies are generally in and out of positions during very short durations of time and end the day flat with little or no exposures to the market, as well as automated and algorithmic trading where positions may be held onto for longer time horizons. HFT, automated and algorithmic trading are grouped together in the definition of HST because each of these styles of trading has the potential to
send thousands of orders per second to a trading venue\textsuperscript{2} that could result in knock-on effects in the market. This is due to the rate at which unintended, large positions can be amassed and out-of-control algorithms and/or erroneous trades can occur. It is important to note, however, that it is difficult to quantify the precise number of orders that would designate a company as being engaged in HST. As an obvious example, an algorithmic trader could execute 100 trades over the course of a day, which would not be considered HST.

There is also the possibility HST may result in negative feedback loops that are caused by a runaway algorithm triggering other algorithms or by numerous HST companies utilising comparable trading models that do not accurately assess and respond to changing market conditions. For example, some companies’ trading models may have been tested using historic data and markets may move in directions that were not anticipated by that data.

At the same time, HST has also brought numerous benefits to the marketplace. In floor-based trading, trade confirmation, allocations, price reporting and risk management were not simultaneous and some of these functions occurred at the end of the day. In contrast, HST can, depending on the frequency with which trading venues provide post-trade data and companies incorporate that data into their risk management systems, result in an immediate, electronic audit trail of transactions. HST also provides companies with automated hedging capabilities and removes human emotions that might otherwise enter the trading process. Investors, too, have benefitted from HST as supported by a number of studies that indicate one or more of the following: (i) the narrowing of bid-offer spreads (the difference between the price at which a buyer is willing to purchase a financial instrument and the price at which a seller is willing to sell it); (ii) the reduction of short term volatility; and (iii) the increase in market liquidity.\textsuperscript{3,4,5,6}

Some may also argue that the implementation of computerised trading strategies and processes is part of a natural, evolutionary outcome within financial markets not unlike the adoption of automated or computer-based systems in every other major industry. Nevertheless, one may also observe that examples of computerised processes in other industries resulting in a company rapidly being brought to the brink of bankruptcy, such as the 45 minutes it took Knight Capital to lose US$440 million, do not readily spring to mind.\textsuperscript{7}

However, it is a given that technological advances are here to stay, that HST is increasing in most markets (Exhibit 4.1), and that reverting to traditional, floor-based trading is not an option. The question then is how do you keep markets safe in a highly complex technological environment?

Until the 1990s, trading in organised financial markets in the US occurred almost exclusively in a physical, floor-based environment. Orders were called into BDs or FCMs by telephone or some other electronic means, transcribed by humans, and passed on to others for execution, confirmation and settlement. During each step, a person would look at the order and, knowing where the market was trading, make a value judgment as to whether a clerical error might have been made. Certainly, errors still occurred, but ample opportunities existed to detect and correct them. HST trading requires a similar level of monitoring, but it needs to happen a lot faster.

As such, HST requires automated pre and post-trade risk controls at every step in the life cycle of a trade with human beings overseeing the process. Consequently, trading companies, trading venues, BDs/FCMs and clearing organisations have an essential role to
play in preventing, detecting and responding to unusually large and unintended positions, out-of-control algorithms and/or erroneous orders.

This chapter will discuss: (i) the confluence of events that has profoundly altered the microstructure of financial markets; (ii) some technology glitches that have brought HST to the forefront of media attention; (iii) the lifecycle of a trade; and (iv) the key findings from a HST study conducted by the Federal Reserve Bank of Chicago Financial Markets Group (FMG) and recommendations from that study.

**Impact of technology, regulation and competition on market microstructure**

For more than a decade, various market, regulatory and technological events have radically changed the microstructure of financial markets. At the same time, trading companies, BDs/FCMs and trading venues are all operating in an environment of increased competition, declining volumes in many markets and reduced bid-ask spreads. Taken together, competitive pressures and regulatory and technological changes have altered the way companies at every level of the trade life cycle manage their business.

One notable technological change in recent decades is the deployment of computer algorithms in the trading process. Long used in US equities markets to control the timing, size and price of orders, algorithms are also used to route orders to the trading venue with the best execution price in compliance with Securities Exchange Commission (SEC) regulation.
Equities trading strategies must also take into account the complex and highly fragmented nature of US markets, where liquidity is unevenly spread between numerous public exchanges and non-public trading venues like dark pools and BD internalisers. Dark pools minimise the risk of large orders moving the market price by allowing buyers and sellers to submit orders anonymously and report price and size information only after the trade has been completed. BD internalisers match customer orders against each other or against the BD’s own proprietary orders. It is currently estimated that about one third of all trades in US equities markets are executed away from public exchanges in dark pools and BD internalisers.8

Technology and regulation have also impacted trading in other ways. Trade sizes have decreased as the result of orders being split into smaller increments to reduce the impact on market price and perhaps as a consequence of more complex trading strategies. Decimalisation of US equities markets has led to smaller tick sizes with prices quoted in pennies rather than sixteenths of a dollar. In options markets, the existence of puts and calls, penny pricing, multiple delivery months and numerous strike prices based on market moves has resulted in thousands of options series available for trading. Assimilating the increased volumes of market data resulting from these changes exceeds the capabilities and response times of human traders. In contrast, computers are ideally suited for handling thousands of data points per second.

Over time, some trading companies also began to adapt algorithms that were used in equities markets for trading of other asset classes. Algorithmic trading strategies, which seek statistical arbitrage and other types of trading opportunities, are now used in the trading of every major asset class, but this transformation did not happen overnight.

Throughout most of the history of floor-based trading, traders stood in a single pit or post and traded one product like Colgate-Palmolive stock. Later, some began to trade multiple products across pits at a single exchange, such as soybeans versus soybean oil. A progression to multi asset class trading across exchanges ensued, such as trading S&P options at the Chicago Board Options Exchange against S&P futures at the Chicago Mercantile Exchange. Some US traders then adapted their trading strategies to a single non-US exchange such as Eurex, where they traded multiple products like DAX versus Eurostoxx. Today, many companies trade globally using complex baskets of products that are driven by multifactor models.

As noted in a recent white paper, ‘We are not just talking here about Chicago traders in Europe or Dutch traders in Asia. We are now seeing Asian trades arbitraging US and European markets or even other emerging markets, such as Latin America or Eastern Europe. A truly global geographic trading landscape is fast becoming a very small and crowded world.’9

One factor important to certain trading strategies is reducing latency, which includes the speed at which market data and signals from the marketplace are processed as well as the time it takes to transmit an order to a trading venue and for a trading venue to acknowledge that order. Some trading strategies even utilise ‘black boxes’ that are capable of reacting to market data, transmitting thousands of order messages per second, cancelling and replacing orders based on changing market conditions, and capturing price discrepancies with little or no human intervention.10 At some companies, computer programs may even incorporate machine readable news into trading decisions.

Another way HST companies may reduce latency is by placing (co-locating) their server(s) in the same data centre that houses a trading venue’s server(s). Doing so provides companies
with a competitive edge by significantly reducing the time it takes to access the central order book\textsuperscript{11} and to transmit orders to the trading venue’s matching engine, a computer(s) where buy and sell orders are matched. Any slippage in trade price or in filling orders can impact the trading strategies of latency sensitive trading companies because price takers are exposed to market risk prior to receiving confirmation that their order has been filled. Price makers are exposed to the risk that their prices will remain in the market at a time when the market has moved in the opposite direction of their trading strategy and before their cancel orders are processed.

Co-location also enables trading companies to streamline their relationships with telecommunication and power providers, leasing agents, network engineers and so on, into a single point of contact. The co-location provider manages these relationships and generally maintains multiple data centres around the globe, which enables trading companies to access markets with multiple redundancies without making significant technology investments.

In the past, co-location services were provided by independent service providers. More recently, some trading venues have begun to compete directly with these service providers by building their own co-location facilities and housing their matching engine(s) at that site. By doing so, trading venues may be finding new ways to generate revenues from their latency sensitive customers that lease space within the trading venue’s co-location site.

Trading venues may try to attract HST companies for other reasons. The highly fragmented structure of US securities markets means that a multitude of trading venues\textsuperscript{12} are vying for market share in an environment of declining trade volumes since the financial crisis (Exhibit 4.2). Designing technology, services and pricing structures in such a way that they appeal to HST companies, which generate large numbers of orders, may enable trading venues to increase volumes and thereby deepen liquidity in their own market. Such practices may be used by trading venues in other markets as well.

In terms of competition through technology and services, trading venues may sell market data, including detailed information about trades like price information, which trading companies incorporate into their trading strategies. The speed at which companies receive and process this data impacts the timelines in which their trading strategies are executed. An outstanding issue is whether some trading venues provide data feeds to certain customers faster than others. In 2012, the SEC fined the New York Stock Exchange (NYSE Euronext) US$5 million for improperly delivering market data to some proprietary trading companies faster than it delivered the information to the consolidated tape. Such early access to data could have provided these companies with a substantial competitive advantage over retail and long-term investors.\textsuperscript{13}

Another way trading venues compete is through the logic they use to match orders. One type of matching algorithm some trading venues use is first-in-first-out (FIFO), also known as price/time priority, whereby orders are matched against the best price and the time the order arrived in the queue. If there are multiple orders at the same price, the order with the earliest time stamp will be matched first. FIFO algorithms may advantage trading companies with fast systems whose orders arrive in the front of the queue.

Some trading venues may also try to attract liquidity through the pricing structures they employ, such as volume-based pricing and rebates, as well as through the special order types they offer. In 2012, a whistleblower reported to the SEC that stock exchanges were offering
HST companies special order types that allowed them to trade ahead of less sophisticated investors.14

Concurrent with an increasingly competitive environment, many exchanges have converted from member-owned to for-profit organisations in recent years and interests of shareholders and members may not necessarily be aligned. Moreover, some HST companies have equity ownership stakes in certain exchanges. An open question for regulators and policy-makers is whether these ownership stakes allow HST companies to influence the structure of trading venue’s systems to their own advantage but to the detriment of optimal risk management. An alternative possibility is whether these companies’ familiarity with risks of HST causes them to influence trading venues to adopt more rigorous risk controls.

It should also be noted that competition is not limited to trading venues. Certainly, some HST companies are engaged in a costly technology arms race as they compete for profits using low latency trading strategies and systems. In addition, some BDs/FCMs may have invested heavily in technology as they compete directly with other HST companies.

At the same time, BDs/FCMs are operating in a highly competitive environment that includes decreased income from: (i) commissions for execution services because more and more of their customers are adopting electronic trading; (ii) smaller trade volumes and shallower

---

**Exhibit 4.2**

**US equities share volume (in billions)**

![Graph showing US equities share volume from 1990 to 2013 with peaks in 2007 and 2008.](source: Tabb Group)

---

14
investment pools in some markets; (iii) tighter spreads; and (iii) lower interest rates. With regard to the latter, one of the ways BDs/FCMs make money is by investing customer money overnight and the low interest rate environment since 2008 has cut into these profits.

The competition for speed has also caused some BDs/FCMs to tout how quickly they can conduct pre-trade risk checks on customer orders before they reach the trading venue. The risk is that some BDs/FCMs may establish less stringent pre-trade risk controls to attract and retain latency sensitive customers.

**Technology glitches**

In recent years, stories of seemingly well-managed companies at various levels of the trade life cycle which have failed to properly control the risks of HST have been appearing in the press with some frequency. One of the earliest reported incidences occurred in 2003, when a US trading company became insolvent in 16 seconds after an unauthorised employee switched on an algorithm. It took the company 47 minutes to realise it had gone bust and to call its clearing bank, which was unaware of the situation.15

More recently, the NYSE Euronext levied a US$150,000 fine on Credit Suisse in 2010 for an incident that occurred in 2007, in which the company failed to ‘adequately supervise development, deployment and operation of proprietary algorithm, including failure to implement procedures to monitor certain modifications made to algorithm’. The episode resulted in hundreds of thousands of cancel/replace orders jamming the exchange’s systems.16 Specifically, the company violated NYSE Rule 401 by implementing a ‘proprietary algorithm that did not have appropriate checks designed to prevent submission of hundreds of thousands of erroneous messages or to alert Company in event of rejected messages’.17

In 2010, Chicago-based Infinium Capital Management was fined a total of US$850,000 for violating Chicago Mercantile Exchange (CME) rules in 2009 and New York Mercantile Exchange (NYMEX) rules in 2010. The latter incident occurred in February 2010 when the company sent 6,767 buy orders to NYMEX for light sweet crude oil. Over the next 24 seconds, 4,612 contracts were filled before the program could be cancelled.18 The algorithm causing the mishap had only been tested a few hours before it was introduced into the market19 and caused the price of oil futures to surge by about US$1 in less than a minute.20

The Osaka Securities Exchange (OSE) disciplined Deutsche Bank in June 2010 after its automated trading system sent US$182 billion worth of Nikkei 225 contracts to the exchange. Luckily, only 0.3% of the orders were executed before the program was cancelled.21 Nevertheless, the error resulted in a 1% decline in the Nikkei 225 contract, which fell 110 points.22

Technology glitches are not limited to trading companies and BDs/FCMs, however. In September 2010, the CME sent 30,000 orders from its test environment into production and promised to pay any resulting losses from the trades.23 A month later, a system upgrade at NYSE triggered nearly a 10% drop in the SPDR S&P 500 ETF. The trades were later voided by the exchange.24 In November, 2010, the exchange, Direct Edge, made untested computer code changes that resulted in US$773 million in incorrect trades for three customers. Contrary to its own rules, Direct Edge then tried to trade out of the positions itself.25
Exchanges also experienced a rash of technology issues throughout 2011. In February, the London Stock Exchange (LSE) Group suspended trading on two of its exchanges due to market data issues. The same month, the Australian Securities Exchange (ASX) closed early as a result of a problem with its messaging system and the CME halted trading in natural gas when it experienced a technical issue on its Globex platform. Trading was suspended on NYSE Arca Europe in March 2011 due to data feed issues. In April, a technical glitch at NASDAQ OMX fed incorrect quotes to market-makers and a market data outage halted trading in more than 80 securities. Traders complained it took the exchange hours to disclose which trades would stand and which would be broken. Trading was disrupted at NYSE Euronext three times in June, once in July and once in August due to technical issues.

Exchanges were not the only companies with technology problems in 2011, however. In March, a programming error at Deutsche Bank caused it to attempt to sell US$1.3 billion worth of SPI futures contracts on the ASX in two minutes and pushed the relevant SPI contract down by 17 points before the program could be cancelled.

January 2012 began with another technical glitch at NYSE. In February 2012, a NYSE Euronext’s Amex options market-maker had a computer malfunction that flooded the markets with quotes that led to over 31,000 accidental executions. Some were allowed to stand, some were cancelled and some had their prices adjusted. In March 2012, the US Financial Industry Regulatory Authority (FINRA) fined Getco US$450,000 for failing to adequately supervise its algorithmic and high frequency trading from a period of 7 May 2010 to December 2011. Also in March 2012, the US exchange, BATS, withdrew its IPO following a software glitch that also caused a 9% decline in the price of Apple shares and paused the buying and selling of Apple for five minutes.

In May 2012, a technical error at NASDAQ delayed the start of trading for Facebook’s initial public offering (IPO) and prevented some investors from learning whether they had successfully purchased or sold shares. This prompted some of them to enter orders numerous times and led to uncertainty about their overall positions in the market. As a result of the error, the Swiss bank UBS is said to have suffered a loss of more than US$350 million. Knight Capital reportedly lost over US$35 million in the Facebook incident as well. As already mentioned, Knight Capital was brought to the brink of bankruptcy when it lost US$440 million in August 2012.

In March 2013, Nikkei 225 stock average futures volume plunged more than 90% due to a software error at the OSE. In April 2013, computer algorithms that read and incorporate news reports in their trading strategies responded to a false Twitter report that the White House had been bombed and President Obama injured. With the markets already on edge following bombings at the Boston Marathon, the report caused a 140 point plunge in the Dow. Following the incident, the Associated Press news agency said its account had been hacked. The Dow recovered from its decline within six minutes with triple digit gains.

The frequency with which these incidents are occurring in seemingly well-managed companies as well some of the resulting losses is focusing attention on the need for risk controls at every level of the trade life cycle, from order submission to clearing and settlement.
Life cycle of a trade

To better understand where mechanisms to manage the risks of HST may reside, we begin with the life cycle of a trade and review the ways in which trading companies, trading venues, clearing organisations and clearing BDs/FCMs each have responsibilities for establishing and administering risk controls. Clearing BDs/FCMs are members of an exchange clearing organisation, where trades are cleared and settled. The types of trading companies engage in, the ways in which orders are submitted to trading venues, and how those orders are checked for errors, matched, cleared and settled are also discussed.

Trading

In the US, some trading companies and BDs/FCMs only trade for their own (proprietary) account, take positions in the market based on their trading strategies and risk tolerances, and use their own capital to trade. In the US, the Volcker Rule which is scheduled to be implemented in July 2014 is designed to restrict banks from engaging in proprietary trading.44 Other BDs/FCMs may have customers that trade with their own money. Customers may include companies (such as airlines that utilise futures markets to lock in a price for fuel), private equity companies, pension funds, mutual funds, retail customers, and so on. BDs/FCMs may also have customers that are members of an exchange, but not members of the exchange clearing house (non-clearing members). BDs/FCMs set limits on the exposures they are willing to accept from each customer’s trading activities. Finally, some BDs/FCMs participate in both proprietary and customer trading.

In a manual trading environment, customers telephone their orders to their BD/FCM or convey them by some electronic means and the BD/FCM executes those orders on the trading floor. As trading migrated from a floor-based environment, some BDs/FCMs began to offer their customers electronic, point-and-click trading systems that provide faster execution capabilities. One of the ways BDs/FCMs may manage potential losses that might arise from their customers’ point-and-click trading activities is by setting parameters, like credit and other limits, on these trading systems.

Over time, some customers found these legacy point-and-click trading systems were not keeping pace with current technological advances and began to seek alternative solutions that would provide even quicker execution in order to reduce latency and to increase the likelihood of obtaining their desired price in the market. In response to this demand, some trading venues and BDs/FCMs began to allow certain customers to access the markets directly (send their orders directly to the trading venue without using the BD/FCM’s trading system) using proprietary trading platforms the customer developed and/or vendor-provided trading platforms the BDs/FCMs approved.

A host of industry and regulatory reports recommend that trading companies that access the markets directly establish and enforce various pre-trade risk checks such as order size, intraday position and other limits that are appropriate based on a company’s capital, experience and risk tolerance.45,46,47,48,49,50,51,52,53,54,55,56 Well-designed trading systems include pre-trade risk checks, which if enabled, do one or more of the following: (i) alert a company when a trade(s) is approaching a pre-set limit; (ii) stop trades entirely once a limit is breached; or (iii) require traders to take opposite positions when a limit is hit.
At times, algorithms deployed by HST companies that access the markets directly may go awry due to corrupt data, programming errors, hardware failures, network issues, or any number of other unforeseen events. Depending on the pre-trade risk checks performed by the trading system, out-of-control algorithms and/or erroneous orders may be detected and stopped. If they are not, the clearing BD/FCM serves as the next point in the trade life cycle to manage risks before orders are transmitted to the exchange matching engine(s).

**Clearing BDs/FCMs**

Clearing BDs/FCMs are financially responsible to the clearing organisation for their proprietary trades and for the trades of any customers they may have. Consequently, it is of paramount importance that clearing BDs/FCMs implement comprehensive risk controls to manage these exposures.

Prudent business practices and regulatory requirements motivate clearing BDs/FCMs to ensure pre-trade risk checks are applied to customer orders before they are sent to a trading venue. As already discussed, these controls may be placed on the point-and-click trading systems provided by the BD/FCM or by a vendor the BD/FCM approves.

For customers that access the markets directly, clearing BDs/FCMs may conduct pre-trade risk checks in one or more of the following ways.

- By remotely accessing the customer server(s) at the co-location facility and checking whether there are established risk limits on that server(s).
- By setting risk limits on the clearing BD/FCM’s server(s) at the co-location facility and requiring customers to connect to that server(s).
- By using functionality provided by the trading venue that enables them to set risk limits for the customer on the trading venue server. Not all trading venues provide such functionality, however.

**Trading venues**

If an out-of-control algorithm and/or erroneous order is not detected and stopped by the risk controls of the trading company or of the clearing BD/FCM, the trading venue serves as the last line of defence to halt orders before they reach the matching engine. As such, some trading venues establish pre-trade risk checks on their own servers and automated systems compare all orders against these limits.

Trading venues that impose pre-trade risk checks increase latency equitably and uniformly for all market participants at that trading venue. It is important to note, however, that the types of pre-trade risk checks individual trading venues conduct vary. Moreover, pre-trade risk checks increase latency and trading venues may compete on how quickly these checks can be executed in order to attract and retain latency sensitive trading companies.

As outlined under ‘Clearing BD/FCMs’, some trading venues also offer functionality to clearing BDs/FCMs that enables them to set limits for the orders of customers that access the markets directly. This functionality may be mandatory or optional or may not be offered at all, depending on the trading venue. Based on their assessment of risk, some BD/FCM may
set fewer/no pre-trade risk limits at the trading venue level. This may result in faster order submissions for some customers of BDs/FCMs than for others. Finally, some trading venues provide functionality directly to customers that enable them to set pre-trade risk limits below the levels set by their BD/FCM.

Once buy and sell orders are matched in the matching engine, trading venues send post-trade information on matched orders (filled trades) to trading companies and to clearing BDs/FCMs to help them manage their exposures. Some trading venues also provide information on orders that are still working in the market in these data. Of critical importance in the management of risk, is the time frame in which trading companies and BDs/FCMs receive these post-trade data. Delays in receiving post-trade data hinder optimal risk management at trading companies and BDs/FCMs.

### Clearing organisations

Trading venues also send post-trade data on matched trades to the clearing organisation where the trade is settled. Some trading venues send this post-trade information to clearing organisations in real-time and others send it once daily. Clearing organisations calculate and monitor the intraday exposures of their clearing BDs/FCMs using different methodologies and frequencies, but in general, using the most up-to-date information they receive from trading venues. When risk tolerance levels are breached, clearing organisation staff responds on a case by case basis, and/or alerts senior management, and contacts the relevant clearing BD/FCM.

Clearing organisations also provide clearing BDs/FCMs with various types of post-trade data, such as information on open positions, performance collateral and collateral substitutions, cleared trades, and so on, so that they can monitor their financial exposures. Clearing BDs/FCMs may compare some of these post-trade data from the clearing organisation to information from the trading venue and other sources to identify discrepancies in customer positions. If any are detected, clearing BDs/FCMs contact the customer.

To summarise, when an order is originated in the life cycle of a trade, it may get checked against risk limits set at the trading company, clearing BD/FCM and trading venue before it reaches the matching engine. Once an order is matched, post-trade information on the trade is communicated by the trading venue to trading companies that access the markets directly, clearing BDs/FCMs and the clearing organisation. The clearing organisation also provides post-trade data on the trade to the clearing BD/FCM. With controls at so many levels, how can things go wrong?

### Key findings from Federal Reserve Bank of Chicago study

To better understand the risk controls used in the marketplace, staff from the Federal Reserve Bank of Chicago’s Financial Markets Group conducted over 30 interviews in 2010 and 2011 with primarily US domiciled technology vendors, proprietary trading companies, BDs and FCMs and clearing organisations. Non-US entities interviewed include one exchange, one clearing organisation and one foreign BD/FCM. A particular focus during the dialogues was how the risks posed by trading companies that access the markets directly are managed. Other topics of interest or concern to these companies were also discussed.
The interviews revealed there are tensions regarding the responsibilities for setting and administering risk controls at all levels of the trade life cycle. For example, some trading companies and clearing BDs/FCMs may rely solely on risk controls at the trading venue level instead of implementing their own pre-trade controls. The problem with this approach is that trading venues’ risk controls may be structured in a way that does not stop erroneous orders before they are filled. In such circumstances, trading venues may have publicly available policies for busting clearly erroneous trades, but these policies may vary by trading venue and by product. The timeframe in which trading venues decide whether or not to bust trades may also vary. Delays in making such decisions create uncertainties for trading companies with regard to the status of their orders and market exposures.

As already mentioned, some trading venues also provide BDs/FCMs, and their customers that access the markets directly, functionality which enables them to set risk limits on the exchange server, thereby helping them to better manage their risks. However, staff at one trading venue said some market-makers set these risk limits at a level where protection is nil.

Of particular concern, staff at one options exchange said credit risk was the clearing organisation’s concern, exonerating any responsibility for setting credit limits at the exchange level. This rationale is driven by the fragmented nature of US equities and options markets and by the fungible nature of their products, which can be bought and sold on numerous trading venues. Equities and options exchanges are reluctant to set credit limits for their customers because they can only see the positions in their own markets and do not typically have information on their exposures at other trading venues. As such, they cannot determine whether a position on one exchange is offset at another.

This is consistent with another finding of the study, which revealed that there may be times when no single entity in the trade life cycle has a complete picture of an individual company’s exposures across markets. For example, some trading companies and BDs/FCMs are unable to calculate their enterprise-wide portfolio risk. Many trading companies and BDs/FCMs also trade products on multiple trading venues, and these trades are settled at multiple clearing organisations. This prevents any single trading venue or clearing organisation from seeing the total positions amassed by such companies. Moreover, many trading companies use multiple BDs/FCMs to clear trades, which results in no single clearing BD/FCM being able to establish the trading company’s total exposures across markets. Although some clearing BDs/FCMs may require trading companies to submit statements from their other clearing BDs/FCMs to assess such exposures, these statements may be only provided once a day.

Similarly, the fragmented structure of US equities and options markets precludes the discovery of manipulative practices across multiple trading venues. While each trading venue interviewed can and does monitor for manipulation in its own market, no single trading venue has an immediate, horizontal view of all market activity. Staff at one trading venue explained, ‘There is no way to tell if a trading company is bidding up the prices on two exchanges and “pounding” on a third’. It is hoped the SEC’s plans to implement a consolidated audit trail will aid in the detection of manipulative market practices in the equities and options markets.

In US futures markets, where some products are traded almost exclusively at a single trading venue, exchange staff are able to view the majority of trades in these products. However, like equities and options trading venues, futures markets only see their portion...
of trades and are unable to detect manipulative practices that companies may be engaging in across markets. As such, it is imperative that regulators develop state-of-the-art tools for market surveillance and information sharing.

Federal Reserve Bank of Chicago staff also found that incidences of companies experiencing out-of-control algorithms were more common than anticipated prior to the study. Two of the four clearing BDs/FCMs, six of the nine proprietary trading companies and every exchange interviewed experienced one or more errant algorithm(s). One exchange said it could detect an out of control algorithm if it had a significant price impact. However, the exchange could not identify an out-of-control algorithm if it slowly added up positions over time, although exchange staff had heard of such occurrences.

As might be expected with incidences of out-of-control algorithms, some companies lacked stringent processes for the development, testing and deployment of code. For example, a few trading companies interviewed said they deploy new trading strategies quickly by tweaking old codes and placing them into production in a matter of minutes. In fact, one company had two incidents of out-of-control algorithms. To address the first occurrence, it implemented additional pre-trade risk checks. The second out-of-control algorithm was caused by a software bug that was introduced as a result of someone fixing the error code that caused the first situation.

Another finding of the study is that the periodicity with which trading companies, BDs/FCMs and clearing organisations are able to calculate and monitor intraday exposures is based in part on the frequency with which they receive up-to-date matched trade information from trading venues. However, not all trading venues globally even offer this post-trade data, and for those that do data formats and time frames for delivery are inconsistent. While some trading venues send post-trade data in near real time (microseconds), others provide it within hours or at the end of the trading day. Delays in receiving this information hinder trading companies, clearing BDs/FCMs and clearing organisation’s ability to optimally manage their risks.

At the same time, some BDs/FCMs back-end systems are unable to process post-trade data from trading venues and clearing organisations in the time frames they are delivered. It is recommended that trading companies and clearing BDs/FCMs maintain automated systems that can process information from trading venues and from clearing organisations at the speed in which it is provided.

One of the most surprising findings from the study is that staff at all levels of the trade life cycle called for regulators to work with the industry to develop best practices for the risk management of HST. This includes specifying the types of pre and post-trade risk checks that trading companies, clearing BDs/FCMs and trading venues should use. Best practices for clearing organisations should also be addressed. Once these best practices are established, regulators should: (i) clearly communicate them to the industry; (ii) specify the penalties for non-compliance; and (iii) audit market participants’ adherence to them.

**Recommendations**

The interviews also highlighted a number of best practices. To better detect out-of-control algorithms and errant trades, one trading company developed a volume detecting mechanism
that restricts the number of orders that can be submitted per second to a trading venue. The mechanism also identifies rapid fills that are out-of-the-ordinary, which may indicate trading against another company’s out-of-control algorithm or erroneous orders. Such a capability is considered a best practice for trading companies and BDs/FCMS.

It is also recommended that trading companies, BDs/FCMs and trading venues develop and maintain a ‘kill switch’ functionality to stop trades that exceed historic norms by some measure. However, it is recognised that it may be more beneficial if a human makes the decision on whether or not to activate it since an automatic kill switch has the potential to worsen a situation by breaking one side of a multi-legged trade. Nevertheless, it is also recognised that the erroneous positions can build up rapidly, as in the case of Knight Capital, and it is of the utmost importance for trading companies, clearing BDs/FCMs and trading venues to identify, assess and respond to problems in the shortest timeframe possible. Consequently, it is imperative that companies at all levels of the trade life cycle institute and maintain robust risk management procedures and subject them to periodic review. These practices should take into consideration the interaction of an array of systems, processes and staff. For example, one best practice observed was the establishment of a cross-functional team, comprised of trading, IT, legal, compliance and other staff, which periodically meets to discuss new and emerging risks. Formal risk management training processes may also be beneficial.

Companies at all levels of the trade life cycle should also have controls to ensure that unauthorised staff cannot access systems or change risk limits. Strict separation of duties and access to trading, risk management and middle/back office systems should be maintained.

Quality management procedures related to the development, testing, modification and deployment of code into production are also needed at all levels of the life cycle of a trade. The efforts of the AT 9000 Working Group in developing quality management standards and guidelines for automated trading as an alternative to regulation may be beneficial in this regard.

In addition, risk managers need a clear understanding of the assumptions and the source data that feed risk models. For example, estimating a market event similar to the flash crash occurring within the next month is a very different proposition than estimating the likelihood of such an event occurring in the next 10 years. Risk managers also need to know how frequently the data that feed risk models and post-trade risk controls are updated. A common misperception may be that all risk management systems have the capability to calculate exposures in near real time whereas additional time may be needed to calculate VaR and Greeks for options and to receive data from multiple trading venues and sources, including information on OTC and obscure products, and so on.

**Suggestions for regulators**

As already mentioned, regulators need to work with the industry to develop best practices for pre and post-trade risk controls at all levels of the trade life cycle. It is also recommended that regulators evaluate the methodology for how trading companies, clearing BDs/FCMs and trading venues establish risk limits and the reasonableness of the limits set, as they could be fixed so high that they are ineffective and circumvent the reason that they were put in place.
In addition, regulators should review the procedures that are followed when risk limits are breached. The adequacy of BDs/FCMs and trading companies’ access controls for source code, trading systems and trading limits should also be assessed to ensure that unauthorised staff cannot access/modify/deploy trading algorithms or adjust trading limits. Because trading companies, BDs/FCMs and clearing organisations rely on post-trade data from trading venues to manage their risks, regulators should work with the industry to standardise the formats trading venues use to deliver post-trade information and provide guidance on the frequency in which these data should be delivered. Similarly, regulators should provide guidance to trading companies, BDs/FCMs and clearing organisations on the frequency with which these data should be incorporated into their risk monitoring systems due to the speed at which HST companies can generate large, unintended positions. To that end, regulators should also assess the methodology trading companies, BDs/FCMs, trading venues and clearing organisations use to monitor out-of-the-ordinary exposures and the frequency with which these exposures are monitored.

Trading venues may also have differing policies for defining error trades and for deciding how quickly error trades are broken or adjusted. Some trading venues make a decision on whether to break/adjust error trades in a matter of minutes while others may take hours or even days. Lengthy processes for breaking/adjusting error trades not only create uncertainties for trading companies during times of market stress, they also make it difficult for companies to ascertain their exposures as breaking trades can impact one leg of a multi-hedged position. It is, therefore, recommended that regulators and trading venues work together to standardise the definition for error trades and the time frames in which decisions are made regarding whether the error trades will be broken or adjusted.

Federal Reserve Bank of Chicago staff have a number of recommendations for regulatory audits of clearing BDs/FCMs as well. The interviews revealed some clearing BDs/FCMs raise customer order size limits to increase the likelihood that customer orders will be filled by trading venues that use pro-rata matching algorithms. Under such circumstances, regulators should ascertain if clearing BDs/FCMs employ other pre-trade risk checks, like reasonable credit limits, to offset the increased order size limits. Audits should also ensure that clearing BDs/FCMs have the capability to cut off trading for any customer and maintain exclusive control of such hardware.

Staff also discovered during the interviews that some clearing BDs/FCMs allow their HST customers to access the markets directly with no pre-trade limits on their trading platform so long as the trading venue has pre-trade risk controls in place. As pre-trade risk controls utilised by trading venues vary widely, Federal Reserve Bank of Chicago staff recommend that regulators evaluate whether risk controls at the trading venue accessed directly by customers are sufficiently robust for prudent risk management.

Interviewees had additional suggestions for regulators. The majority felt strongly that regulatory agencies should: (i) hire staff with practical experience in and deep knowledge of the markets, especially HST; (ii) concentrate on solutions rather than problems; and (iii) work with industry participants to make the markets better. Thoughtful policies that focus on the global picture, avoid unintended consequences, and are not politically reactive are needed. Moreover, domestic and international regulatory agencies should work together to harmonise rules in order to prevent regulatory arbitrage.
**Suggestions for trading companies**

In addition to working with regulators to define best practices for risk management of HST, trading companies should be encouraged to adopt a number of operational risk controls, such as subscribing to cancel on disconnect at each trading venue where it is offered. It is also recommended that trading companies estimate the costs and benefits of including working orders in their risk calculations. Finally, trading companies are encouraged to reconcile their trading activity using data from three sources: (i) trading venue(s); (ii) clearing member(s); and (iii) the trading company’s internal databases.

**Suggestions for clearing BDs/FCMs**

It is of paramount importance that clearing BDs/FCMs implement comprehensive risk controls to manage their exposures, including proprietary and/or customer trading activities, as they are financially responsible for all of the trades they clear. It is, therefore, recommended that clearing BDs/FCMs work with regulators to develop best practices for pre and post-trade risk management of HST.

The interviews revealed clearing BDs/FCMs conduct credit reviews during the onboarding process for new customers. If a new customer engages in HST and desires to access the markets directly, some BDs/FCMs subject such companies to additional screening. However, most of the clearing BDs/FCMs interviewed do not have formalised audit procedures to periodically monitor HST customers that access the markets directly subsequent to the onboarding process. It is, therefore, recommended that clearing BDs/FCMs review the adequacy of client’s HST risk management controls and procedures during the onboarding process and during subsequent periodic audits. Reviews should include access controls, change management procedures and pre and post-trade risk controls. BDs/FCMs should also consider requiring HST customers that access the markets directly to subscribe to cancel on disconnect functionality at the trading venues where it is offered.

It is also recommended that BDs/FCMs obtain a high level understanding of their HST customers’ trading strategies during the onboarding process and monitor whether there is a significant divergence from those strategies. For example, if a HST company indicates it engages in statistical arbitrage strategies and later begins to mass quote in options, the risk profile would be significantly different than originally anticipated and additional collateral may be appropriate.

Staff also learned during the interviews that some BDs/FCMs’ risk management systems could not process information as quickly as it was provided by trading venues and clearing organisations. As such, BDs/FCMs should estimate the costs and benefits of updating their risk systems so that risk calculations can be made at the frequency within which data is received from trading venues/clearing organisations. This estimation should also include the costs and benefits of including working orders (from those trading venues that provide such data) in risk calculations.

**Suggestions for trading venues**

In addition to working with regulators to develop best practices for pre and post-trade risk controls, trading venues are encouraged to consider a number of other risk management...
procedures like standardising the definition for error trades and the time frames in which decisions are made regarding whether error trades will be broken or adjusted. Lengthy processes for breaking/adjusting error trades create uncertainties for trading companies during times of market stress such as whether or not one or more legs of a hedged position may be broken.

In addition some trading venues offer functionality to clearing BDs/FCMs and/or their customers and/or market makers to set pre-trade limits. There may be a role for such trading venues to conduct periodic audits of the types of pre-trade risk controls companies utilise, the methodology for how these controls are applied, and the reasonableness of the limits set. Trading venues are also encouraged to obtain a high level understanding of HST companies’ trading strategies before permitting such companies to access the markets directly.

It is also advised that trading venues consider a number of recommendations for order management. Assessing the costs and benefits of including working orders in the post-trade data provided to trading companies and clearing BDs/FCMs, could provide these companies with enhanced capabilities for calculating their potential exposures. To address extreme scenarios, trading venues offer cancel on disconnect and kill switch functionalities. How these functionalities are activated and what orders get deleted from the order book may vary by trading venue, which may create uncertainties with regard to which orders are left working in the market. Some exchanges delete all orders while others provide clients with the option to select what order types should be cancelled. Trading venues may also want to consider standardising the definition for cancel on disconnect and establishing written policies regarding the type of orders that are deleted when cancel on disconnect and kill switch functionalities are activated. Trading venues are also encouraged to provide price tolerance checks to market orders or eliminate market orders.

**Suggestions for clearing organisations**

The periodicity with which clearing organisations are able to calculate and monitor intraday exposures is based in part on the frequency with which they receive up to date matched trade information from trading venues. Clearing organisations that receive information from trading venues in real time are able to continually conduct intraday risk monitoring of their clearing BDs/FCMs, including those that engage in HST and those that clear for HST companies. Clearing organisations that receive data from trading venues once daily are only able to conduct comprehensive risk management at that point in time. Using historic data on clearing BDs/FCMs’ intraday positions and comparing that information to current open positions could help detect unusual trading patterns, which could result in intraday calls for additional performance collateral. Clearing organisations should also assess the frequency in which they revalue clearing BDs/FCMs’ portfolios throughout the day to determine if it is adequate for capturing volume spikes and performance collateral utilisation outside the norm.

Margin deposits are used as an ex-ante safeguard against unfavourable price movements and as protection against potential liquidation losses that could arise if a clearing member was unable to fulfil its financial obligations. Assessing margin surcharges on concentrated positions and on positions in less liquid products to protect against the potential losses that
could arise if it were to take an extended period of time to liquidate such positions should be considered a best practice.

It is also recommended that clearing organisations with a vertical clearing structure monitor open positions at the customer account level, even if the customer clears through multiple clearing BDs/FCMs. Margin surcharges on concentrated positions would apply to all clearing BDs/FCMs where that customer had such positions.

Guarantee funds also serve as a financial safeguard to protect the clearing organisation when potential liquidation losses might exceed the margin deposits of the defaulter. Clearing organisations may want to consider including executed and cleared volumes in their calculations for determining guarantee fund contributions to better manage exposures to their clearing BDs/FCMs, including HST companies, which may have high volume levels but low margin requirements.

Conclusion

HST environments have the potential to generate errors and losses at a speed and magnitude far greater than that in a floor or screen-based trading environment. Because losses arising from HST have the potential of impacting the financial conditions of trading companies and, depending on the magnitude, clearing BDs/FCMs and possibly clearing houses, determining and applying the appropriate balance of financial and operational controls is crucial. It is also critically important that each company involved in the life cycle of a high-speed trade has its own risk controls and does not rely solely on another company to manage its risk.

Nevertheless, the types of risk-management tools employed by individual trading companies, clearing BDs/FCMs, trading venues and clearing organisations vary. Regulators should, therefore, work with the industry to define best practices for the risk management of HST and to audit companies’ compliance with these best practices. Each level of the trade life cycle should also have a risk manager who can respond quickly if exposures exceed pre-set limits.

Finally, it is recognised that issues related to risk management of HST are numerous and complex and cannot be addressed in isolation within domestic financial markets. For example, placing limits on HST within one jurisdiction might only drive trading companies to another jurisdiction where controls are less stringent.

The author wishes to acknowledge the helpful comments on this chapter by John McPartland and Richard Heckinger and the contributions to the Federal Reserve Bank study referenced therein by Rajeev Ranjan and John McPartland. The opinions expressed in this chapter are those of the author and do not reflect those of the Federal Reserve Bank of Chicago or the Board of Governors of the Federal Reserve System.

1 Algorithmic trading has two generally accepted definitions. (1) Algorithms were originally used as a tool to aggregate liquidity and to determine the optimal trade execution least disruptive to a stock’s price in the equities markets. Based on various statistical measures, they slice up big orders and route them to various trading venues where they are executed. (2) In some markets, algorithmic trading fully automates the decision making process by using mathematical models that determine which instruments to buy or sell.

2 Trading venues include exchanges and alternative trading systems (ATS) like Electronic Communication Networks (ECNs) and dark pools.
Operations and IT risk


10 Black box trading strategies are 100% automated, pre-programmed, and traders cannot interact or modify the algorithm.

11 The central limit order book is where electronic information on quotes/orders to buy and sell as well as current market prices is warehoused.

12 Currently, there are 16 securities exchanges registered with the SEC, see www.sec.gov/divisions/marketreg/mrexchanges.shtml. In 2010, the SEC estimated that there were 5 ECNs, approximately 32 dark pools and over 200 broker-dealer internalisers; see Federal Register, ‘Securities Exchange Commission concept release on equity market structure’, 17 CFR Part 242, 21 January 2010.


18 The Economist, ‘The fast and the furious: high-frequency trading seems scary, but what does the evidence show?’ 25 February 2012.


35 Steinert-Threlkeld, T, ‘Stock plunge leads to glitch on NYSE’s European derivatives exchange, again’, Securities Technology Monitor, 4 August 2011.
43 CNBC, ‘False rumour of explosion at White House causes stocks to briefly plunge; AP confirms its twitter feed was hacked’, CNBC.com, 23 April 2013.
47 FIA Asia, ‘Profile of exchange and FCM risk management practices for direct access customers’, 3 December 2007.
48 FIA-FOA, Clearing risk study, 2009.
53 See also Futures Industry Association, ‘Software development and change management recommendations’, Principal Traders Group and European Principal Traders Association, March 2012; Futures Industry Association, ‘Market access risk management recommendations’, April 2010; and Futures Industry Association ‘Recommendations for risk controls for trading companies’, Principal Traders Group, November 2010.
54 CFTC, ‘Recommended practices for trading companies, BDs/FCMs and exchanges involved in direct market access’, Pre-Trade Functionality Subcommittee of the CFTC Technology Advisory Committee, March 2011.
57 To ensure that BDs/FCMs are appropriately managing the risks of their customers that access the markets directly, the SEC implemented Rule 15c3-5 in July 2011, which among other things, requires BDs to maintain a system of controls and supervisory procedures reasonably designed to limit the financial exposures arising from customers that access the markets directly. The US Commodity Futures Trading Commission (CFTC) also adopted rules in April 2012 to bolster risk management at the FCM level.
63 For information on the AT 9000 Working Group, see www.at9000.org/.
64 VaR (Value at Risk) is a widely used risk measure of the risk of loss on a specific portfolio of financial assets. Greeks are the parameters that represent the sensitivities of the price of derivatives, such as options, to a change in the underlying parameters on which the value of an instrument is dependent.
Pro-rata algorithms match orders at the same price based on a variety of factors. For example, the matching sequence may be based on time stamps assigned by the exchange and orders in the queue may receive a fill proportionate to their size.

Some trading venues monitor whether the connection between their server and trading companies’ servers is lost due to network disruptions, an ungraceful disconnect (termination of a session by the trading company without an authenticated signoff), or the trading venue’s systems going down. When this connection is disrupted, some trading venues have cancel on disconnect functionality that deletes working orders from the order book. However, the policies regarding which open orders are deleted vary by trading venue. For example, some trading venues delete all working orders including good till cancel (GTC) orders and good till date (GTD) orders. Other trading venues only delete all working GTD orders. In addition, there is no standardised definition of cancel on disconnect across trading venues.

For example, GTD orders are non-persistent order types that are only valid for the trading day. At the end of the trading day, these orders get cancelled by the exchange. GTC orders are persistent order types that do not get cancelled by the exchange at the end of the trading day. GTC orders are generally cancelled by the trader/trading company/clearing member, but some exchanges set limits on how many days a GTC order can stay in the market.

For example, if a connection is lost for 30 seconds, cancel on disconnect is activated.

With regard to trading venues and clearing organisations, the SEC proposed Regulation SCI in March 2013, which would replace their voluntary compliance programs with enforceable rules. Among other things, these rules would require them to establish policies and procedures related to the capacity, integrity, resiliency and security of their technology systems.
Chapter 5

Algorithmic trading, flash crashes and IT risk

Philip Treleaven and Michal Galas
University College London

Overview

Algorithmic trading (AT) is growing rapidly, accounting for as much as 73% of US trading volume in 2009. However, it is also an area of major concern due to potential catastrophic ‘flash crashes’, such as the US$600 billion 26 May 2010 crash, and unforeseen behaviour of rogue algorithms, such as Knight Capital Group’s loss of US$460 million in 30 minutes. From a company perspective, the behaviour and risk of algorithmic trading systems (ATSs) and individual trading algorithms is poorly understood, and the efforts of industry regulation through legislation such as MiFID II (Level 2) may in fact increase instability.

For the past eight years University College London (UCL) has worked with the major investment banks and funds developing ATSs, and more recently with the regulators investigating high-frequency trading risk and systemic risk. In this chapter we review algorithmic trading IT risk. We introduce: (i) the structure of algorithmic trading algorithms; (ii) the system architecture of trading systems; (iii) then discuss flash crashes; and (iv) what might be termed ‘model risk’ – namely rogue algorithms. Finally to support UCL’s work we have developed our Algorithmic Trading & Risk Analytics Development Environment (ATRADE) platform which can be used both for virtual and real trading. It has been designed to: (i) speed the development of trading algorithms; (ii) evaluate algorithm risk; and (iii) assess algorithm programmers. ATRADE is described to illustrate the structure of AT systems, which is difficult to find in the literature.

Algorithmic trading and IT challenges

The hugely competitive nature of AT, the scarcity of expertise, and the potentially vast profits to be made, makes this a highly secret and somewhat paranoid community. It is, therefore, extremely difficult for IT professionals to find technical details of algorithmic trading systems’ implementation and for the QUANT developers to understand the complexities of individual algorithms and ATSs. ATSs are normally deployed for highly liquid markets across the majority of tradable asset classes, including cross-trading between the classes. Tradable asset classes include equities, futures, derivatives, bonds (US Treasuries and so on), and foreign exchange (currencies).
There are a number of closely related computational trading terms that are often used interchangeably, such as AR, systematic trading, and high-frequency trading (HFT) and so on. In general:

- **AT** refers to any form of trading using sophisticated ‘algorithms’ (programmed systems) to automate all or some part of the trade cycle. AT usually involves learning, dynamic planning, reasoning, and decision taking on the basis of logical inference from internal analytical, trading, risk and money management models;
- **HFT** is a more specific area, where the execution of computerised trading strategies is characterised by extremely short position-holding periods involving trading speeds in excess of a few seconds or milliseconds; and
- **ultra high-frequency trading** or low-latency trading refers to HFT execution of trades in sub-millisecond times through co-location of servers and stripped down strategies (strategies with minimal amount of functional layers separating it from the exchange, to further reduce latency) at exchanges, direct market access, or individual data feeds offered by exchanges and others to minimise network and other types of latencies.

Due to the commercial pressures on QUANT programmers and the sheer complexity of these IT systems, ATSs are often poorly understood and rushed into use before they have been thoroughly tested. In addition, incidents of flash crashes and rogue algorithms are spurring regulators to introduce controls on HFT,$^9,10$ such as holding periods and transaction taxes, which may in fact increase instability.

**Automated trading algorithms**

As discussed, given algorithmic trading systems’ increasing importance in the financial markets, the escalating sophistication of the algorithms involved, and commercial sensitivities, it is extremely difficult to find technical details of algorithms’ performance, behaviour, and risk. For completeness we give a brief introduction to AT.

In electronic financial markets, AT$^{11}$ is the use of computer programs to automate one or more stages of the trading process: pre-trade analysis (data analysis), trading signal generation (buy and sell recommendations) and trade execution (when and how to trade). Furthermore, trade execution is subdivided into two broad categories: (i) agency/broker execution, when a system optimises the execution of a trade on behalf of a client; and (ii) principal/proprietary trading, where an institution is trading on its own account. Each stage of this trading process can be conducted by humans, by humans plus algorithms, or fully by algorithms.

Exhibit 5.1 illustrates the major components of an algorithmic trading system.$^{12}$ The pre-trade analysis comprises three main components: (i) alpha model – this is the mathematical model designed to predict the future behaviour of the financial instruments that the algorithmic system is intended to trade; (ii) risk model – evaluates the levels of exposure/risk associated with the financial instruments; and (iii) transaction cost model – calculates the (potential) costs associated with trading the financial instruments.
At the ‘trading signal’ stage, the ‘portfolio construction’ model takes as its inputs the results of the alpha model, the risk model and the transaction cost model, and decides what portfolio of financial instruments should be owned and in what quantities, in the next time horizon.

Finally at the ‘trade execution’ stage, after the trading signal has been generated and the portfolio constructed, the ‘execution model’ performs several decisions with constraints on (actual) transaction costs and trading duration: the most general decision is the ‘trading strategy’ followed by the venue and order type, decisions handled by smart order routing.

The prerequisites for AT include:

- centralised order book – shared centralised order book listing the buy and sell orders;
- liquid markets – a highly liquid market and typically one suitable for high-frequency trading of securities (for example, equities, bonds, FX); and
• financial information protocols – financial information XML exchange protocols (for example, the Financial Information eXchange protocol (FIX)\textsuperscript{13} for the computer communication of information.

The complexity of ATSs largely centres on the alpha (forecasting) model and the risk model.

**Alpha models**

For the Alpha model, as illustrated in Exhibit 5.2, there are essentially two alternative basic strategy approaches:

- **theory-driven** – the researcher/programmer chooses a hypothesis for the way the securities are likely to behave and then uses modelling to confirm their hypothesis; and
- **empirical** – the researcher/programmer uses an algorithm to data mine and identify patterns in the underlying data, without any preconceived views on the behaviour of a security.

---

**Exhibit 5.2**

**Alpha models – predicting the future**

![Diagram](https://www.rasabourse.com)

*Source: Authors’ own*
Constructing the alpha model and, more specifically, setting the variables is a highly complex task. Numerous factors influence the actual algorithm implementation: (i) forecast goals (for example, direction, magnitude, duration, probability); (ii) forecast time horizon (for example, millisecond, day, week); (iii) the mix of instruments; (iv) the data available; (v) actual setting of the model’s variables; and (vi) the frequency of running the model.

Risk models

Risk models\textsuperscript{14,15} focus on the risks associated with a target portfolio of financial instruments and the relevant factors which affect the economic climate and hence the current/future value of the financial instruments. It does so by attempting to quantify both the risk associated with individual instruments and with the portfolio. Exhibit 5.3 illustrates the two principal approaches to risk which we will refer to as:

- \textit{limiting amount of risk} – limiting the size of risk is managing the exposure through: (i) limiting by constraint or penalty; or (ii) calculating volatility and dispersion; and
- \textit{limiting type of risk} – this focuses on eliminating whole types of exposure. As with other models, the approaches broadly subdivide into: (i) theory-driven; and (ii) empirical using historical data.

In limiting the above two – we take care of optimising the risk-reward ratio.

Data sources, services, streaming and cleaning

In algorithmic trading there is a strong correlation between the quantity/quality of the data available for analysis and the success of the trading system. Given the large number of variables in an ATS and their sensitivity, if the data is not rigorously cleaned then the behaviour of the system and its trading performance will be unpredictable.

Data sources

Data sources broadly comprise:

- \textit{financial data} – price data on financial instruments from exchanges and electronic communication networks (ECNs), as well as financial information from services such as Bloomberg and Thomson Reuters;
- \textit{economic data} – fundamental economic data, such as the overall state of the countries’ economies (for example, unemployment figures), interest rates, gross domestic product, and national policy; and
- \textit{social/news data} – sentiment data ‘scraped’ from social media (for example, Twitter, Facebook, Google), RSS feeds, and news services.
These sources subdivide into:

- **real-time** – live data feeds from exchanges, ECNs, news services, or steamed social media; and
- **historic** – previously accumulated and stored financial, economic, and social/news data.

These data sources, real-time and historic, are the cornerstone of the research, design and back-testing of trading algorithms, and drive the decision-making of all ATS components. However, buying (raw and especially cleaned) data is hugely expensive and cleaning data is highly time-consuming, but essential due to the sensitivity of trading algorithms.

**Exhibit 5.3**

### Risk models – sizing of exposures

<table>
<thead>
<tr>
<th>Risk model</th>
<th>Limiting type of risk</th>
<th>Limiting amount of risk (exposure)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Size limits (constraints, penalty)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Measuring risk (standard deviation)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Volatility</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Measuring risk (standard deviation)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VaR</td>
</tr>
</tbody>
</table>

**Source:** Authors’ own
Data services
The principal platforms for financial data are Bloomberg (www.bloomberg.com), Thomson Reuters (www.thomsonreuters.com) and Markit (http://www.markit.com/en/). Thomson Reuters is also leveraging its news service and provides a comprehensive social media service.

- Financial data vendors – the major general financial market data vendors are Bloomberg and Thomson Reuters, and for high-frequency data there are specialist providers such as Chi-X and Trading Technologies.
- Social data – platforms and services for social analytics are being provided by Thomson Reuters, Google and SAP amongst others; plus various smaller service providers such as Lexalytics and Socialmention.

As discussed, the Bloomberg and Thomson Reuters platforms are capable of streaming large quantities of data in either live or historical mode, and both architectures provide ‘big’ data\textsuperscript{16} storage facilities on the server-side and buffering capabilities on the client-side. Apart from data streaming, both platforms provide basic data analytics on the client-side and charting functionalities for data visualisation.

Stream processing
Another important IT topic is high performance ‘stream’ processing or ‘complex event processing’ (CEP). Increasingly, analytics applications that consume real-time financial ‘ticker’ data, social media and sensor networks data, and so on, need to process high-volume temporal data with low latency. These applications require support for online analysis of rapidly changing data streams. However, traditional database management systems (DBMSs) have no predefined notion of time and cannot handle data online, in near real-time. This has led to the development of data stream management systems (DSMSs)\textsuperscript{17} (that is, in main memory without storing the data on disk) that handle transient data streams online and process continuous queries on them. Example commercial systems include: Oracle CEP engine, StreamBase, and Microsoft’s StreamInsight.

Data cleaning
Decisions taken by trading algorithms rely on data; consequently, it is one of the most important drivers in AT. To enable a high quality of algorithms’ outputs, high quality input data needs to be ensured. Various techniques for data cleansing have been proposed over the years, with Tan et al.\textsuperscript{18} giving a general overview of appropriate techniques for dealing with the data; Diebold\textsuperscript{19} describing time-series analysis, including data cleaning applications; and Dacorogna\textsuperscript{20} providing in-depth analysis of financial data and adaptive (on-the-fly) filtering methodologies. Most popular techniques for data cleansing involve removal of duplicates and statistical outliers, standardisation or normalisation, as well as de-trending and seasonality removal, where applicable.

Having discussed ‘data’, we now look at flash crashes and rogue algorithms.
Flash crashes

The ‘infamous’ flash crash\(^\text{21}\) was a US stock market crash that took place on Thursday 6 May 2010, in which the Dow Jones Industrial Average plunged by about 1,000 points (about 9\%) only to recover those losses within 20 minutes. It was the second largest point swing, 1,010.14 points, and the biggest one-day point decline, 998.5 points, on an intraday basis in Dow Jones Industrial Average history. It wiped approximately US$600 billion off the US markets in approximately 10 minutes. However, this is just one – albeit the most famous – of a number of flash crashes, such as those occurring in India in 2012.

Market volatility

It is interesting to visualise the amount of high-frequency trading in the stock market from January 2007 to January 2012 – which is illustrated by the NANEX\(^\text{22}\) charts shown in Exhibit 5.4. The various grey tones, as identified in the legend, are all the different US stock exchanges. You might think there are only two stock exchanges in the US; however, while there are only two exchanges where stocks are listed, there are many more exchanges where stocks are traded. What we see here is relatively low levels of high-frequency trading through 2007 ramping up to 2011, where volumes can surge and fall back in pretty much random fashion.
Intriguingly there is still no definitive explanation of what caused the Flash Crash. The joint Securities and Exchange Commission (SEC) and the Commodity Futures Trading Commission (CFTC) report portrayed a market so fragmented and fragile that a single large trade could send stocks into a sudden spiral and detailed how a large mutual fund company selling an unusually large number of E-Mini S&P 500 contracts first exhausted available buyers, and then how high-frequency traders started aggressively selling, accelerating the effect of the mutual fund’s selling and contributing to the sharp price declines that day. While some companies exited the market, companies that remained in the market exacerbated price declines because they escalated their aggressive selling during the downdraft. High-frequency companies during the crisis, like other companies, were net sellers, contributing to the crash.

The joint report said prices stopped falling when at 2:45:28 pm, trading on the E-Mini was paused for five seconds when the Chicago Mercantile Exchange (CME) Stop Logic Functionality was triggered in order to prevent a cascade of further price declines. In that short period of time, sell-side pressure in the E-Mini was partly alleviated and buy-side interest increased. When trading resumed at 2:45:33 pm, prices stabilised and shortly thereafter, the E-Mini began to recover.

Source: Nanex, 'The rise of the HFT machines', 2012; www.nanex.net/aqck/2804.HTML
Early theories

Early theories regarding the 2010 Flash Crash include the following.\(^25\)

- **Fat finger theory** – a fat-finger trading error is a slang term for a typing mistake – in the case of the Flash Crash, the placing of an unusually large number of E-Mini S&P 500 contracts.
- **Impact of high frequency traders** – regulators found that high frequency traders exacerbated price declines; however, they also accelerated the rise once the recovery occurred.
- **Large directional bets** – regulators say a large E-Mini S&P 500 seller set off a chain of events triggering the Flash Crash.
- **Changes in market structure** – some market structure experts speculate that, whatever the underlying causes, equity markets are vulnerable to these sort of events because of decentralisation of trading.
- **Technical glitches** – an analysis of trading on the exchanges during the moments immediately prior to the flash crash reveals technical glitches in the reporting of prices on the NYSE and various ATSs that might have contributed to the drying up of liquidity.
- **Stop-loss triggers** – if the market price falls through the stop loss trigger price, then the order will be activated and the long position will be automatically closed out.
- **Inconsistent halting rules** – if the market falls by a certain percentage then the exchange halts trading. Unfortunately the halting rules across all the exchanges can be inconsistent, especially amongst regional exchanges in the US.
- **Stub quotes** – ultra-low bids order placed well off a stock’s market price. Stub quotes are used by trading companies when the company does not want to trade at certain prices and wants to pull away to ensure no trades occur. In order to make this happen, the company will offer quotes that are out of bounds.
- **NYSE delay** – NYSE went to ‘slow market’ on these stocks.
- **Quote stuffing** – attempt to overwhelm a market with excessive numbers of quotes by traders. This involves placing and then almost immediately cancelling large numbers of rapid-fire orders to buy or sell stocks.

Aftermath

In the aftermath of the 2010 Flash Crash the EU, governments and regulators have scrambled to understand the underlying causes of flash crashes and to put in places procedures to alleviate crashes, such as improved circuit breakers. Various proposals have been put forward as follows.

- **Circuit breakers:** the concept of ‘circuit breakers’ in trading refers to any of the measures used by stock exchanges during large sell-offs to avert panic selling. After an index has fallen a certain percentage, the exchange might activate trading halts or restrictions on program trading. For example, if the Dow Jones Industrial Average falls by 10%, the NYSE might halt market trading for one hour. There are other circuit breakers for 20% and 30% falls.
• Algorithm certification: another proposal to improve the reliability of trading algorithms is that algorithms should be ‘certified’ before use. This is considered infeasible due to time and complexity.

• Transaction tax: the aim of the so-called transaction tax on HFT is to deter spurious trades, and also ‘gaming’ the system by reducing the profitability of certain strategies that are common in HFT, which uses systems that can make trades in milliseconds. For example, to deter unwelcome practices in high-frequency trading, such as where consecutive orders are issued with less than 500 millisecond delay from each other, France is proposing a 0.01% charge on order cancellations if they are above 80% of total orders. There is also a 0.2% levy on all share transactions issued by companies with a capitalisation of more than €1 billion, as well as a tax on naked sovereign credit default swaps.

• Holding period: another proposal to discourage spurious trades is the idea of setting a ‘holding period’ for the buying of an asset. However, many practitioners believe this will in fact increase instability.

• MiFID II: the EU MiFID II, the second version of the MiFID directive, is intended to regulate high-frequency trading. According to the European Parliament’s website, all trading venues must ensure that trading systems are resilient and prepared to deal with sudden increases in order flows or market stresses. These could include ‘circuit breakers’ to suspend trading. Additionally, all orders should be valid for at least 500 milliseconds, that is, must not be cancelled or modified during that time. The second version of MiFID is intended to tackle some of the issues missed by the original document. It also addresses concerns over HFT. MiFID imposed new rules for equity trading across 30 European countries when it became law on 1 November 2007, but traders argued that it should be expanded. Some suggested regulators should focus on the over-the-counter (OTC) markets.

Model risk and rogue algorithms

There is no better example of a rogue trading algorithm than market-making company Knight Capital’s US$460 million trading loss in 30 minutes on 1 August 2012, when a technology glitch in one of their ATSs started pushing erratic trades through on nearly 150 different stocks, bombarding equity exchanges with erroneous orders. Besides the US$460 million trading loss, shares of Knight plunged by 60% from US$11 to around US$4, bringing the company close to bankruptcy.

The above example illustrates potential shortcomings with current practice, when a QUANT programmer develops a new algorithm, for example an alpha model for forecasting, they firstly back-test the algorithm on historic data, then trade it virtually and finally start trading ‘for real’ with small amounts of funds that are steadily increased as the algorithm ‘appears’ to work.

Algorithm or model risk

So called rogue algorithms can result from a number of issues. Algorithms, or the models they embody, are approximations of the real-world. If the programmer makes a mistake, the algorithm/model will fail to correctly model the real world and is prone to what is called
‘model risk’. Model risk might involve a small error in the specification or a programming error. Problems and inaccuracies may develop slowly and (in the extreme) suddenly ‘explode’ during a period of frantic trading. The most common model risk problems are:\(^2^9\) (i) inapplicability of a model; (ii) model misspecification; and (iii) implementation errors.

**Inapplicability of model:** QUANT programmers use computational statistics, machine learning, and complexity techniques for modelling, but the computational precision of these models can create a fool’s paradise. Unfortunately, as the saying goes: ‘if the only thing you have is a hammer, everything looks like a nail.’ The two common forms of model inapplicability are: (i) applying an inappropriate modelling technique to a problem; and (ii) misapplying a valid/appropriate technique to a problem. An example of the former might be using a simple neural network to attempt to forecast a complex financial time series.

**Model misspecification:** model misspecification is where the QUANT programmer has built an algorithm that fails to correctly model the real world. A common area of model misspecification is where many Value-at-Risk (VaR) models used as a measure of the risk of loss on a specific portfolio of financial assets perform poorly. ATSs also struggle to model systemic risk involving endogenous and exogenous risk.

**Implementation errors:** the third area, and probably the most common, are programming errors in building the AT algorithms. Implementation errors cover: (i) single algorithm errors – misprogramming a specific algorithm, such as a simple case of sign reversal; and (ii) system errors – where algorithms might be running on different servers that fail to synchronise; for example, when one server goes down and the other continues to trade with erroneous values.

**Algorithm behaviour**

One of the greatest areas of concern, and the one least understood, is the behaviour of individual trading algorithms in the marketplace, or the complex interactions between multiple algorithms.

**Single algorithms**

An algorithm frequently embodies a trader’s trading strategy and concerns, such as the time horizon considered: high-frequency, intraday or long-term. For example, with the high-frequency approach, traders attempt to benefit from extremely short market inefficiencies with bursts of high frequency transactions. In this approach, the traders look for inefficiencies in pricing, volumes and liquidity with positions being held for no longer than few seconds. With an intraday approach, a few dozen transactions might be executed to adjust and maintain positions throughout the day. Such positions are then closed at the end of every day. In this approach, traders usually look for daily trends and news announcements from which they can benefit. Finally, with a long-term portfolio asset allocation, traders maintain positions on the basis of long-term trends, social sentiments and fundamental economic prognosis. Clearly the effectiveness of the algorithm is closely tied to the trader’s ability and knowledge.
**Algorithm-algorithm interaction**

The biggest concern with regard to algorithmic trading is ‘systemic risk’; namely that the unforeseen consequences of the interaction of a large number of algorithms could cause the collapse of the financial system. The number of financial transactions controlled by algorithms has risen significantly over the last few years. Electronic exchanges provide programmatic access through which algorithms can freely interact; creating sophisticated networks of actions and reactions, where the behaviour of such networks is unpredictable. In addition, the speed of interaction between trading algorithms is beyond the human reaction threshold and consequently beyond human control. This forms new types of systemic threats for the entire financial system. Since 2008 it has become a governmental concern to manage systemic risk and consequently to model and control algorithms and the way they interact together.

While governments are concerned about systemic risk, financial institutions are concerned about the interaction of algorithms within a single company, where the algorithms may be competing destructively.

Typical causes of risks in algorithm-to-algorithm interaction may be related to:

- **algorithm control** – lack of control over algorithms either due to the previously mentioned speed of processing or due to the complexity of decision-taking/analytic process in the algorithm; where all the possible outcomes may not be comprehended by a quant and consequently tested;
- **algorithms execution process** – which by its collaborative nature is dynamic and relies on multiple factors, including money management (asset allocation) as well as availability of securities, gateways exchanges, and the underlying protocols and network infrastructure;
- **implementation errors** – faulty coding of one or more elements of the algorithm, which due to the feedback element present in most trading algorithms may be only visible after a long period of time (that is, due to error aggregation in variables);
- **faulty infrastructure** – problems caused by damaged hardware infrastructure, non-compatible libraries, and non-optimal work of services or operating systems; or
- **malicious/rogue trading practices** – encoded as part of trading algorithms including, for example, flash trading.

Modelling and prevention of algorithmic (and systemic) risks is particularly hard due to the complexity, scale and distributed nature of the problem. One potential approach is multi-agent modelling of financial systems and economic phenomena, including graph/network theory to shape policy-making processes of such systems.

**Algorithm testing**

Even before implementation, let alone use in real trading, an algorithm strategy is thoroughly back-tested. As discussed, back-testing is a specific type of historical testing that determines the performance of the strategy if it had actually been employed during past periods and market conditions. While back-testing does not allow one to predict how a strategy will perform under future conditions, its primary benefit lies in understanding the vulnerabilities
of a strategy through a simulated encounter with real-world conditions of the past. This enables the designer of a strategy to ‘learn from history’ without actually having to make the strategy with actual money.

If the algorithm performs satisfactorily during back-testing it is then implemented in a trading system and run ‘virtually’ alongside real trading to see how it might perform in current market conditions. If the algorithm appears to perform satisfactorily, then the algorithm is allowed to trade for ‘real’ but with a modest amount of funds. As confidence builds in the trading algorithm then the amount of funds are increased. This might appear to alleviate risk, however the actual market microstructure might change fundamentally from day to day, and render the algorithm loss-making or susceptible to risk.

Algorithmic trading, IT system architecture

To understand ATSs from an IT perspective, it is useful to introduce the trading floor architecture within which an ATS operates.

Trading platforms can be considered three-tier systems, where the first tier consists of external communication elements, the second tier is responsible for internal communication and, finally, the third layer provides the actual business logic. Cisco’s Trading Floor (Exhibit 5.5) presents a good representation of a typical architecture of a trading system.

The bottom layer of Cisco’s trading floor represents the wider, external communication layer and the low level hardware and network architecture of the platform, as well as the external systems with which it exchanges information. The performance of the platform depends on many interrelated factors but the most important one is the speed of the underlying hardware and networks. Consequently, when considering the architecture of this type of financial system, the emphasis needs to be put on the hardware and connectivity. Apart from the hardware and network, the operating systems on which the platform relies must also be chosen with care. In the majority of cases the platform is distributed amongst many servers in a cloud; the operating system on the servers needs to be able to deliver such distributed functionality. When speed is of the essence (and in majority of cases it is crucial), the customisation of the underlying hardware, software, and network architecture is an option to consider. One can consider dedicated hardware for specific needs of a fast execution, storage or multi-processing. Kernels of utilised operating systems can be tailored to specific needs of the underlying hardware as well as the platform. All the unnecessary drivers may be removed, and the kernel may be optimised and recompiled to support a particular need. The network architecture may also be optimised by reorganisation of its structure, collocation of services with markets and exchanges, and also by utilisation of dedicated connection lines. In addition to the architecture and organisation of the network, the communication protocols also play a major role; they need to be able to transport information – quite often a substantial volume of it – in the quickest possible way, with a guaranteed delivery.

Cisco’s trading floor also defines the external communication layer from the data-streams perspective, and it specifies the ‘market data’ stream and ‘trading orders’ stream. The ‘market data’ stream, to quote Risca et al., carries pricing information for financial instruments, news, and other value-added information such as analytics. It is unidirectional and very latency sensitive, typically delivered over UDP multicast. It is measured in updates/second.
Exhibit 5.5

Cisco's trading floor architecture

High performance trading cluster and ticker plant

- Price engine
- Algorithmic trading
- Risk modelling
- Order management
- Execution monitors
- IP phone alerting
- Trading systems

Message bus

- Application virtualisation
- Data virtualisation
- OS virtualisation
- Storage virtualisation

Services

- Latency monitoring
- Quality of service

Market data providers (exchanges, market data aggregators, ECNs)

AON-instrumented high availability multicast network

End-user applications

- Performance dashboards
- Compliance surveillance
- Mobile alerting
- IP phone alerting
- Performance dashboards

and in Mbps. Market data flows from one or multiple external feeds, coming from market data providers like stock exchanges, data aggregators and ECNs. Each provider has their own market data format. The data is received by feed handlers, specialised applications which normalise and clean the data, and then sent to data consumers, such as pricing engines, algorithmic trading applications, or human traders. The ‘trading orders’ stream is a type of traffic that carries the actual trades. It is bi-directional and very latency sensitive. It is measured in messages/second and Mbps. The orders originate from a buy side or sell side company and are sent to trading venues like an exchange or ECN for execution.

**Architecture layers**

The top layer of Cisco’s trading floor describes the logic behind all the modules of the platform. Typically, this includes: the connectivity engines, data aggregation units, data processing elements, order routing, execution elements, trading strategy elements, graphical and programmatic interfaces, monitoring tools and many more, depending on the requirements and specification.

The middle layer of Cisco’s trading floor architecture is the service layer that allows internal communication within the platform. The internal communication system is the backbone of every platform. It needs to be reliable and fast; it also needs to be flexible and modular. The internal communication system is necessary as a result of the distribution of elements of the platform. A modern approach to development of trading platforms incorporates both the service and event-oriented methodologies.

From the functional point of view, the internal communication layer of an enterprise trading environment can be divided into three types of application components. These are: (i) the components that publish information; (ii) the components that subscribe to information; and (iii) the components that both publish and subscribe. Every middleware communication bus provides multiple channels that allow communication of publishers with subscribers. The channels make sure that a group of subscribers receives only the relevant information provided by a group of publishers connected to a particular channel.

**Core modules**

*FIX engine module:* the most common format for the order and market data transport is the FIX protocol. The applications which handle FIX messages are called FIX engines. They are designed to send/receive high volumes of FIX messages, translate the messages to relevant internal formats of the architecture and interface to all the relevant modules of the platform. The FIX engine must be able to handle multiple streams at once, and be able to deal with the reconnections, re-quotes and message synchronisation issues. The FIX engine also needs to be capable of handling different versions of the FIX protocol and non-standard FIX messages customised by the information providers.

*Feed handlers:* when the market data is handled by the FIX engine and passed to the rest of the elements of the system, it needs to be cleansed, processed and aggregated. The Feed handler module provides functionalities to clean the data from outliers, gaps and other statistically significant problems to make sure that it is of high quality. The Feed handler
module is also responsible for processing of data to provide data analytics, it also provides aggregation facilities to store the cleansed and processed data in the database, for a future use by other elements.

**Price engine:** the goal of the engine is to allow an automatic pricing of securities, using inputs from different sources and adopting financially consistent pricing models. The price engine needs to be designed to support concurrent, real-time calculations for hundreds of assets at the same time. The engine relies on financial models and they in turn need to allow only a minimal number of market operations to efficiently control a large amount of securities. The engine processes the prices from different sources and performs data filtering to obtain an internal best price upon which all subsequent calculations are based. When the reference securities change their market prices, the engine handles the logical dependencies, sending recalculation signals to the correct models and instruments.

**Risk modelling:** the risk engine is a management system designed to identify the risks associated with a given set of assets. The engine provides a set of comprehensive measurement tools to assess, control and communicate risk. It incorporates advanced statistical analyses, proprietary valuation models provided by the price engine and advanced data aggregation provided by the feed handler module.

**Order management:** the engine is used for rapid order entry and processing, it facilitates and manages the order execution, typically through the FIX engine. The order management systems often aggregate the orders from different channels to provide a uniform view. The engine allows input of single and multi-asset orders to route to the pre-established destinations. Apart from issuing the orders, the engine also allows changing, cancelling and updating orders. Additional functionality of the engine is the ability to handle execution reports and to access information on orders entered into the system, including details on all open orders and on previously completed orders and general profit and loss (P&L) information.

**Algorithmic trading:** the module supports the use of trading strategies that issue and manage orders with use of an embedded logic coded within the algorithm of the strategy. This allows the logic of the algorithm to decide on aspects of the order such as the optimal timing, price or quantity of the order. The module supports a variety of strategies, including market making, inter-market spreading, arbitrage and directional or pure speculation. The investment decision and implementation may be altered at any stage and is usually supported by analysis of data provided by the feed handlers, the price engine, the risk modelling and the order management. The module plays the role of a manager of all the handled strategies, providing functionality to enable/disable selected strategies and modify factors of the strategies on the run.

**Trading systems and alerting modules:** these modules provide a means of manual trading by allowing the users visual or audio access to the platform. The two module types describe (in visual or audio ways) the states of the markets and selected assets, and also provide a graphical representation of analytical data (in case of the trading systems). The trading systems module also provides graphical interfaces to execute orders and maintain positions on the market. This module type is characterised by strong charting capabilities and functionalities that allow users a quick response to different states of the markets. The alerting module is an extra means of communicating with clients and is usually not part of the architecture, but needs to be considered as a part of a trading floor.
Execution monitors, performance dashboards and compliance surveillance: these modules allow monitoring of the architecture’s performance. The modules also allow the compliance of the performance and behaviour with predefined policies. This is to reduce the risk of malfunction of modules of the architecture and allow quick response to threats and issues within the architecture.

High-frequency and ultra-low-latency IT systems

High frequency and ultra-low latency IT systems are two tightly coupled technological approaches that imply minimal human traders’ interaction with markets.

HFT algorithmic trading systems

High frequency trading is a technological concept to trade securities on a basis of very frequently issued order instructions, allowing instant modification of positions held in traded securities. The HFT computerised quantitative models are largely driven by their ability to simultaneously process volumes of information. In HFT, a traded position is typically held without modifications for only fractions of a second with the majority of positions being traded on a binary (open/close) basis. HFT is inherently an algorithm-based, quantitative trading style with order instructions modifying positions thousands or tens of thousands of times a day. HFT systems compete for very small, consistent profits relying heavily on the processing speed of their trades, and on their access to the market.

Typical HFT strategies would involve statistical arbitrage as well as variations of hedging and momentum approaches.

Ultra-low latency algorithmic trading systems

Ultra-low latency trading systems are a technology concept where speed of order-instruction transfer is critical for the purpose of best execution. It is a means of executing trading flow by bypassing intermediary latency obstacles, where market co-located trading algorithms transmit orders via a direct access into the market’s matching engine for execution. Necessary prerequisites in most circumstances involve ‘direct market access’,32 collocation, dedicated network line, customised hardware/software infrastructure, and stripped-down algorithmic trading strategy.

The IT architecture for the low latency system involves a stripped-down algorithmic trading strategy being integrated directly with connectivity engine. Typical ultra-low latency trading strategies would involve arbitrage (price arbitrage, event arbitrage, triangular arbitrage and others) and mean-reversion cross-trading.

UCL’s ATRADE platform

To give the reader a detailed understanding of an algorithmic trading system, we present the system architecture of the UCL ATRADE platform, which follows the CISCO architecture described above. As discussed, for the past eight years UCL has worked with the major
investment banks and funds developing algorithmic trading systems, and more recently with the regulators investigating high-frequency trading risk and systemic risk. To support this work we have developed our ATRADE platform which can be used both for virtual and real trading research. ATRADE is described to illustrate the structure of ATSs, which is difficult to find in the literature.

ATRADE has the following properties.

Simulation and real trading: the platform allows users to trade virtually or with real money, with use of a dedicated application programming interface (API) able to support algorithmic trading. To provide such capability the developers of the platform have designed a set of interfaces that allow issuing and maintenance of orders, extraction and processing of market data to generate trading signals, extraction and processing of order data, and finally management of information related to P&L and users’ account information.

Rapid prototyping: the platform is a framework for developing, trading, testing, and evaluating the ‘algorithm’, and especially its risk. The majority of ATRADE users come from academia and are relatively inexperienced in the field of the algorithmic trading. Consequently, to boost their experience they need a framework for development, trading, testing and evaluation of their algorithms. Rapid prototyping requires usage of a simple programming language and also simple and powerful interfaces that provide as much embedded functionality as possible without a need of low-level programming. ATRADE meets all the requirements.

Data processing and aggregation: the platform is capable of aggregating and processing data in real-time. Information is a key to success in algorithmic trading, therefore, ATRADE aggregates and processes data in real-time and delivers it to the users for analysis. Apart from delivering the data in real-time, the platform allows retrieval of historical data through a set of dedicated interfaces.

R and Matlab: the platform provides interfaces to the R and Matlab statistical/mathematical computing environments. Statistical and mathematical computing concepts are important for modelling and implementation of algorithmic trading strategies. Therefore, incorporation of the two environments may simplify many, otherwise complex, objectives set by the users. The two environments may be incorporated with ATRADE by the users and utilised in the functionalities that analyse data, generate trading signals, evaluate risk and many more.

Black box and multiple models: a particularly useful functionality of the platform is a module for automated evaluation of risk of ‘black box’ models that allows ranking of the models on such a basis; without the need of handing the source of the models to anyone. This feature is useful in eliminating the intellectual property rights (IPR) issues and ensures safety of the algorithmic trading models from potential third party threats. The platform supports automated evaluation of multiple models concurrently, and also enables generation of statistical-performance reports for every evaluated model.

Secure remote access: ATRADE provides an API that allows remote access to the major functionalities of the platform. This is to provide the users with a simple and easy way of access, allowing users to work remotely with the API. This involves making sure that the developed models and strategies are able to securely communicate with the platform.
ATRADE components

The ATRADE platform (Exhibit 5.6) consists of a set of distributed, multi-threaded, event-driven, real-time Linux services communicating with each other via an asynchronous messaging system. The platform allows multi-user real and virtual trading. It provides a proprietary API to support development of algorithmic trading models and strategies. It allows an advanced trading-signal generation and analysis in real-time, with use of statistical and technical analysis as well as the data mining methods. It provides data aggregation functionalities to process and store market data feeds. Finally, the platform allows back and forward testing of trading strategies. The following modules of the platform summarise the undertaken work.

Exhibit 5.6

ATRADE platform and modules

Source: Authors’ own

Back-end services: this module provides the core of the platform functionalities. It is a set of services that allows connection to a set of brokers and data providers, propagation processing and aggregation of market data feeds, execution and maintenance of orders, and management of trading strategies in a multi-user environment.

Front-end GUI client: this is a graphic user interface (GUI) that allows visual management of trading strategies. It allows the user to visualise market data in the form of charts, provides a market watch grid with a capability of manual trading and modification of positions generated by trading models, provides a dynamic visualisation of an order book and allows users to observe information on events relevant to the user’s environment.
Front-end competition evaluation module: this module supports trading competitions and allows organisations to monitor/display the trading results of users utilising the platform. The framework is a web-based graphic user interface. Its functionalities allow the generation of a ranking of users (on the basis of statistics of performance of their trading models) that allows classification of users. It also provides a dossier with records of a selected user.

Front-end API client: this provides a set of interfaces that can be used via a user’s favourite integrated development environment (IDE) to implement and test trading models. The API provides strategy templates to simplify access to some of the interfaces and defines generic structure of the strategies. The API also provides means of accessing market data, issued orders, and P&L of performed trades.

The business logic of the ATRADE is similar to the one proposed on Cisco’s trading floor. The following elements can be considered as the platform’s modules.

Connectors: this functionality provides a means of communication with the outside world; with the brokers and the data providers. Each of the outside venues utilised by the platform has a dedicated connector object responsible for control of communication. This is possible due to the fact that each of the outside institutions provide either a dedicated API or is using a communication protocol that is, the FIX protocol. The platform provides a generalised interface to allow standardisation of a variety of connectors.

Internal communication layer: the idea behind the use of the internal messaging system in the platform draws from the concept of event-driven programming. The trading platform utilises events as a main means of communication between the elements within the platform. The elements, in turn, are either producers or consumers of the events. The approach significantly simplifies the architecture of the system while making it scalable and flexible for further extensions.

Aggregation database: this provides a fast and robust DBMS functionality and enables the platform to store, extract and manipulate large amounts of market data. The storage capabilities of the aggregation element not only allow a back-testing of the strategies with use of historical data but also enable other, more sophisticated, tasks related to functioning of the platform, including data analytics on the basis of market data, data analytics on the basis of trades, variety of risk analysis, evaluation of performance of trading models and many more.

Order management engine: this provides four main functionalities to the platform. First, it allows execution/maintenance of orders triggered by trading strategies. Second, the orders are synchronised with the selected brokers. Third, it plays a role of an order buffer, providing a quick access to active and historical orders. Finally, it enables P&L as well as statistical evaluation of orders with the omission of brokers.

API client: this is an API that enables development, implementation and testing of new trading strategies with use of the developer’s favourite IDE. The API allows connection from the IDE to the server-side of the platform to provide all the functionalities the user may need to develop and trade.

Shared memory: this provides a buffer-type functionality that speeds up the delivery of historical data to strategies and the analytics-related elements of the platform (that is, the technical analysis library of methods), and, at the same time, reduces the memory usage requirement. The main idea is to have a central point in the memory (RAM) of the platform.
that will manage and provide the historical data of performance of the financial instruments (or mainly a time-framed selection of historical data from the current point of history up to a specified number of timestamps in history). Since the memory is shared, no trading strategy will have to keep and manage the history of instruments required by the strategy. Moreover, since the memory is kept in RAM rather than in the files or the DBMS, the access to it is instant and bounded only by the performance of hardware and the platform on which the buffers work.

**Strategy templates:** the platform supports two generic types of strategies; push and pull. The push type registers itself to listen to a specified set of financial assets during initialisation, and the execution of the strategy is triggered each time a new data feed arrives to the platform. This type is dedicated to very quick, low-latency, high-frequency strategies and the speed is achieved at the cost of small shared memory buffers. The pull strategy type executes and requests data on its own, based on a schedule. Instead of using the memory buffers, it has a direct connection to the database and hence can request as much historical data as necessary, at the expense of speed.

**Risk management in ATRADE**

Next we examine the ATRADE approach to the vital area of ‘risk’. To manage algorithmic risk in the platform we attempted to identify, evaluate, and prioritise vulnerabilities and threats. We distinguish four different risk management layers: (i) market provider layer; (ii) connectivity layer; (iii) order management layer; and (iv) strategy layer. We evaluate the layers from the perspective of the risk they present to the outer, surrounding layers.

**Market provider layer:** consists of exchanges, clearing houses and brokers; each differently affecting risk characteristics. The layer is not as such part of trading platform but is part of a larger marketplace system.

The algorithmic risk on this layer is related to the provisioning of data to the outer layers (for example, connectivity, order management, and so on), provisioning of order instructions issued by the outer layers and management of positions and accounts affected by the issued instructions. In algorithmic trading the brokerage companies are responsible for providing access to exchange and clearing facilities. Their main concern is, therefore, on persistence of connection between their facilities and clients’ systems, availability of information the clients may require and access latency of requested information. The clearing companies, from the perspective of AT, accept responsibility for transaction (insolvency) risk and updating the client on the state of his/her positions and accounts. In worst case scenarios, the clearers suspend acceptance of users’ margin-increasing order instructions and, depending on the current policy, liquidate part or the entirety of the user’s active positions. Furthermore, the clearers are likely to hedge active positions held by their users. This procedure is prone to users’ assets abuse, especially when clearers actively participate in trading to minimise their own hedging loses. Consequently, measures on hedging actions should be applied to make such actions transparent and accountable to the users. Finally, the exchanges in AT match and execute order instructions and, consequently, are required to ensure fair access to best asset prices, to liquidity and information.
Connectivity layer: one of the key algorithm risk threats is an inability or lack of persistent communication of outer layers with the market provider layer. Persistent communication between the layers facilitates continuous stream of information with the most recent security prices, but most importantly enables the outer layers to receive order-status updates (called execution reports) without which trading algorithms lose control of their P&L status. The algorithmic risk, in case of non-persistent connection, arises from the possibility of delayed, and consequently mispriced, orders and from the possibility of loss of information (that is, on the state of currently held positions, due to missed flow of execution reports). If connectivity problem happens to propagate on a larger scale, it may cause serious systemic risk where large amount of misinformed trading algorithms may attempt to trade, consequently destabilising parts or the entire system.

ATRADE’s connectivity layer enables the platform to establish persistent sessions with data providers and brokers, clearing houses and exchanges. If such persistent connection is not possible due to wider network infrastructure problems the platform is capable of providing information about the problem to the outer layers, consequently giving them the chance to react and suppress potential issues.

The platform, we believe, provides a diversified approach to account management that aims to reduce insolvency risk through separation, diversification, and on-the-fly account management. ATRADE allows one session to an account per one trading-algorithm. Such accounts, held with a clearing house, are typically sub-accounts of a global account of a trading party, with an assigned fraction of the global funds. This ensures control of trading algorithms over their accounts (and positions held on the accounts), simplifies P&L calculations and enables risk managers to control accounts of particular algorithms. Multiple sub-accounts, one per algorithm, enable diversification of risk, giving more chance of capital preservation in case one of the algorithms go rogue. Finally, this approach enables utilisation of meta-algorithms for hedging sub-account positions against each other on a portfolio-basis.

Order management (OME) layer: is the key P&L risk management layer in the ATRADE platform. ATRADE’s integrated order management engine encompasses the following functionalities: (i) trading venue availability information, including securities available for trading on the venue; (ii) security information, including current best bid/offer prices and volume availability; (iii) order status, including active and historic order information; (iv) P&L calculator, enabling position and account status calculations; (v) margin allocation functionality, with basic account warning elements; and (vi) P&L management functionality, with dynamically set policies making it possible to keep account statistics within set thresholds.

Strategy (SDK) layer: on this layer the risk management functionality is incorporated as part of every strategy in the platform, with quants and quant developers bearing responsibility for the risk characteristics of the layer. ATRADE assumes that strategies will have advanced risk measures as part of their architectures and will use outcome of the measures in their decision rules and control mechanisms. Such coupling enables testing, simulation and optimisation of strategies together with their risk mechanisms to fine tune the algorithm. In order to enable functioning of the advanced risk metrics, ATRADE provides to each strategy a constant stream of account status updates as well as execution reports. Account updates contain P&L statistics of an account including margin information as well as status of all the positions held in the account.
Risk management between strategy and OME layers

To minimise the potential impact of a rogue algorithms, risk management is provided between the strategy and OME layers. The OME’s trading venue availability information and security information functionality enables programmatic requests from the strategy layer on the status of connections to exchanges as well as information about available traded securities. Such functionality is useful in minimising risk while cross-trading between various exchanges and in obtaining availability of securities while trading in securities with various maturity dates. This is useful for algorithms that require information on whether particular exchange/security is available at the time.

OME’s order status functionality is designed to be used by the strategy layer in the process of activation or reactivation when strategies need to rediscover the status of their orders to establish their risk and money management characteristics.

Based on the securities information and order status functionalities, the OME is able to provide P&L status provisioning in its functionality. The ability to provide P&L status of a user’s account and currently held positions on the account is important for the P&L risk management in the strategy layer. Moreover, based on the P&L calculator (as described in point (iv) above) and the fact that all the order instructions and execution report traffic goes through the OME, the engine is able to provide margin allocation functionality (as described in point (v) above) designed to warn and prevent the strategy layer from bankrupting the account.

The P&L management functionality is a logical extension of the margin allocation functionality, enabling the strategy layer to fine tune thresholds and responses of OME to potential P&L risks.

Risk management policies

As previously mentioned, ATRADE SDK enables algorithms to set risk policies that reside in OME that controls execution or order instructions and provisioning of account update information to algorithms. With respect to risk management, SDK enables an algorithm to: (i) get active policies of an account; (ii) set active policies for an account; (iii) send an order instruction; (iv) receive execution reports for issued order instructions; and (v) receive account updates with warning flags.

OME can handle multiple policies per account where available policy-type enables it to set: (i) allowed maximum amount of accepted orders within defined timeframe (for example, accept maximum 10 orders per second); (ii) allowed maximum traded quantity within defined timeframe (for example, allow trading of maximum 100 lots per second); (iii) allowed maximum amount of positions in different securities (for example, open positions possible in only three different securities at a time); (iv) allowed maximum size of position in particular security (for example, given position can only use 5% of available margin for the account); and (v) allowed minimum remaining margin on the account, in percentages (for example, remaining available margin for the account cannot fall below 10%).

An algorithm can also set a policy level for a particular policy type with four available levels: conservative, liberal, social and custom. The OME actions with respect to a particular
level will vary, and will depend on the algorithms setting policy instruction for a level. Key available policy instructions are: (i) reject current order instruction and do not accept any further new/modify orders; (ii) liquidate oldest element in positions with a market order until threshold is no longer meet; (iii) liquidate all open positions; and (iv) set a warning (social) flag in an account update but do nothing.

In case the threshold of a particular policy has been passed, OME will execute the policy-instruction and will send an account update with policy identifier, a policy-level flag appropriately set and textual information on why the flag was set. This enables algorithm with flagged account to react to the risk.

**Risk management responsibilities and ATRADE risk layers**

It is useful to introduce the so-called stakeholders of algorithmic trading systems. The stakeholders can be loosely defined as follows.

- **Quantitative analysts** – the creators of different AT models and measures of AT risk and performance characteristics, they work mainly on applications of complex numerical methods (statistics, financial mathematics, econophysics, machine learning) in finance. Quantitative developers (programmers) are a particular type of ‘quants’ specialising in low-level implementations of numerical methods.

- **Traders** – are active market participants specialising in various investment methodologies involving tradable securities. AT traders utilise work of quants to automate one or more stages of their decision-making, trading, and evaluation process.

- **Trading floor managers** – are responsible for managing larger groups of traders and their strategies. Responsibility of this stakeholder is to organise and co-ordinate work of the trading floor and to ensure optimal performance of the group with respect to some predefined goals and quantitative measures evaluating the goals.

- **Portfolio managers** – are responsible for management of larger portfolios of tradable securities on the basis of longer investment horizons. While traders typically engage in arbitrage, speculation and high-frequency or intra-day trading, portfolio managers work on the basis of stable, long term returns with minimal appetite for risk.

- **Risk managers** – are not actively engaged in the investment process but are rather assessors of different types of risks for traders, portfolio managers and trading floor managers. While traders and portfolio managers often work with quants on AT models, risk managers engage with quants on building various risk measures.

We recognise quants as the key parties interested in the development process while traders, portfolio managers, risk managers, and trading floor managers are key parties interested in usage of systems created with our development process and logical framework. Consequently, quants, traders and risk managers are responsible for risk management on the SDK layer, traders, portfolio managers and risk managers are responsible for risk management on the OME layer, and finally, trading floor managers and risk managers are responsible for management of risk on connectivity and market provider layers.
Quant, IT and policy-making challenges

It is also appropriate to discuss the various ‘challenges’ that exist in developing reliable systems.

Algorithmic trading impact

There have been a number of Government studies on the impact of computer-based trading (CBT) and HFT trading\textsuperscript{34,35} which have reached broadly similar conclusions.

- **increased volatility** – no evidence that CBT/HFT increases volatility or instability;
- **increased risk** – occurrence of ‘risk’ events (for example, price jumps) become accepted as normal, but can be off-set through circuit breakers;
- **rational behaviour** – algorithms reduce the irrational trading behaviour of human traders;
- **risk drivers** – underlying algorithmic rules may overreact and increase risk during periods of volatility; and
- **algorithm instability** – heterogeneity in HFT strategies may cause catastrophic instability of the entire financial systems (for example, the collapse of the economic ecosystem).

Policy issues to be addressed are as follows.

- **What are the effects of CBT on the markets** – this includes measures such as liquidity, transaction costs, price discovery, and so on?
- **Impact on market making** – this covers new providers such as HFT traders and new technologies such as ultra-fast HFT and co-location.
- **Liquidity supply** – the argument is that CBT/HFT liquidity is ‘transient’ and evaporates during a crisis.
- **Unproductive messaging** – order cancellations, revisions, and price requiting leads to misinformation and possibly market abuse (see below).
- **Flash crashes** – it is generally agreed that CBT/HFT did not trigger the 2010 Flash Crash, but their responses to the unusually large selling pressure on that day exacerbated market volatility.
- **Market abuse** – CBT facilitates market manipulation, such as quote stuffing, layering and spoofing.
- **Appropriate legislation** – given the above negative views of CBT: what are the appropriate definitions of CBT-market abuse and how should they be addressed, where overregulation may be more destructive than under regulation?

Various challenges in algorithmic trading and risk can be differentiated either on a quant/analytical level, IT level or regulatory/policy-making level: (i) we recognise that handling scale in terms of frequency, latency, big data and model quantity is one of such challenges; (ii) we also recognise validation, evaluation and testing of received (and generated) data, utilised models and trading strategies to be an important challenge; (iii) we recognise importance of sound principles, procedures and policies in quantitative/analytical applications, in IT
infrastructure, in brokerage/clearing/exchange policing; and (iv) finally we recognise a need for sound financial/economic modelling and systemic/algorithmic policies.

Handling scale – scale is an important issue due to the growing amount of data required to be analysed, the frequency of appearance of data, and consequently the frequency of issued order instructions, and the latency required to be ensured by the hardware and software infrastructure-supported trading and analytic processes. All the data and analytic processing outcomes need to be stored and available for further processing or simulation, hence storage and computing infrastructure needs to be able to cope with such requirement. Finally, the number and complexity of trading algorithms continuous to grow; requiring more resources, introducing more degrees of freedom and as a consequence posing more systemic threats.

Validation, evaluation and testing – another challenge in algorithmic trading and risk is presented by the processes of validation, evaluation and testing of data, models and algorithms. Data validation/evaluation is required to ensure good quality of results on algorithms utilising the data, therefore elements of trading architecture responsible for data provisioning should consist of in-built algorithms providing data filtering and data quality evaluation that would provide meta-information on the state of requested information. Such algorithms should be able to handle high frequency data in real-time (guaranteed response within time constraints) and should enable algorithms using the data to make a decision about whether provided quality of data is acceptable. Model validation/evaluation is considered a semi-manual process performed by quants when a new model is created. The quality of a model depends on how well the mathematics behind a model (the theoretical side) agrees with results of repeatable experiments (measurements). Therefore, it is crucial to evaluate whether or not a given mathematical model describes a system accurately in all circumstances. Given the complexity of trading models, it is challenging to ensure models do what is expected of them in all circumstances. Algorithm validation and evaluation is a functional and non-functional process of ensuring that the code comprising an algorithm is correct with respect to the set requirements. Model validation and evaluation should happen periodically and, although challenging, this process should be automated as part of an SDK of a platform enabling algorithm to trade.

Sound principles, procedures, and practices – sound software engineering practices should be enforced in the process of development, testing, deployment and administration of quantitative applications and IT infrastructure, and in the process of policing the brokerage, clearing and exchange organisations.

The quants should have a full understanding of the data features they are using, should always model behaviour of applied processes, and comprehend and respect boundary conditions of their models. While developing IT infrastructure and algorithms, software engineers should apply good practices (for example, test-driven development, continuous integration, model-based development, agile development, self-documentation and code-reviews) as well as perform periodic conformity tests.

The brokerage, clearing, exchange organisations should not only follow the established technical guidelines, principles, and procedures, but should be transparent on how the guidelines are implemented. Given our experience with ATRADE, we also believe that such organisations should be obliged to provide programmatic access to users’ account, position
and P&L characteristics to enable propagation of such information from the point of their origin to trading algorithms and functionalities that require such information to minimise algorithmic risk.

**Sound financial and economic modelling** – although AT is still a relatively new technology, national banks and financial authorities should adopt a more scientific approach to their legislation process\(^{37,38}\) to ensure their policies actually improve trading and reduce (systemic) risk. Scientific fields of systems theory, cybernetics (the study of self-governance) and control theory offer mature approaches and tactics applicable to financial and economic systems. Such approaches, combined with financial computing (for example, computational simulation techniques), offer possibilities of exploring regulatory systems, their structures and constraints; giving potential solutions to policing of financial systems.

Various publications on the subject consider the following new proposals.

- **Notification of algorithms**: investment companies engaged in AT would be obliged to periodically provide description of their strategies to the home regulator. This is generally considered unworkable.
- **Minimum resting time**: a minimum holding or resting period for contracts before they can be closed.
- **Larger tick sizes**: there are arguments to increase tick sizes. The two main effects of a tick size are that it: (i) creates a floor for the bid-ask spread; and (ii) impacts the price-time priority rules of the limit order book.
- **Financial transaction tax**: a financial transaction tax (FTT)\(^{39}\) is a tax levied on transactions in financial assets, but can also be imposed on unfilled orders, derivatives, repos, currency trades and any financial transaction.
- **Order-to-execution ratio (OER)**: defining a trading ‘ratio’ to reduce excessive message traffic and cancellation rates, or imposing a minimum transactions to order ratio. This means that any trader would need to participate in a trade before they could submit a new order once they hit the rule limit.
- **Batch auctions**: this is an alternative trading mechanism that can be based on periodic auctions, which can be designed to minimise the advantage of speed and to mitigate other negative outcomes of the continuous trading model, such as manipulative strategies.
- **Pro-rata instead of time-priority**: current CBT/HFT operates under a price-time priority auction, which gives a huge advantage to market participants that can leverage speed. An alternative is a ‘pro-rata auction’; namely sequential call auctions at random times, for example, roughly once a minute that only provide limited indicative price information.

**Lessons learned**

As discussed, algorithmic and high-frequency trading is growing rapidly, accounting for as much as 73% of US trading volume in 2009. However, it is an area of major concern due to the unforeseen behaviour of ‘rogue algorithms’ and catastrophic ‘Flash Crashes’, which have the potential to bring down the financial markets. From both a company and regulatory perspective, the behaviour and risk of AT systems and individual trading algorithms
is poorly understood, and the effect of industry regulation through legislation may in fact increase risk and instability.

**Market impact**

Regarding CBT and HFT in general, our view is that it should be considered the ‘industrialisation of trading’ and definitely not a transient phenomenon. Although there is considerable negative publicity about CBT we believe it is generally beneficial to market quality:

- **liquidity**: CBT/HFT has contributed to liquidity when measured by bid-ask spreads and so on;
- **transaction costs**: costs both for institutional and retail investors have reduced; and
- **market efficiency**: CBT/HFT has driven improvements in information and infrastructure efficiency with better prices, analytics and market integration by exposing already existing inefficiencies.

**Company impact**

The main company concerns are algorithm or model risks (for example, rogue algorithms) which can result from a number of issues that need to be increasingly addressed by companies:

- **inapplicability of model**: applying an inappropriate modelling technique to a problem; or misapplying a valid/appropriate technique to a problem;
- **model misspecification**: where the QUANT programmer has built an algorithm that fails to correctly model the real world;
- **implementation errors**: programming errors in building the AT algorithms; and
- **algorithm interaction**: the interaction of trading algorithms within a market or even a single company is poorly understood.

**Regulatory impact**

A number of regulatory changes for CBT/HFT are being implemented or under consideration both to control systemic risk and what is seen as market abuse. These include:

- **financial transaction tax**: this is a tax levied on transactions;
- **minimum resting time**: this is a minimum holding or resting period for contracts;
- **larger tick sizes**: to create a floor for the bid-ask spread; and
- **order-to-execution ratio**: defining a trading ‘ratio’ to reduce excessive message traffic and cancellation rates. Already the SEC is considering bringing market abuse charges against HFT companies that place spurious orders.

The overall conclusion is that regulator policy needs to catch up with realities without undoing the advantages brought about by CBT/HFT. We believe that the lack of appropriate national data facility and infrastructure with reasonably policed availability restricts academic research
into CBT/HFT market behaviour and phenomenon such as flash crashes, and consequently slows development of appropriate legislation.

The implementation of the ATRADE is a team effort involving over 40 UCL and King’s College London (KCL) students and a number of companies. In particular we would like to thank Microsoft for sponsoring the AT competition in 2010, Trading Technologies and Knight (HotSpotFX) for granting us access to their data-feed streams; CitiGroup, BNP Paribas Fortis and BAML for their advice and support; and the Investment Banks for disrupting the work by hiring away our best students at incredible salaries. Amongst the supporters we would particularly like to thank Nick Edwards (Trading Technologies), Martin Frewer (BNP Paribas Fortis), Kevin Gillespie (HotSpotFX) and Dr Piotr Karasinski (EBRD).

6 Narang, RK, Inside the Black Box: the simple truth about quantitative trading, 2009, Wiley Finance.
8 Quantitative analyst trader or programmer trader.
9 Markets in Financial Instruments Directive II, MiFID II (Level 2).
19 Diebold, FX, Elements of Forecasting, 2006, South-Western College Publishing.
26 Markets in Financial Instruments Directive II, MiFID II (Level 2).
29 Narang, RK, Inside the Black Box: the simple truth about quantitative trading, 2009, Wiley Finance.
Algorithmic trading, flash crashes and IT risk

37 Markets in Financial Instruments Directive II, MiFID II (Level 2).
Chapter 6

Cultures of risk?

Paul Willman  
*London School of Economics and Political Science*

Mark Fenton-O’Creevy  
*Open University Business School*

Overview

This chapter discusses the concept of risk culture. It surveys the literature on corporate culture which sees cultures as complex and all-embracing, thus difficult to change. Further complexity is added by surveying the literature on risk perceptions and propensity, which stresses both inter and intra-personal differences in risk behaviour. Taken together, these literatures indicate that the management of risk taking behaviour in financial organisations is a complex managerial task, not something subject to a quick regulatory fix. The chapter looks at five empirical examples of rogue trading to illustrate the dynamics of risk taking in trading situations. Although individual acts were the subject of legal action, the examples are all of a failure of management process. The chapter closes with reflections on management process and practice in financial institutions.

Introduction

It has become fashionable recently to talk about cultures of risk within large organisations. Particularly since the financial crash, it has become a term commonly applied to large financial institutions, and often in a negative way; specifically, following some adverse event or disclosure, media organisations remark on a ‘culture of risk’ as a causal factor. In popular usage, the term is often used as a shorthand description for a set of management practices and/or individual behaviours that are seen retrospectively to have contributed to disaster.

This usage influences views on regulation of financial institutions and their markets. The recent Kay Report on UK equity markets notes: ‘A lack of trust and poorly aligned incentives, have helped create a culture of short-termism in our financial markets…. This is undermining their role of supporting innovative, sustainable long-term business performance. We must create cultures where businesses and finance can work together to create high performing companies and earn returns for investors on a sustainable basis. This means moving away from a focus on short-term transactions and trading to an environment based on long-term trust relationships.’

www.rasabourse.com
Here, culture is seen as a consequence of practices such as reward structures, the implication being that if we alter the practices, cultural change will follow.

The usage has not been confined to financial institutions. It has been applied also to safety risk. In recent years, there has been substantial public reference to the culture of risk in BP following the Deepwater Horizon accident that led to substantial leakage in the Gulf of Mexico; again, the notion of ‘culture’ aggregates a set of practices and a general attitude to risks which are retrospectively seen as inappropriate. Risk and blame are closely associated. In more academic literature, Vaughan has analysed two of the most documented disasters of modern times – the two space shuttle crashes – trying to show how assumptions about and perceptions of risks can be pervasive and enduring in organisations, even where the individuals involved change over time.

However, financial institutions differ from these examples in at least one major respect. Whereas the analysis of safety risk concerns itself with minimising the down side as much as possible, financial institutions are often more concerned with pricing risk correctly rather than eradicating it. Put differently, a culture of risk in a financial institution may lead to adverse consequences not only if it encourages too much risk seeking behaviour but also if it encourages pervasive risk aversion. The experience of failure may differ, the first leading to crisis and the second to loss of competitiveness, but understanding a risk culture in an organisation remains important in both cases.

This chapter seeks to examine the notion of a risk culture in a financial institution. To do this we need some conceptual clarity. First, what is a ‘culture’ within an organisation and how might it affect how people act? Second, what affects individual behaviours in the face of risk? Third, how do these two levels of analysis interact? Fourth, we introduce some empirical material on instances of ‘rogue trading’ over recent years to illustrate the argument.

**Culture in organisations**

The concept of culture is central to anthropology and, therefore, almost by definition, contested, so in the interest of economy let us begin with the definitions of culture offered by those organisational theorists who have tried to import the concept into the analysis of modern business. Two definitions stress what culture does.

...the importance for people of symbolism – of rituals, myths, stories and legends – and about the interpretation of events, ideas and experiences that are influenced and shaped by the groups within which they live.

...the shared rules governing cognitive and affective aspects of membership in an organisation and the means by which they are shaped and expressed.

It is cognitive (inside peoples’ heads), social (experienced as an external constraint) and historical (tradition is often a source of legitimacy). In this view, culture is not a variable or subsystem but within an organisation it is a ‘root metaphor’ or a manifestation of human consciousness. This is a reflection of the anthropological roots; in the study of Pacific Islanders,
for example, culture is more or less everything about the society, not something that can be consciously tinkered with in order to effect a desired outcome.

Unfortunately, those who seek to manage culture in large modern companies often do so with desired outcomes in mind. Normally these outcomes have to do with some measure of performance. In this sense, economic or structural variables can be seen as manifestations of culture, not dependent variables, for example, accounting variables such as cost, revenue and margin indicate cultural choices about what to measure and how. Organisational size, which is a potent explanatory variable in almost all quantitative work on companies, may be seen as cultural preference. Strategic decision-making may similarly be seen as a cultural manifestation. Both Schein and Pettigrew have examined the assumptions and processes of strategic decision-making as cultural phenomena; both how decisions are made (for example, collectively or individually) and what the decisions are about (for example, the identification of competitors) may be culturally influenced.

This tension, between culture as everything and culture as a performance lever, is significant. Managers wishing to use culture to change an organisation or regulators seeking to blame a transgression on cultural factors need more specifics on how organisations, as cultures, work.

Alvesson has rather heroically attempted a summary, using the notion of eight ‘metaphors’, as follows.

- **Culture as exchange regulator** – culture is seen as a set of rules or conventions for handling complex intra-organisational transactions. Typically this embeds notions of fairness, integrity and community to spell out what should and should not happen. This is entirely consistent with a transaction costs perspective on organisations. Culture replaces monitoring by socialising employees into the avoidance of opportunism and helps reduce transaction costs. This notion is most similar to the limited treatment of the idea of culture in organisational economics.

- **Culture as the organisation’s compass** – this view emphasises the role of culture as a set of values that guide the organisation’s ultimate goals and the strategies and processes designed to attain them. Empirically, mission statements and ethical policies provide guidance to this aspect of culture. It is close to the marketing conception of ‘organisational brand’.

- **Culture as glue** – this view emphasises the role of culture as a promoter of social cohesion, helping avoid fragmentation and conflict. Culture is seen primarily as a set of social norms and values which may emerge spontaneously through interaction or, more sinister, the conscious articulation of such norms by senior managers intent on control.

- **Culture as ‘sacred cow’** – this conception emphasises value commitments that control strategies. Organisations are seen as having an identity which dictates the adoption or rejection of specific strategies and the subordination of rationality to emotional attachment. Strategies then become subordinate to the maintenance of identity, even where other options may appear economically more beneficial. Here are two examples. When Apple was a computer company it refused to licence its operating system (the ‘crown jewels’) to other manufacturers. Microsoft gave Windows, its operating system away, and thus became the industry standard. Oxford University maintains an expensive tutorial system,
in which very small class sizes (two to four) are maintained, at vast cost, in deference to its commitment to a particular style of education.

- **Culture as affect regulator** – organisations control and discipline the display of emotions in order to sustain some valued organisational practice or achieve a business objective. A good example is Van Maanen’s work on Disney as a ‘smile factory’; employee emotions are actually the product that is consumed by paying customers. As Hochschild notes, many job roles come with implicit scripts which define display rules for appropriate emotion; we have more to say about this below.

- **Culture as disorder** – given the evidence of cultural differences within organisations, and the ensuing conflict, one needs a metaphor which is in some ways the opposite of culture as exchange regulation to characterise such circumstances. These circumstances include (but are not restricted to) companies which are the result of merger or acquisition, and research exists to show this conflict and tension is a feature both of cross-border merger activity and intra-country mergers; that is, it is relevant to studies of both corporate and national culture.

- **Culture as blinders and culture as world closure** – we have merged Alvesson’s last two metaphors, since they seem close to the same idea that culture acts as a lens making current practice seem like the natural order of things, rather than the outcome of contingencies, and thus blinding organisational members to the possibility of alternate states of things. It is central to the anthropological venture that culture is a concept that enables the reader to understand what might seem bizarre tribal practices as part of a consistent world view. Conformism and the inability to question may be features of modern organisations also. To take two examples, in both oil and tobacco companies, the notion that the core product and central purpose of both industries generated harmful outcomes in the form of climate change and cancer respectively was harder for members to address than outsiders. An interesting corollary of this aspect of culture is that companies affected often have to be pressured into changes by external bodies, and debate on the organisation’s core identity is conducted by non-members. This is largely true of financial institutions after the crash.

Where does this leave us? Let us summarise as follows.

- An organisation is likely to develop a culture or cultures whether its leadership consciously seeks to or not. These cultures may have positive, negative, or neutral effects on any performance measure. Empirically, there is likely to be a mix of such effects where culture is not explicitly managed.

- Where it seeks to manage its culture to improve performance it must be able to articulate a likely relationship between some attitudes and behaviours and one or more performance outcomes. The conventional way to do this is by developing a ‘business model’ in which certain non-financial indicators are inferred to predict future cash flows. For example in a retail bank, customer service measures may be seen to predict customer retention or profitability through increasing customer satisfaction, or in an automotive business quality measures may predict profit per vehicle through increasing productive efficiency.

- Managing culture is demanding. In order to effectively manage culture one must pay attention to control of inputs (for example, selection of employees on cultural values), processes
(for example, evaluating employees on ways of doing work) and outputs (the desired performance variable). This has substantial set up and maintenance costs. Organisations that manage through performance outcomes alone find it extremely difficult to manage culture. For example, Roberts describes how in the years before the Deepwater Horizon disaster, BP successfully extended the lifespan and reduced the amount expended on fixed assets through rigorously monitoring and rewarding on the basis of return on capital employed. This was a financial success but, ultimately, a safety disaster. You get what you pay for.

- Organisations that do not manage inputs and processes (that is, employee selection and employee behaviours) will not only fail to manage culture, they will expose themselves to unmeasured operational risk. In a study of trading floors in investment banks in London prior to the crash, we found little evidence of effective management of inputs or processes, but rather a determined focus on outputs (trader or desk profits). In fact there are no studies of trading in financial markets that show a focus on inputs or process. This may be particularly problematic given what we know about learning from feedback. Feedback is an important component of any learning process. Yet, there is considerable evidence that many feedback interventions are ineffective. For example, a major review and meta-analysis of over 3,000 studies found that while, on average, feedback interventions improved performance, up to a third of feedback interventions actually reduced performance. Simple feedback on outcomes is particularly problematic when tasks are complex or important antecedents of outcomes are beyond the control of the feedback recipient. This is very much the case for trading and investment decisions. Poor strategies can produce good outcomes due to unpredictable changes in the market and vice-versa. In these circumstances, negative outcome feedback may tend to trigger defensive processes (in order to avoid negative emotions) and positive outcome feedback may generate false beliefs about the value of particular strategies or the nature of market processes. In contrast to outcome feedback, which concerns the accuracy or correctness of response, process feedback concerns the how and why that underpins the outcomes. There is evidence that (properly designed) process feedback has significantly greater utility for learning than outcome feedback in complex real world domains. Some of this evidence is specific to trading tasks. For example a study of the decision-making of professional security analysts found better performing security analysts to be more likely to ignore outcome feedback and more likely to focus on information about underlying issues such as fundamental market factors.

So, a culture of risk (or risk management) would focus on who is taking risks and how they do it. Analysis of the first requires that we know something about individual differences in approaches to risk.

**Risk takers**

There is a substantial literature in applied psychology on variance of perception and behaviour in the face of risk. Individuals may vary in their perception of the degree of risk they face, and in their appetite for risk (risk propensity). In addition to these interpersonal differences, there are also intrapersonal differences; specifically, individuals may vary their behaviour...
Cultures of risk?

in the face of risk over time (for example, depending on their recent history of gains and losses), or across domains (for example, we might have different appetites for health risks versus financial risks).

The literature relating to risk propensity has three main themes. The first theme relates to expected utility theories, of which prospect theory is a much-cited example. Prospect theory proposes that risk taking is asymmetric about a reference point: people will be risk averse when they perceive themselves to be in the domain of gain, and risk seeking in the domain of loss. Prospect theory has stimulated numerous research studies into risk preferences and risk taking. One premise of the theory is that individual level risk taking is relatively inconsistent across situations – a person will take risk in some circumstances, and avoid risk in other circumstances. The prompt for behavioural change could be as simple as the semantic presentation of data, for example whether a choice outcome is framed as a loss or a gain. A meta-analysis of studies of such ‘framing effects’ found that framing is a robust phenomenon, particularly when reference points and outcomes are manipulated.

Factors that might influence risk behaviour and framing effects are considered in more detail in a stream of research that considers the individual difference factors that could influence risk taking. Among psychological factors, personality seems to be the strongest contender for major effects on risk behaviour. For example, Highhouse and Yuce suggested that findings which apparently contradict prospect theory, that is, risk taking in situations of gain and risk aversion in the domain of loss, are due to individual differences in perceptions of what constitutes threat and opportunity. Risk propensity could be more a characteristic of an individual than their situation. In this area, sensation-seeking has been found to be particularly important. Since Zuckerman pioneered the study of this concept a stream of research has confirmed its importance as a highly consistent predictor of various kinds of risk taking, including compulsive gambling and participation in high risk activities.

A related individual approach has been to consider risk propensity in terms of the variance in within-individual measures of risk. An early example of this work is Weinstein and Martin, and other studies have adopted the same approach in more recent research. These empirical works focus attention on the inter-correlation of scores across a range of measures of risk taking in different decision areas. Findings have typically shown correlations between different measures of risk to be weak, suggesting that people do not have generalised tendencies to take or avoid risk. However, research on managerial decision-making by MacCrimmon and Wehrung showed that this pattern of results does not preclude the possibility of strong intra-individual convergence of different measures of risk taking. They found that a small number of people showed consistent responses on different measures of risk taking, and could be categorised as consistent risk seekers, or consistent risk averters.

What data we have on trader personality seems to suggest that there is no especial trader personality type. In these terms, traders are ordinary rather than exceptional. Trader personality types span pretty much the full range of types to be found in the general population, with some tendency, on average, to be somewhat more introverted and conservative and more risk averse than the general population. However this ‘ordinariness’ of personality types also implies that extreme types are as likely to crop up in trading as elsewhere and there is little in trader selection processes which is screening out these extremes.
The work of Weber and Milliman, and subsequent work by Weber et al., represents an important development in this field. These authors found that while the degree of risk perceived in a situation can vary according to the characteristics of the situation, attitude to perceived risk (the degree to which people find perceived risk attractive) remained stable across situations for a significant portion of their sample. This work is part of the third stream of literature – the combination of situational and individual approaches to risk propensity through consideration of individual responses to different risk domains. Work in this area shows that it is possible to be risk seeking in some areas of one’s life and risk averse in others while having a relatively consistent view of risk. Data from perceived risk attitude studies, such as Weber et al., suggests that both general (for example, sensation-seeking) and domain-specific (for example, perceived risk) risk propensities are possible.

Let us return to the managerial implications of this, again using our earlier example of trading. In order to effectively manage the operational risk emerging from psychological differences in risk perception and preference among traders on the desk, the trader manager should:

- **select** traders based on the (stable elements of) their approaches to risk. This includes their risk preferences and propensities, and the extent of sensation seeking; as Fischer Black once noted, if risk taking enters the utility function, it becomes difficult to predict anything. This selection probably should be on a portfolio basis; a desk of variable risk preferences enables hedging of behaviours in a manner parallel to that of an asset portfolio;
- **monitor** trader decision-making with an eye to the avoidance of decision biases and framing effects. This might be facilitated if traders themselves self-monitored for decision bias; and
- **be aware** of the trading history of each trader, specifically with respect to the recent history of loss and gain (as perceived by the trader).

We could list more, but this is enough to illustrate the difficulty of the task.

**Emotions and risk**

We have characterised cultures as mechanisms of emotional control. In this section, we relate emotional control to risk behaviour. Money fractures marriages, drives wars, inspires art, motivates some to great achievements, leads others to despair. Fear, desire, love, hate, jealousy, anger, anxiety, relief, shame and many more shades of emotion may attach to money in the course of an ordinary day. Yet, economic accounts of human financial behaviour focus on reasoned weighting of information and the utility of different courses of action. Emotions are typically absent from economic accounts of markets and financial behaviour except as an element of future utility. However, emotions are a pervasive and necessary element of human decision-making, including trading, providing an important aid to navigating complex risk environments and often providing early information about risks based on pattern recognition before the slower conscious cognitive systems catch up.

Consider the example of driving to work. You are driving automatically on a familiar route, chatting to a passenger and listening to a news report on a radio. Suddenly a small child runs into the road. It is the instant emotional reaction which directs all of your
cognitive and physical resources to avoiding hitting the child whilst avoiding collisions with other vehicles. In fast-paced decision environments like trading it is also emotions which help direct attention to the most salient aspects of complex information flows. However, just like other elements of our thinking processes, emotions can mislead. Many of the traders we have interviewed described emotion driven errors and described the disabling impact of strong emotions, especially in their early career, to the extent sometimes of terror driven vomiting in the bathroom.

Thus the self-regulation of emotion is an important challenge for any trader; striking an appropriate balance between the necessary use of emotion in directing attention in highly dynamic and complex environments, whilst avoiding the decision traps which can arise from the carry-over of emotion from one set of trading outcomes to the next trade.

There is growing research evidence that effective regulation of emotions is an important element of trader expertise which links both to financial performance and risk management; and that there is considerable variation among traders in their emotion regulation capabilities. In one study, a physiological measure was used on the trading floor to measure the moment by moment emotion regulation of traders in two investment banks, throughout the trading day.

A particular source of emotional pressure for traders is associated with market volatility. In volatile markets, outcomes become more uncertain and risks greater. A commonly used measure of market volatility (the VIX) is colloquially known as the ‘fear index’. High levels of the VIX are coincident with high degrees of market turmoil and spikes in this measure can be seen at times of stock market decline, the threat of war and other major events which create great economic uncertainty. The period of data capture in the study was particularly interesting in terms of market volatility, following shortly after the 2008 financial crisis and during a period of great uncertainty about sovereign debt in Europe. VIX values ranged from 17.42 to 42.15, representing large variation in expectations about market volatility relative to historic levels. The findings not only confirmed prior research on the link between effective emotion regulation and trader expertise, but also showed traders to have greater difficulty in regulating their emotions at times of greater market volatility, regardless of experience.

Of course, in all studies one finds exceptions to general trends; and one such exception was particularly revealing. When we looked at the physiological data from one highly experienced trader it was clear that he was showing signs of intense stress in periods when the market was quietest. This was so contrary to our expectations that we discussed the results with him. He was entirely unsurprised and explained that he had experienced an unusual run of poor performance and was in real danger of failing to make his targets; thus likely to receive a considerably reduced bonus. This, he told us, meant he felt under intense pressure to recover lost ground. Periods of low market activity meant fewer opportunities to trade at a profit but he felt under great pressure to engage in ‘forced-trading’ that is ‘...to trade anyway or take stupid risks’. He ascribed the stress to resisting this pressure.

This illustrates an important intersection between incentive structures and traders’ emotional experience. Trading is skilled work, but good traders often lose money and poor traders can get lucky. There is a lot of noise in the relationship between skill and performance. This can be a source of enormous performance anxiety for traders. The trader who
gets lucky early on and develops an undeserved reputation (and bonus) for good judgment will feel under great pressure to maintain a level of performance that is beyond them. The good trader who is having a period of bad luck and diminished confidence can see earnings and reputation suffer. But traders build up financial commitments that reflect their earnings and faced with an inability to sustain performance (and status) can feel under tremendous pressure to take unreasonable risks or bend the rules.

This is exacerbated by the human capacity to fool ourselves. Not only are we loss averse, taking risks to avoid loss, but commonly we seek to avoid negative emotion. This can be useful in focusing us on behaviours likely to resolve the situation causing distress. However, we sometimes engage in less productive strategies to resolve the pain; hiding behind self-protective illusions. Individuals deprived of a sense of control make active efforts to restore it cognitively. Given internal goal states embodying desires for control and faced with evidence of a lack of control some may cope with the discrepancy by denial and retreat into self-protective illusion. One such common self-protective illusion which has been studied in investment banks is the illusion of control. This is the tendency to assume we have greater control over a situation than is reasonably possible. Early studies on illusions of control by Langer showed that people often behave as if chance events are accessible to personal control. Key elements of the trading environment are particularly conducive to the development of illusions of control as we describe below.

**Noisy feedback:** markets are in practice very ‘noisy’: there is a lot of trading going on that is not based on information genuinely relevant to the underlying value of an asset. On any individual trade it will be difficult to tell whether an outcome (positive or negative) is the result of trading on information or of essentially unpredictable market movements. Hence, it will often be difficult to determine whether an outcome was contingent on a trader’s information and skill. At the same time traders are highly motivated to establish causal relationships between information they hold and market movements.

**Stress:** illusions of control are more common in circumstances of stress. Investment banking is a highly stressful occupation in terms of workload, time pressure, visibility and uncertainty coupled with limited control opportunities, and this occupational group suffer significantly higher levels of free-floating anxiety than the general population.

**Competition:** illusions of control are more common in competitive environments. The process of trading is innately competitive. The markets in which traders deal are founded on competition between market actors. Furthermore, dealing rooms are often highly competitive environments.

**Implemental mindset:** illusions of control are both more common and more severe when conditions induce an implemental (focus on goals) mindset than when they induce to adopt a deliberative (reflection on action-outcome contingencies) mindset. The bonus system and associated targets are designed to keep traders goal-focused. The short-term nature of information advantages also means that traders are unlikely to forgo the opportunity to trade in order to learn more about the value of information or a strategy, by observing the market.

**Choice, involvement and familiarity:** choice, involvement and familiarity can act as skill cues, leading to an illusion of control. Trading involves continually making choices but more importantly requires close focus on a particular type of instrument or market. Traders are often highly identified with the instruments or markets in their area of expertise.
In a study of 107 traders in four City of London investment banks, we used a computer based task to measure propensity to illusions of control. The findings were striking. Trader performance was inversely related to susceptibility to illusions of control as measured by managerial ratings of risk management performance. Traders with high illusions of control were paid significantly less, contributed less on average to desk profits, and were rated by managers as poorer at analysis and risk management.

Illusions of control are just one example of the human capacity for self-deception. As any conman knows, properly motivated by strong emotions such as fear, status anxiety, or greed, people are endlessly creative in their capacity for self-deception. Intelligence is no defence since it simply implies greater cognitive resources to be marshalled in the cause of self-deception. As we noted earlier, for good or for ill, emotions guide what we pay attention to; and these effects do not just apply to middle ranking traders. It seems at least plausible that senior managers in large financial institutions who failed to ask what extraordinary risks were attached to extraordinary profits being made in sub-prime mortgages or who failed to notice the systematic rigging of the Libor rates were in the grip of self-deceiving illusions fostered by intense pressures for success and fear of failure.

Paradoxically, those who show the least emotion may often be at greatest risk. There is evidence that people who manage their emotions by suppressing them may, in the end show stronger effects of those emotions than those who consciously attend to them and use strategies such as ‘reframing’ to modify their emotions.

To return to our discussion on culture at the start of this chapter, organisations are not simply arenas for the production and display of emotions they also produce particular kinds of expectations about what are legitimate forms of emotion. In most organisations there are strong cultural expectations about appropriate emotion display and banks are no exception. In our studies of professional traders this has manifested itself most strongly in interviews with junior traders who were typically very concerned to present a particular kind of script in relation to emotion display. For example: ‘I’m a bit of a cold fish; I don’t think emotions greatly affect my decision making. If you are making money, you are achieving your objective.’

However, typically much later in the same interview, with a little more trust established, a more complete picture would emerge. The same trader, 30 minutes later, told us: ‘When you lose money, it can be horrendous, violent mood swings. You do not know what to do when you lose money... Managers on the desk are an added pressure, and my colleagues. So this emotion can encourage you to make decisions more quickly... It is important to have a recovery period from losses to be more emotionally stable before you go back in [to trading].’

What then does this imply for management?

- Emotions are a necessary part of human decision-making; attempts to eliminate them are at best doomed to failure and typically counter-productive.
- However, there is building evidence that effective emotion-regulation skills are an important component of expertise on the trading floor underpinning effective performance and inoculating against a tendency to succumb to self-protective illusions.
- Managing the emotional climate and understanding the nature and consequences of traders’ emotions is a key management task.
In our work, we did come across some notable examples of trader managers who saw regulating traders’ emotions as a core part of their role. Take, for example, this highly experienced trader manager: ‘I care deeply because I have tremendous pride in the people here... The stress is generated by fear entirely and it is fear of what? I guess everyone has their insecurities whether it is being fired, losing lots of money and appearing stupid in front of their peer group. Whatever it is, it is definitely fear and if you can take the fear out of the situation they perform better.’

However, in our experience, such attention to the emotional agenda is uncommon and certainly highly inconsistent across managers. Further, cultural assumptions which associate overt emotion with weakness or irrationality can act to make much emotion undiscussable. We would argue that given the strong contribution that poor emotion-regulation can make to operational risk, that emotions and emotion-regulation deserve greater managerial attention.

People, risk and environments

We now turn to consider an approach that models the interaction between culture and environment on the one hand, and risk preferences on the other.\(^{54}\)

In general people are not risk seeking in the true sense of the word; they are willing risk-bearers or stimulation seekers. Both groups often only have a hazy sense of what real risks they are bearing. Risk cannot be ‘sought’ because it is not a meaningful goal for a person to seek. Let us use the model in Exhibit 6.1 to explain.

At the top of Exhibit 6.1 are two key sets of factors whose nexus is the substance of risk behaviour. On the one side are individual differences, that is, the repository of risk relevant person-specific drives, dispositions, tastes, abilities, styles and so on. On the other side is the risk environment; the array of risks that exist at any point in time for an individual or group, which can be analysed as a set of opportunities and threats, with a profile of stability or change.

Self-selection (arrow 1), represents the tendency to orient towards certain environments and away from others. People enter risk domains because of other features that attract them, or sometimes because they have no choice. They may take jobs and join organisations that bring with them a degree of unavoidable risk. Conversely, environments select people (arrow 2) because they want to make use of their skills and character. In some high-risk occupations – such as firefighting – there is a conscious exchange at the point of engagement about the person’s interest and willingness to bear the risks of the job. In other occupations this may not occur, such as the job of publican (bar owner), statistically one of the most unhealthy jobs in our society, but one about which one can infer that many incumbents did not take on the position with much thought about the risks they would be bearing.

Agency (arrow 3) represents how individual purposes, goals and intentions arise from unique identity elements, for example, emotional sensitivities and personal capabilities. Arrow 4 represents the flow of information, imagery and impressions about domains that capture attention. Our orientation to risk is entirely dependent upon this representation, which is filtered or amplified by media and other agencies. Constraint and opportunity (arrow 5), represents those features of the environment that, often with quite inadequate information,
Cultures of risk?

Exhibit 6.1

**A model of individual risk behaviour**


expose people to risk, as in our example of the publican who remains quite unaware of the health risks of his occupation throughout his tenure.

*Goal-seeking behaviour* (arrow 8), represents the heart of the argument: the idea that most people are exposed to risks not because of any liking they have for risk but because of the nature of their wants. Loss avoidance is the greatest inducement to risky behaviour; as is known and feared in trading environments. The capacity for self-consciousness affects this risk calculus, as shown by *bias* and *accommodation* (arrows 6 and 7), respectively. *Bias* (arrow 6), is the process identified within self-regulation theory through which people misperceive risks in order to remain aligned with their goals as well as how they are framed. Wanting something very much focuses attention on the reward and blinds one to degree of risk exposure. Interestingly, in the most extreme cases of want the connection is short-circuited, without even the need for judgment. Human behaviour has a non-calculative automaticity in some circumstances: for example, humans, like other animals, will instinctively put their lives at risk to save offspring from danger. *Accommodation* (arrow 7) is the moderation of goal

www.rasabourse.com
orientation in the light of risk information (that is, from whatever processes are occurring under arrow 4). The implication of arrows 6 and 7 is that ultimately it is in the control of the individual how much they choose to be aware of and attend to risk information, except perhaps in life and death crises.

Learning/adaptation (the last arrow 9), represents learning from experience. This plus other feedback processes represents the engine that drives a lot of risk behaviour, by reducing risk perception. It explains why people will run great risks on streets and motorways, as drivers, passengers and pedestrians, without even thinking about them. As has been pointed out, there is a kind of homeostasis in our everyday risk taking: our quiet and comfortable cars loaded with safety features can lull us into a dangerously passive stance in relation to risks.

Finally, it is important to note that some of the most important risk-taking determinants come from biases that infect the aftermath of a risk episode. If people bear risks and suffer no harm then this may reduce attention to risk information and bias perceptual awareness. One major input to this danger is people’s unfounded beliefs that they are able to exercise personal powers to control or recover the consequences of uncalculated risks. This may come as much from dispositions, such as high self-esteem, as from faulty learning.

Let us summarise the model and its implications for risk management.

It indicates that risk behaviours are induced principally by four causes: (i) location – such as an imposed risk environment; (ii) loss avoidance – the drive to avoid some known threat; (iii) goal-seeking – powerful orientations towards known gains; and (iv) sensation-seeking – the desire for stimulation through danger. The first of these, location, points directly to the processes of selection and self-selection into the trading context.

As we have discussed in our trading examples, management systems try to desensitise traders to excessive fear of loss, whilst establishing norms of vigilance and control over acceptable risk. Trading environments are as information-rich as any in the business world. Traders do not require an appetite for risk; they are constrained (and incentivised) to take risks (arrow 5) by virtue of their roles and positions, albeit in many different forms according to the nature of markets and trading roles.

The challenge for management in this analysis is how to counter the three hazards of too much risk bearing, avoidance of risk and errors of judgment about risk. The model suggests several paths to doing so, only some of which are currently practiced in financial institutions.

• First, it suggests more attention could be given to selection, with a more intelligent modelling and matching of individuals to roles.
• Second, the information environment needs to be actively managed to ensure that traders are accurately apprised of actual risks (and avoid the clutches of the actuarial fallacy).
• Third, their goal orientation needs to be monitored and managed. An important issue here is to avoid too close a coupling of risk with individual reward, such that risk bearing becomes unduly amplified.
• Fourth, and this is probably the biggest gap, managers need to understand the psychology of decision-making, bias and risk, to the degree that they can help train traders to avoid major pitfalls.
• Fifth, the model stresses the importance of monitoring what traders are learning. Inaccurate inferences can easily be drawn from apparent consequences of actions.
Case study: rogue trading and cultures of risk

In this section, we use the example of rogue trading to illustrate the dynamics of culture and risk management within financial organisations. It may be an extreme example, but such examples recur at intervals and across companies and markets. What we specifically mean here are examples where an individual appears to have taken on significant and inappropriate risks leading to losses and has then succeeded in covering up losses to the point where, on discovery, the employing organisation has incurred significant financial and/or reputational damage. The examples we cite here are the result of casual empiricism; that is, this is a biased selection and includes only those cases which have gone public, and about which sufficient data have subsequently been disclosed by regulators or legal process. It is highly likely that the underlying dynamics of risk, loss and cover up are more common in trading activities.

The instances are as follows.

1. The trading activities of Nick Leeson in Singapore futures markets that led to the collapse of Barings Bank in 1995 (over US$600 million losses).
2. The losses incurred at Daiwa Bank over an extended period from 1984 to 1995 in bond trading by Toshihide Iguchi (over US$1.1 billion).
3. The commodity trading (copper) losses of Yasuo Hamanaka employed by Sumitomo Corporation at the London Metals Exchange (eventually over US$2.6 billion).
5. The trading in equity derivatives by Jerome Kerviel at Société Générale between 2006 and 2008 (over €4.5 billion).

The instances have been chosen because they are in different markets and companies and because they have been relatively well documented, primarily as a result of retrospective analysis by courts or regulators; we are not concerned here with documenting the trades themselves. It is significant that, despite numerous advances in technological capabilities for monitoring trades, and learning by both companies and regulators, such examples repeat.

The cases we cite have the following characteristics.

• The trading activities went on for a considerable time. For example in the Iguchi and Kerviel cases, there were thousands of trades.
• Losses continued for a considerable time, and losses escalated as the traders sought to rectify their positions.
• Losses were apparently undetected by settlements and compliance in the organisations concerned, again for a considerable time.
• Except in the case of Rusnak, there was no evidence of direct personal gain other than the rewards for apparently successful trading issued by the corporations concerned.
• However, the ‘verdict’ on all of these events was that the individual acted alone. This was reflected in the punishments delivered on individuals subsequently.
We sought to answer the following questions about this set of cases.

1. How was it possible to evade detection by compliance or settlements departments within the organisation?
2. Why did the individuals concerned not voluntarily cut their losses when trades went sour?
3. Did the individuals act for personal gain?
4. Why did trader managers not spot the problems?

Our main findings are summarised in Exhibit 6.2. In four out of five cases the traders concerned had either experience of internal regulatory procedures or some measure of direct control over back office processes. In all five cases, there is evidence from the trading histories and trader revelations of loss aversion – increasing risks to eradicate losses. However, in only one case is there evidence of direct personal gain outside the employment package. In four out of five cases the traders had not taken holidays over an extended period, and in all cases the corporation concerned had examined the possibility of fraudulent trading but, finding no evidence, had dropped the investigation.

---

**Exhibit 6.2**

**Rogue trader data**

<table>
<thead>
<tr>
<th></th>
<th>Leeson</th>
<th>Iguchi</th>
<th>Hamanaka</th>
<th>Rusnak</th>
<th>Kerviel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Settlements/compliance experience</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Control of back office</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Loss aversion</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Personal gains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Holidays</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Early detection ignored</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Source:** Authors' own

What does one make of these findings? First, experience of settlements and compliance does not only provide expertise relevant to the detection of trades but also a network of contacts and, perhaps, influence on an ongoing basis. Someone with such experience operating in a trading role arguably embeds greater operational risk than someone without such experience and contacts. Second, loss aversion is, presumably, a decision bias about which any trader should be aware. Either this awareness was ignored, or perhaps the payoff matrix is biased in favour of loss aversion. Once losses reach a certain level, voluntary or involuntary disclosure has the same payoff: jail. Non-disclosure and increasing operational risk emerge as the rational strategy for the individual but, of course, not for the corporation.
This may be one type of agency problem but, in other respects, agency theory does not work simply because the individuals concerned were seeking to make substantial profits for the corporation, rather than themselves. However, both reward and arguably status issues align the individuals’ interests with corporate profit and, to the extent that alignment exposes corporations to unobserved operational risk, agency theoretic approaches to management are dysfunctional in that they create precisely the problems they are designed to avoid.\textsuperscript{56} In all these cases, trader rewards were output (profit and loss (P&L)) based, rather than focusing on (and thus monitoring) decision making.

The fourth and fifth items seem to point towards simple risk management failures, again based on process. Taking regular holidays is a regulatory requirement so that someone else might examine traders’ positions. If it does not happen it is a simple compliance failure. Early detection then ignoring subsequent monitoring has two aspects. First, there is clear indication in this set of cases that management oversight should have been more thorough. Second, there is evidence, not least from Vaughan’s earlier cited work,\textsuperscript{57} that a problem, once identified, is deemed ‘solved’ once investigated; malfeasance is protected by the desire to protect the initial investigator.

Overall, these findings point away from the idea of a lone gunman firing off trades to the detriment of the organisation. However, they do not point directly at any other complicit agent either. Rather they point to a culture in which the valuation weighting on apparent trading success is much higher than that on effective risk management. As Kahneman has noted in several publications, optimism and a focus on the upside is generally a much more successful personal career strategy than a Cassandra-like focus on possible future disasters. Particularly in cultures where individual success is highly valued as well as rewarded, ‘yes, but...’ is not a career enhancing opening to discourse.

A second issue concerns the nature of management and regulation; specifically, it all sits within the same corporate culture. Our earlier study found that trader managers were overwhelmingly former traders, and that they valued the autonomy of trading activity very highly. A ‘light’ managerial touch in the absence of overt losses was favoured.\textsuperscript{58} In addition, corporations that employ traders also employ trader regulators. It is an industry that overwhelmingly has called its own fouls. Ultimately, those who might feel the need to stop trades report into managers whose bonuses depend on successful trades. There is enough moral hazard here to go around.

Conclusion: ‘fixing’ risk cultures

We opened this chapter by saying that many observers feel that the cultures of some financial markets and institutions needed ‘fixing’ to generate better performance outcomes for the sector and the economy as a whole. Cultures were seen to be sources of operational risk, and dealing with this risk was seen as a regulatory task. The argument in this chapter agrees with the proposition that such cultures are the source of regulatory risk; moreover, it is not quantified but probably substantial.

However, we have complicated the ‘fixing’ argument by laying out two broad arguments of our own. First, we have introduced a broader and more embracing notion of culture. This is consistent with thinking in the management field, where the relationship between
culture and organisational performance has generated a massive literature in the 30 years. This literature would, broadly, see the issues of trust and incentives mentioned in the Kay Report as outcomes of an organisational culture or cultures and thus see the argument that, for example, changing incentives would change culture as mistaking the cart for the horse.

Second, we have argued that to understand how culture generates operational risk one must understand how people perceive risk and respond to it, and the environmental triggers for these responses. This makes the management task more complicated. Because of inter and intra-personal variations in risk propensity, one must pay attention to the ways in which employees who expose the organisation to such risks are selected, developed, rewarded and punished.

This points attention less towards the regulator and more towards the quality of management and the rigour of management processes within financial institutions. Managing by outcomes – for example, bonuses for traders – is, quite simply, the lazy way to manage operational risk. Our rogue trader cases illustrate the importance of management processes and information, since the outcome data can be regularly manipulated.

In our field study of investment banks in London prior to the crash of 2008, we found that, compared to organisations outside the financial services industry of similar scale, large financial institutions did not put a high priority on management development. A strong emphasis on technical rather than managerial skills, and a preference for putting high performers into leadership roles independent of leadership abilities has characterised the sector for many years. Without addressing fundamental questions about managerial roles and capabilities, the changes to culture which many outsiders advocate will be extremely difficult to achieve.

Bibliography


7 Schein, EH, Organizational Culture and Leadership, 1985, Jossey-Bass.
10 Willman, P, Learning the Business; the social scientific foundations of management, forthcoming, Pearson.


15 Willman, P, Learning the Business; the social scientific foundations of management, forthcoming, Pearson.


52 Subsequent events at UBS are being researched.
Chapter 7

Reputation management after 2007

Roland Schatz
Media Tenor International AG

Overview

This chapter looks at the damage that has been inflicted on the reputations of financial services companies in recent years, some of which predate the current financial crisis, and provides some prescriptive advice on how it can best be repaired. In-depth market information is used to analyse just how the media and citizens of a number of Western countries view financial services institutions and why it is that this industry has managed to reach a point where it is in certain locations viewed as negatively as the tobacco industry.

Introduction

The financial services industry has had to live through a tough few years. They have been pretty much blamed for a large number of the crises that have engulfed the world in recent years. Politicians and the media have, rightly or wrongly, ensured that the blame for the recent financial crisis is placed squarely at the door of financial institutions. One week ahead of the 2010 World Economic Forum in Davos, President Obama publicly threatened Wall Street, saying, ‘If these folks want a fight, it is a fight that I am ready to have’. They have even had to carry some of the blame for the crisis in the Eurozone. Of course, the recent Libor fixing crisis, the class action lawsuits against misinformation and wrongful repossessions have not helped. Neither has the excessive pay for failure that has continued throughout and post crisis.

Together, all these reasons are enough for financial executives to think long and hard about why the reputation of their industry has reached such a low point.

Banking executives and their boards of directors need to ask themselves whether they have managed to provide convincing responses to the public for the following questions.

• What have we put in place to ensure that we communicate our financial positions as accurately and transparently to our clients and stakeholders as possible?
• Have apologies for our negligence been clear enough for customers to both hear and understand them?
• Have new policies been put in place to strengthen risk management policies?

While a number of first movers managed to gain some trust back, most financial institutions instead managed to stay in the headlines by increasing the fees for current and savings account

www.rasabourse.com
transactions. Home owners in the US, who lost the roofs over their heads, increasingly saw statements in the news from their old contractual partners stating that it was their own fault they took out loans they were unable to pay back. From an objective and technical point of view, this was perhaps correct, but in the context of 2007 this was not only cynical, but also exposed the incompetence of bank employees who had checked and countersigned these contracts. Additionally, headlines regarding bankers’ bonuses captured the media’s attention, including UBS’s announcement that it planned to pay out CHF4 billion in bonuses to its managers despite the company being in the red.2

This generated public outrage then and the fact that they continued on this path in 2013 certainly did not make it any better. The reputation ranking of the bankers appearing in the Financial Times as well as global prime time news speaks a clear language (Exhibit 7.1).

### Exhibit 7.1

**Visibility and tone of coverage: banking industry executives (2012)**

<table>
<thead>
<tr>
<th>Name</th>
<th>Financial Times</th>
<th>TV news</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diamond, Bob</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimon, James</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hester, Stephen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agius, Markus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pandit, Vikram</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blankfein, Lloyd C.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jain, Anshu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gulliver, Stuart</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ackermann, Josef</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ermotti, Sergio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fitschen, Jürgen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gorman, James</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jenkins, Anthony</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dougan, Brady</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sands, Peter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Botin, Emilio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goodwin, Fred</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walker, David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diamond, Bob</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimon, James</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fitschen, Jürgen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ackermann, Josef</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jain, Anshu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hester, Stephen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breuer, Rolf-Ernst</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adoboli, Kweku</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agius, Markus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halver, Robert</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ermotti, Sergio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gribkowsky, Gerhard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Härter, Holger</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rapanos, Vassilis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Source:** 657 / 364 reports on the banking industry managers listed above in the Financial Times and on 22 international TV news

At the same time, Goldman Sachs (GS) CEO, Lloyd Blankfein, was quoted as saying that the financial meltdown ‘could be compared to a hurricane’,3 rejecting all responsibility for the bank’s role in what happened. The statement came only a few weeks after Financial Times referred to GS as ‘the bank we love to hate most’. This should have been an important wake up call to not just those at Goldman Sachs’ headquarters. In fact, it took GS until
April 2013 to realise that they have to actively work on their media relations if they want to overcome this perception. This was illustrated years ago in a German conservative business paper. Cartoons are generally a late indicator of reputational damage, and *Frankfurter Allgemeine Zeitung* (FAZ) tried to assist by letting banking executives understand just what their reputation had become (Exhibit 7.2).

Exhibit 7.2

**Mother to her children: ‘I don’t want you playing with the bank directors’ kids’**

![Cartoon](source: Frankfurter Allgemeine Zeitung (FAZ), 19 January 2010: Business section, p. 1, cartoon)

People still remember the banking crises from 20 years ago in which the government in Washington, as part of its rescue action (US$300 billion was needed back then), took responsibility and dismissed top management, without notice or bonuses.

Today when Joe Average reads the headlines that Wall Street bankers are still receiving millions of dollars in bonuses and getting promotions while he sits on the sidewalk with mortgages he cannot pay, a job he has lost and Washington using his tax dollars to bail out the architects of these problems, it is no wonder there is uproar and distrust in the system. Even GS paying their taxpayer-funded loan back at a higher interest rate than required was not enough to re-establish public trust.

Every citizen on every continent whose public swimming pool is closed because of low municipal coffers, whose children can no longer visit local libraries, or whose schools have shut down due to lack of funds will judge their politicians in the coming elections on what actions they took to ensure that financial institutions were held accountable for the damage

---

*www.rasabourse.com*
they caused as well as what policies were set in place to avoid another market crash triggered by greed.

People always look for scapegoats during times of crisis. The most suitable targets have generally been in the minority in whatever community has faced turmoil. And, if this minority has reputational issues, then the finance industry and politicians simply cannot continue with the business-as-usual attitude they have engaged in since 2007.

Exhibit 7.3 presents the ratings of the banking industry across a number of countries. In order to understand the relevance of these results, one needs to know that a company which has filed for bankruptcy is usually rated by the media on average 40 to 50% negative. The way the finance sector was portrayed throughout all of 2012, with ratings close to 60% negative on balance, puts the banking sector at close to broke and suggests its managers as guilty of corruption, for far more than a few short weeks.

---

**Exhibit 7.3**

**Tone of rating of the banking industry by country**

![Diagram showing the tone of rating for different countries across two years: 2010-2011 and 2011-2012.](image)


*Source: 6,329 / 5,864 reports on 21 international TV news from the countries listed in the diagram*
To demonstrate that this is no longer just a reputation issue, we will choose only three concrete actions taken by Main Street which were direct responses to the fact that politicians were now willing to really get tough on Wall Street.

1 Moveyourmoney.org, created by The Huffington Post, is a typical initiative that allows the marketplace not to have to wait for legislative action. The appeal during the 2009 Christmas season to all who continued to allow the large banks to administer their money has been taken up by media worldwide and is easy to implement. Similar to the Shell boycott following the Brent Spar affair, it costs the knowledgeable customer little to withdraw their money from a bank and to transfer it to a bank with local references and control.

2 Customers had already begun turning away from the financial system 10 years ago. Entire territories in Europe abandoned the euro and created their own regional currencies. In the beginning, when the diehards were sneered at across the old continent, ‘global’ concepts were developed that achieved massive acceleration after the financial meltdown. In the beginning these concepts largely followed the anti-globalisation ideology and therefore featured only limited, but considerable, volumes. In the meantime, the regional currencies made the connection to the existing currencies and became a relevant alternative for both the middle-classes and the wealthy.

3 Up until recently, micro-financing was regarded as irrelevant by most major banking institutions. Today, with more than 2,000 different micro finance operations in place, banking executives have come to the realisation that 10 million low risk US$50 loans are, on the whole, more sustainable than a US$500 million loan gone wrong. At the same time, microfinance has historically been considered a financial instrument only for poorest of the poor, but the media is now reporting that the micro-finance concept has found a niche in Spain and several other mature markets.

These early initiatives above are only first examples of money moving away from large global financial players to local, more trusted, transparent and accountable institutions with regional references or personally liable shareholders. This trend has become more intense in 2013 and will continue as long as trust is not restored on Wall Street. The latest letter sent on 14 May 2013 by Senator Elisabeth Warren to the US authorities speaks a clear language.

Meanwhile, executives of the ‘too big to fail’ banking institutions have created the impression that their bonus negotiations are still more important to them than making sure their customers and stakeholders understand how the finance sector must change to prevent another melt down. To add more fuel to the fire, news headlines have accused several banks and their audit companies of misrepresenting their financial positions to investors.

Certified public accountant (CPA) companies facing massive class action lawsuits running into the billions because they provided assurance and opinions on the largest sub-prime credit providers despite notifications of irregularities and missing vouchers from regional accountants. In the press, there are reports that the accounting industry is asking Washington to delay court decisions because otherwise the entire industry could collapse. Is this just another ‘too big to fail?’
Month after month, Americans faced a previously unknown unemployment rate of 10%, and entrepreneurs were been told that there is not enough money in the bank to continue their loan request or for new loans to be available for business expansion. Yet the headlines make it clear that banks appear to have sufficient funds to distribute bonuses to the tune of billions. At least top tier universities, such as Harvard Business School (HBS) in 2009, have apologised publicly for their role in fostering short-term thinking through their curriculums, by having the new Dean, Nitin Nohria, implementing a public oath similar to the one every doctor gives in public after receiving his degree at university. However, for people who have lost their homes or jobs as a result of the recent crisis an apology from Harvard is of no consolation. It certainly did not help avoid the spread of the Occupy movement to Harvard.

Their disappointment and anger towards the Ivy League-educated professionals would at least be reduced if such professionals apologise in a way that is visible and audible to them and other aggrieved parties without any ifs, ands, or buts. They could all learn from scholars, such as Nitin Nohria of HBS.

Michel Ogrizek, a former communications director of a prominent investment bank, refers in the first edition of the ‘Trust Meltdown Report I’ to the problem that many groups of professions are represented on the boards of banks, but not those who understand communication and the feelings of people. The advice to ‘not overdo it with apologies’ must come from extensively-staffed legal departments which, as always in such cases, refer, and will always do so, to the risk of damage claims derived from apologies. But who has asked how much business has, in the meantime, been lost by financial institutions because apologies have not been perceptible to the customer? Expertise in business administration science, often criticised for many mistakes, can, at least on this topic, refer to extensive results in the area of opportunity-cost calculation. The falling share prices and the billions that were withdrawn from a number of these institutions could give bank CEOs and their boards of directors cause for thought the next time they staff relevant posts.

Exhibit 7.4 demonstrates how negative reports on banks in general did not improve, but became even worse in the first year after the financial meltdown.
The reason why, in surveys, Wall Street bankers ‘enjoy’ only 4% esteem amongst all respondents, why school children rarely state ‘banker’ as an aspiring profession, and why politicians from all countries will trim any room for play that financial institutions have with legislation is, however, not only the result of lack of awareness regarding blame and admission of guilt. Faith in the strategic competence of managers had already long been in doubt in opinion-leading media. But it took a while for Wall Street to understand that they had not caught a cold, but had a much more severe ailment, which requires a different treatment (Exhibit 7.5).

In September 2009 at the G20 summit, executives from global financial institutions did not want to bring themselves to a clear global agreement upon commitment on common guidelines on how to have minimum liquidity and bonus structures. Instead, following the G20 summit, most financial institutions seemed to look for the most pleasant regulation, proposing similar conduct to its competitors, and, as they were not prepared to comply, to again completely delay any actions in matters of control and sustainable management. What very seldom succeeds appears to be what bank executives and their boards of directors want to apply to themselves – pointing fingers at each other and then doing nothing.
Times may be a thing of the past, when some banks believed they could prevent unpopular reporting by closing their own press offices, while others believed they could eliminate unpopular reporting in the media by having employees buy up all the relevant editions from kiosks. However, the way in which managers in the finance industry take such little care to ensure that their own press department does not sanitise clippings does not imply that there is a real interest in the facts. In the meantime, it is not only The Wall Street Journal or Financial Times that are closely watching every (in)action taken by managers of these institutions; it is every newspaper, magazine, radio station and TV camera, not to mention online activists. In 2009, the year in which the top management of banks stated that the situation would slowly calm down, Main Street’s judgment of Wall Street was comparable to the reputation of the tobacco industry. Trust in the industry had been badly damaged (Exhibit 7.6).

Invitations to the White House, Downing Street, the Federal Chancellery and all other government leaders’ houses were misinterpreted by Wall Street. Just because governments were now themselves having liquidity difficulties – thereby becoming consumers of bonds, that is, banking customers, by handing out taxpayers’ funds to rescue the banking industry – does not in any way mean that the customer was satisfied or even grateful.

Seldom have more contracts been signed with a clenched fist in the pocket than they were in the aftermath of 2007. The president of Iceland, with his non-signing of the
legislation for repaying monies due, positioned himself beyond any market logic, because he saw his re-election assured and, with his conduct, stimulated other parties to act similarly in the future. President Obama’s legislative initiative in January 2010 followed this example. Other governments will likely follow. If banks in England want a sense of security due to their market dominance, they would be advised to not sleep too deeply. Neither Gordon Brown nor David Cameron could afford, from a political point of view, to not take action against the financial institutions and were, therefore, looking for alternative penalties such as levying taxes on executive bonuses. As Brown failed to clearly communicate on this, Cameron won the election. However, he did not take a clear advantage afterwards because he repeated Brown’s mistake of not finding the right tone to address the public’s frustration towards the banking sector, while also finding a compromise to keep the sector in the City.

At the same time, the crisis has had some positive side effects. What has up until now only been discussed at conferences by experts has finally seen the light for the broader public.
Experts, such as Professors Bob Eccles and George Seraphim from Harvard Business School have been suggesting for years that due to previously inadequate accounting regulations, only 30% of the value of a company is reflected on its balance sheets. In other words, 70% of the performance is assessable by personal contact with the respective company alone. However, companies that reduced their customer consultants for cost reasons, as did the banks, and thought they could replace them with cheap software programs for evaluating creditworthiness, reduced their own ability to make judgments to exactly this 30% figure, which has been what is reflected on balance sheets. This does not have to lead only to misjudgments when assessing private or even corporate loans. The text by Professors Eccles and Seraphim in May 2013 on ‘The performance frontier, innovating for a sustainable strategy’, in *Harvard Business Review* marks a clear milestone and ends a phase in which calls for integrated reporting have only been heard at niche conferences of experts.

Consequently, asset managers and floor traders have little other information access and their decisions are made based on this limited 30%. Thus, it has become known to the general public that the entire business of trading in shares has been based on a very narrow database of the sort that no patient would let their doctor use in making a diagnosis. Additionally, banking CEOs did not distance themselves from this substantial systems error and neither did their supervisors press for improvement. Only after the SEC’s December 2009 decision, to make transparent communication via Extensible Business Reporting Language (XBRL) legally-binding, was the first step to bind disclosure by management interactive.

Banks are not yet on the frontline in order to actively and publicly proceed against auditors who incorrectly gave assurances that financial statements are presented according to generally accepted accounting principles. A major Swiss bank’s decision to recall their own balance sheet within seven days of the publication of their annual report in 2007 also hit the headlines in publications such as NZZ, *Handelszeitung* and others. A number of state-owned and controlled German Landesbanken provided yet another example of fundamentally undermining every basic trust principle, when in 2009 they adjusted their need for financing requirements upwards by several billion euros and within a few days made themselves look pretty bad in the eyes of both credit providers and the general population. The economic impact of the restatement of their annual report, which has remained in the memory of the general public, comes from the fact that almost every newspaper reader is himself a borrower and can well remember his own meetings with his bank’s so-called customer consultants. Upward adjustments are just as well-liked as the submission of incorrect figures, and most people remember these well.

In addition to this divide, which has developed between Wall Street and Main Street, the finance industry will not regain trust unless it starts evaluating companies at their real economic value or be willing to face the risk and liability imposed by regulatory authorities. They should remember that among countless numbers of victims are enough judges, even in the highest courts, who did lose money on these inadequate investments schemes.

Previous references to the uncertainties of stock trading are, however, only now applicable when the person acting in that capacity can prove that they have taken all required aspects into account regarding the sale. But, the public will sooner or later no longer accept this since they are aware that banks only take 30% of all relevant factors when making investment decisions.
Banks have also been caught giving ‘sweetheart loans’ to those that should be protecting the public from fraud and mismanagement by banking institutions, according to *The Wall Street Journal*. For example, Connecticut Senator and Chairman of the Senate Banking Committee, Chris Dodd, received preferential treatment from Countrywide. All this led to an increase in calling for action by the bank supervision authorities (Exhibit 7.7).

Exhibit 7.7

**Banking supervision – share of all reports**

![Bar chart showing the share of all reports on central banks and European supervisory authorities.](chart)

*Source: 113,902 reports in more than 100 international media*

On the 20th anniversary of the fall of the Berlin Wall, many studies were published concerning the construction and conquest of the ‘anti-capitalist barrier’. While in the beginning too little information flow relating to each side did not strengthen the defences, in the second half of the 1980s the web of deceit could not be hidden anymore thanks, in particular, to television. A continuous flow of information in both directions is, after all, the basis for trust.

Experts in the finance industry remain confronted by two questions: (i) why did they not provide the continuous information regarding their own situation which the market
and the customers urgently needed so that they could form their own judgments? And, (ii) if necessary, which could be the alternative supplies that were not as enmeshed in sub-prime or other making money out of money campaigns? At the same time, people would have been grateful if highly paid analysts had given them precise information. The horror scenarios they offered were only intended to get the governments to approve bailout scenarios more quickly than those which would actually correspond to data on the world economy.

People react according to the information offered to them. Accordingly, it can come as no surprise to anyone that the Germans, as far back as the last general elections in 2009, having lost their leading role as the export world champion would place less and less trust not only in banks, but in the entire system. Approval rates for democracy in general and for the social market economy sank by more than 20% in 2009. In view of this information offering this comes as no surprise (Exhibit 7.8).

Things have not changed since then. In April 2013 a new party was founded by a group of highly respected and outspoken economic scholars in Berlin to participate in the year’s general elections in order to give the German voters a choice on 23 September on whether to remove the country from the Euro currency union.

There have been enough examples within the finance sector showing how blue-chips can act as true entrepreneurs with a clear understanding of the needs of all stakeholders: the examples of Allianz and Commerzbank in Germany demonstrate just that. When, more than 12 years ago, Holocaust class-action suits by Ed Fagan dominated media headlines around the globe, managers at Allianz did not hide behind their relatively secure legal positions, but remained available to journalists day after day. They also opened their archives to qualified researchers so that they could understand the mistakes made then and draw conclusions for the 21st century. Similarly open was the communication by the people of Munich concerning the joys and sorrows that their relationship with Dresdner Bank had bestowed on them. For Allianz, openly speaking about their own mistakes is not a sign of weakness but part of their usual dealings with stakeholders. The communication activities of Commerzbank are similar, although everyone understands that the yellows (that is, Commerzbank) will not have to pay entertainment taxes for the years since 2007.

But as long as these examples are more the exception than the rule, fundamental doubts in the social contributions of financial institutions will remain. This will remain all the more true so long as they do not, sooner rather than later, eliminate three capital blocks from the wall between them and the rest of the world.

As it stands in spring 2013, anyone who believes that they can, on their own, offer the market something trustworthy with 800,000 ‘products’ and 500,000 certificates in Germany alone without having these inspected by an independent authority will time and again come up against the same problems.

Additionally, anyone who thinks that they can, without a compelling reason, betray their clients in order to optimise premium proceeds in the short term underestimates communication by victims. Finally, anyone who continues to communicate externally that no qualified person is prepared to come to work to fulfil his duties for less than US$500,000 in annual income, should not be surprised to find that 90% of all professors, doctors and owners of small and medium-sized business owners are upset by their actions.
Exhibit 7.8

Salience and tonality of the Social Market Economy in reports about the state of Germany with regard to political values, ARD and ZDF main evening news, 2001–2013

(a)

Share of all reports

(b)

Source: 7,388 reports about the situation in Germany with regard to political values, 551 focusing on the economic system
So, what should financial services institutions do?

There is a way back to those times in which bankers were perceived so positively that the representatives of leading finance institutions in each town were actually viewed similarly to priests, doctors, or school teachers.

Professor Robert S Kaplan explained in his recent book why he no longer works at Goldman Sachs and is back at Harvard Business School. Whoever is unable to bear to see his mirror image must, in principle, change. This change has to be both fundamental and transparent. Michael Diekman, the CEO of Allianz SE, addressed this issue recently in a simple way in front of his top managers, saying, ‘In case you are not sure, if a product, project or any other of your activities are good or bad, just quickly imagine this will become a front page story on WSJ’.

As this chapter has to focus on the reputational side of the problem, we will not elaborate too much on the urgent need for the banking sector to implement better accounting standards, making sure integrated reporting becomes, sooner rather than later, the new gold standard. Making money out of money is definitely no longer a path any bank should walk – as reputation follows facts.

Therefore, the dramatic reputational loss for the banking sector will not be fixed by continuing with questionable products and non-transparent balance sheets but by no longer operating as self-centred and acting with the media based on respect. As the BP example shows: a US$100 million advertising campaign will not help build trust on a long-term bases – just ask their stakeholders, how this helped them during the last oil spill.

Let us start with the most obvious of issues that the financial services sector needs to deal with. Without respecting corporate governance rules such as a clear split between Chairman and CEO, there is no need to even think about gaining trust among those who have been waiting since 2007 in order to see the financial services sector recognise its role in the crisis. Checks and balances are not only a basic need in politics but even more in the corporate world.

But let us move to the centre of the trust meltdown: without changing the remuneration system the banking system will lose their licence to operate sooner rather than later. Bonuses paid while a company produces losses is simply a no-go, no matter what the traditions in the banking sector have been. Deutsche Bank, under leadership of its new chairman, Dr Paul Achleitner, created a commission with the former CEO of BASF, Dr Jürgen Hambrecht, and the former Minister of Finance, Dr Theo Waigel, at the top in order to implement new standards not only for the board, but for the first and second levels of executives. A reasonable percentage of the bonuses as of 2013 will be linked to clear and transparent improvement from the top management towards improving the banks’ reputation among the key stakeholders as publicly discussed during their shareholder meeting 23 May 2013. One cannot imagine a stronger symbol illustrating the move towards a clear and transparent stakeholder position. Others, like Barclays or HSBC are working towards similar definitions. UBS, under the new leadership of the former German central bank president implemented different strategies, but with a clear target in common. Without getting the public to understand how a bank is defining its bonuses the house will vanish sooner or later.
The keynotes at the 2012 10th Corporate Governance Conference in Berlin formulated this more precisely. Whether it was the president of the German Parliament, Norbert Lammert, or the Chairman of Commerzbank, Klaus-Peter Müller, the mirror that they held up to the business elite in Germany and elsewhere showed managers the subtle difference from the entrepreneur: in *dubio pro* company. Only when all stakeholders of the company are under the impression that the owner has served the company and the community in an exceptional way does he get paid a bonus.

In ancient Greece, as it still is, it was a tradition to ask questions of the Oracle of Delphi. The truth sometimes hurts, especially if it is so easy, as in the case of bonus rules, to see the problem. Anyone who has to hide their payments in three different places in an annual report must know that they are not deserving of what they are being paid. In today’s world, the Oracle of Delphi sits at every news kiosk: the headline news provides a clear indicator for the do’s and do not’s. And this is not just about tabloids: both *Financial Times*, the English as well as the German, have been clear with regards to the writing on the wall (Exhibit 7.9).

---

**Exhibit 7.9**

**Tone of coverage – banking industry**

All topics January 2003 to September 2012.

*Source: FT* 191,144 stories, *FT Germany* 139,244 stories
It has been the decision of some Wall Street executives to evade this tough question. To be crystal clear: it is not a question of the amount of money one receives as a bonus. Neither Michael Schumacher, Tina Turner, nor Warren Buffett have such problems. Their fans, and even more importantly, society, know why and how much they receive for their activities at the end of a year. As soon as managers achieve this kind of transparency and satisfaction based on performance and ongoing information this debate will come to an end.

Banks expect to see companies’ balance sheets each time prior to granting loans, and rightly so. But they themselves believe that they will not be able to convince stakeholders of their performance unless they have perfect figures. The balance sheet practices of banks were criticised in the media long before Enron, and yet the industry did not implement any improvements. Anyone’s products that are classed as toxic must come to terms with the debate as to whether they are even worth their salaries, let alone any bonuses. Banks are remarkably silent on the current interest rate crisis as well as on speculations regarding the demise of others. But silence never pays off, especially in times when the public is confronted with negative news. In 2012, the media used the term ‘toxic’ more often in connection with products from the finance sector than in connection with products in the chemical industry.

The trust gap is obvious, because expecting others to eat dry bread while feasting themselves has never and will never win anyone over. The financial world presently also lacks a clear self-image. Banks are not the heart of the economy, but are more like the kidneys. The heart is actually an intact legal system. But as long as the finance sector believes that they are the heart of the economic system it gives a false comfort to hope that no change is needed.

The expectations of every director have changed significantly in the 21st century. While positive development of the share price and significant increases in profits were enough in earlier times for directors to expect bonuses, the demands placed on directors since the bursting of the internet bubble and, in particular after 2008, have changed. After all, it is common knowledge that the development of the share price is not just the result of a CEO’s actions.

Even profit expectations are relativised in times of crashing markets and are unfair if they serve purely as a basis for the payment of a bonus. Qualitative, sustainable and, above all, measurable growth in satisfaction amongst the seven key stakeholders is in demand in the 21st century. Only those who, as directors, measurably motivate their own employees to perform better, who drive the further development of core markets forward, whose performance convinces financial market analysts, who make the necessity of their own company clear to politicians, NGOs and experts, and who provide ‘all’ the relevant information via the media, will earn a bonus this year.

**Bonus category 1: employee satisfaction**

Most employee surveys are not worth the paper they are printed on. The irrelevance of the exercise is too obvious to all the participants, as the results are actually published in full, discussed from all angles, and result in consequences in only 1 in 10 cases.

Whoever surveys employee satisfaction and uses the results as a basis for paying their CEO a bonus, sends a clear signal to all that this is not just mere lip service but a transparent assessment of the company’s leadership that is linked to a concrete figure in euros in the end.
The data basis must be correspondingly complex; just one survey every six months will not suffice. Voluntary participation, the ratio of improvement suggestions, and so on, must also be included in the overall assessment. And, last but not least: the feedback system must be visibly put into practice for all.

**Bonus category 2: key markets satisfaction**

Key markets should not only include the five most important areas according to sales, earnings and growth rates, but also the five most relevant growth regions of the last three years.

But, instead of analysing only the conventional business key figures as the basis for this bonus portion, acceptance indicators, without which current positions should not be held permanently, relating to stakeholders from politics, business, science and media should also be included. Toyota and Siemens in the US have demonstrated the risks inherent in the absence of cross-linkages in the most important sales and production regions.

As a benchmark, not only other DAX 30 companies or even only industry competitors should be included, although there should be clearly definable and measurable indicators such as keynotes at relevant events of the respective country, guest contributions in opinion-leading media of the country and participation in TV talk shows.
Bonus category 3: financial markets satisfaction

Quoted analysts now dominate the overall reputation of almost all DAX 30 companies with a share of significantly more than 15% of the overall media image. For this reason alone, not only analysts’ satisfaction with the company but, above all, the quality and speed of the information provided should be included in the bonus assessment.

Only in this way can a situation be avoided where statements not based on fact are glossed over. In addition, the extent to which a company deviates from industry and cross-sectoral standards in its reporting should also be included.

As stakeholders were recently informed of the rulings against Deutsche Bank, over a period, via the media, consideration should be given as to whether this criterion should be classified as the absolute key criterion. The low recommendation rate (only 12% in autumn 2011 by Business Elite) is a fatal blow.

Bonus category 4: politics satisfaction

From time to time, politicians are depicted as surgeons who have to keep the banks alive with emergency operations, as in 2008. Certainly, nobody wants to have a bad relationship with those who could soon turn the scalpel against them. Especially the political shareholder who has since not had its interventions credited as an asset item.

Likewise, satisfaction figures to be newly collected at the political levels should be included in the bonus criteria. This would ensure that, in addition to the many different board positions at the local, national and international level, monitoring would also be included, which would include directors appearing at parliamentary hearings, and so on.

This is not just about measuring how relevant the company’s expertise is for politicians, but about practicing for a worst-case scenario. The Murdoch example has shown how quickly a licence to operate can be lost due to incorrect actions before political committees.

Bonus category 5: NGO/NPO satisfaction

BMW has set a new international standard with its involvement in the UN Alliance of Civilisation 2011. Its participation in a UN unit for religious dialogue as a global player shows that the top management in Munich has an understanding of social developments. The operating climate at the conveyor belt has long not been dominated solely by debates on 35-hour weeks or gender balance. Whoever cannot offer a clear position on the challenges important to society in 2013 will be overtaken by trends quicker than they would like.

Monitoring directors’ dialogues with as many NGOs as possible gives insight on the extent to which the top leadership has prepared the company for all contingencies. This dialogue cannot be delegated to in-house charitable foundations. Whoever wants to succeed in China but thinks they do not need to dialogue with Greenpeace was wrong as far back as 1986.

Bonus category 6: science satisfaction

Whoever, as CEO, is not invited to specialist lectures at top universities must ask the question as to how real his/her thought leadership is. In a world in which neither water nor oil,
specialist personnel, is the scarcest resource, the presence of a company’s top leadership at schools and universities around the world is of central importance for securing the future of the company. These appearances and their effects are measurable in human resources. This is not an exercise in collecting irrelevant PhDs. Rather it is about being concretely quoted in the media or scientific specialist bodies on talks and technical contributions in the scientific cycle.

**Bonus category 7: media satisfaction**

Irrespective of the organisation, whoever during times of clarification communicates below the awareness threshold puts the whole company’s licence to operate at risk.

Press clippings such as key media reputation figures are not provided by a company’s own media department but by independent third parties that are themselves subject to scientific controls. Asking the chief press officer whether a company’s image is good only makes sense if the CFO is able to define his own share price. The same applies to all external and internal communications with the same professionalisation claim as other specialist departments model it before him.

When an executive, such as a CEO, communicates for himself as well as for the company, the company’s positions, products and concepts above the awareness threshold in all relevant media with a share of voice of 35% he has earned his bonus.

**Conclusion**

This chapter looked at how financial institutions have been managing their reputational risks over the years and discussed whether their strategies have been effective.

Having demonstrated the difficulties of financial institutions to fully recognise the extent to which their reputations have become damaged in recent years, we suggested seven standards to help them regain some of their lost credibility.

Once a bank implements these standards, it will be perceived more and more as a transparent and reasonable part of the corporate community. It will regain the status which is the condition *sine qua non* for the whole sector: trust. Without this, the rest of the society will just wait for alternatives and once they materialise, the banking sector of today will face the same experiences as the transport system 100 years ago. The horse industry was not helped by referring to the German Emperor Wilhelm 1, who said, ‘I think these cars are heavily overrated – I will stick to my horse’. Or as Bill Gates put it more recently, ‘banking cannot be replaced, the banks sure’. Just look at the growth of telecoms offering banking products. This is a trend no longer restricted just to Africa.

---

1 Available at: www.whitehouse.gov/blog/2010/01/21/president-obama-never-again-will-american-taxpayer-be-held-hostage-a-bank-too-big-fa.
162

Senator Elizabeth Warren’s letter to the Justice Department, the Securities and Exchange Commission and the Federal Reserve.

Available at: www.exed.hbs.edu/assets/Documents/CorporateDNAReport2012.pdf.

Available at: www.sec.gov/spotlight/xbrl.shtml.


Based on private conversations between the author and the senior management of the institution.


Rupert Murdoch had to stop publishing his newspaper, News of the World, in 2011.
Part 2

Regulatory and enterprise risk
Chapter 8

Theories of regulatory risk: the surprise theory, the arbitrage theory and the regulatory mistake theory

Jonathan R Macey
Yale Law School

Overview

This chapter first provides a taxonomy for use in analysing the various theories of regulatory risk and then proceeds to analyse each type of regulatory risk. Theories of regulatory risk take the form of the ‘surprise’ theory of regulatory risk, the ‘regulatory arbitrage’ theory of regulatory risk and the ‘regulatory causation’ theory of regulatory risk. The surprise theory posits that unanticipated regulatory changes will have an adverse effect on a company’s performance. Regulatory arbitrage refers to the risk that those companies affected by regulation will employ various tactics to avoid regulation that they view as costly or burdensome. The third form of regulatory risk, which we call regulatory causation, is the risk that regulation, by forcing companies to conduct business or to analyse risk in a single uniform way, can actually increase the levels of systemic risk in the economy by causing herding. In other words, by reducing the heterogeneity of company behaviour in an economy regulation – even regulation aimed at reducing systemic risk – can increase such risk. This risk is vastly exacerbated when regulators make errors, because these regulatory mistakes affect all companies, and therefore may have systemic consequences.

In analysing these forms of regulatory risk, we use public choice theory to show that neither the surprise theory nor the regulatory arbitrage theory of regulatory risk is truly a form of risk, because both regulation and regulatory arbitrage are the anticipated outcomes of a bargaining process in which the companies affected by the regulation are active participants. As such, it is inaccurate to assert that either regulation itself, or companies’ responses to regulation, come as a surprise to the companies and industries affected by the regulation. Regulation may come as a surprise to the rationally ignorant general public, but not to the relevant industry groups or to other discrete, well-organised interest groups. Finally, in this chapter we claim that significant regulatory risk does, of course, exist and that such risk manifests itself in the form of the macroeconomic risks caused by regulations that affect not just individual companies but the entire economy. This theory is dubbed the regulatory causation theory of regulatory risk because the regulation is what is causing the risk. Such risk is not only unforeseen; but it also cannot be mitigated by individual companies. Regulatory
causation risk results from the herding behaviour that regulation causes. In order to illustrate this point, we discuss various regulations, including minimum capital requirements for banks, regulations concerning Value at Risk (VaR) models for financial institutions, and the rules regarding credit rating agencies.

**Introduction**

The concept of regulatory risk is poorly understood and has been described in diverse ways. In fact, there is no universally recognised definition of regulatory risk. At least three different conceptions of systemic risk have been advanced by regulators and academic commentators.

The first and perhaps the most widely known form of regulatory risk is the risk that unanticipated regulatory changes will have an adverse effect on a company’s performance. This form of regulatory risk can be dubbed the surprise theory of regulatory risk because it posits that the increasing scale and pace of regulation is unparalleled and that there are risks that the burdens and compliance costs of new regulation are a significant threat to banks.

The surprise theory reflects a number of flawed assumptions. Most prominently it reflects the assumptions that regulations occur as exogenous events over which the companies regulated have no influence or input. Similarly, the surprise theory takes no account of the fact that industry input and back-and-forth bargaining have significant impact on regulation. In other words, the problem with the surprise theory is that regulation rarely comes as a surprise to the companies and industries most affected by the regulation.

The second form of regulatory risk is called regulatory arbitrage. This refers to the risk that companies affected by regulation will employ various tactics to avoid regulation that they view as too costly or burdensome. Regulatory arbitrage can be accomplished using a number of strategies, such as structuring mergers, leveraged buyouts and other transactions to avoid the consequences of certain tax or bankruptcy or licensing rules. Various forms of financial engineering and even decisions about where to locate subsidiaries and plants and corporate headquarters are made in order to affect regulatory arbitrage.

In one important respect the ‘theory of regulatory arbitrage’ is the mirror image of the surprise theory of regulation. By this we mean that while the latter posits that new regulations surprise even sophisticated market participants, the former suggests that regulators are surprised by the reactions of market participants to regulation. In this way, the arbitrage theory of regulation confuses what is known as the ‘law of unintended consequences’ with the foreseeable reactions by regulated companies and their lawyers to new regulations such as taxes. The ‘law of unintended consequences’ posits that the actions of government generally have one or more unintended consequences and harks back to the work of the philosopher, John Locke. Locke, writing towards the end of the 17th century, opposed a suggestion by Sir Josiah Childs that the legal rate of interest be lowered to 4% from 6% in order to help borrowers. Locke argued that lowering the maximum interest on borrowing would impose more costs than benefits on the weakest members of society due to the unforeseen consequences of the regulation. In particular, Locke pointed out that the law would shrink the supply of available credit and reduce the income of ‘widows, orphans and all those who have their estates in money’. Of course, a single regulation can have consequences that are unforeseen to some but that are entirely foreseen to others. And, while it certainly is the case
that many regulations have consequences, sometimes dire consequences, that were unforeseen and therefore unintended, it is highly unlikely that the loopholes and exceptions that provide opportunities for regulatory arbitrage are systematically unforeseen. In fact, it is far more likely that these loopholes and exceptions were entirely foreseen and were indeed part of the special interest ‘deal’ that was struck in order to get the statue passed in the first place. Moreover, even the most unforeseen loophole or exception does not remain unforeseen after companies have begun to exploit them. Thus, regulatory arbitrage ultimately cannot exist without the co-operation, or at least the acquiescence of regulators who observe it and, for whatever reason, do nothing to stop it.

The third and most important form of regulatory risk is the risk that regulation, by forcing companies to conduct business or to analyse risk in a single uniform way, can actually increase the levels of systemic risk in the economy by causing herding. In other words, by reducing the heterogeneity of company behaviour in an economy regulation – even regulation aimed at reducing systemic risk – can increase such risk. This final form of risk is best described as the regulatory causation theory of regulatory risk. This risk is vastly exacerbated when regulators make errors, because these regulatory mistakes affect all companies, and therefore may have systemic consequences.

In this chapter we argue that this three-part taxonomy of regulatory risk is conceptually flawed. Under ‘Public choice theory and regulatory risk’, we argue that the first two forms of regulatory risk described above (the risk that unanticipated regulatory changes will have an adverse effect on a company’s performance and the risk of regulatory arbitrage) are not really risks at all for a variety of reasons. Sometimes companies favour regulation, and when they do not favour regulation they can find ways around it. Regulation in this context is not a risk, it is a commodity. Drawing on some of the basic lessons of public choice theory, we point out that regulation is supplied and demanded like any other commodity or service, and that bureaucrats are not saints but human beings who pursue their own rational self-interest either simultaneously with or instead of pursuing some notion of the public interest or broad social welfare.

Under ‘Regulatory causation risk’, we argue that the real risk associated with regulation is a form of macroeconomic risk that is unforeseen, and, because it is a form of systemic risk, cannot be mitigated via a diversification strategy. This risk results from the herding behaviour that regulation causes. In order to illustrate this point, we discuss various regulations, including minimum capital requirements for banks, regulations concerning Value at Risk (VaR) models for financial institutions and the rules regarding credit rating agencies.

Public choice theory and regulatory risk

The theory of public choice, also known as the economic theory of legislation, makes the same basic assumptions about self-interest for politicians and bureaucrats that standard economic analysis makes for private sector actors. Thus, public choice theory views regulation as a commodity like any other. As Richard Posner has observed, ‘the interest group theory asserts that legislation is a good demanded and supplied much as other goods, so that legislative protection flows to those groups that derive the greatest value from it, regardless of overall social welfare’. According to the public choice theory of legislation, market forces
provide strong incentives for self-interested politicians to enact laws that serve private rather than public interests because, over a wide range of issues, these private groups can provide politicians and bureaucrats with the political support they need to serve their objectives of achieving re-election, or of maximising their bureaucratic turf. In a nutshell, public choice theory posits that laws and regulations are supplied by lawmakers and bureaucrats to the political groups or coalitions that outbid competing groups.

Public choice theory has attracted a wide variety of adherents among welfare state liberals and Marxists, as well as among free-market economists. In general there has been a ‘shift in scholarly thinking about legislation from a rather naive faith in the public interest character of most legislation to a more realistic understanding of the importance of interest groups in the legislative process’. The widespread acceptance of public choice theory is linked to increasing suspicion about much of what Congress does. The public choice literature has attempted to specify more precisely the process by which organised special interest groups obtain rules that transfer wealth from weaker political coalitions to themselves. Concentrated interest groups demand special benefits for themselves. Individual citizens and less organised groups supply this legislation by paying the taxes and incurring the costs of complying with the regulations that transfer wealth to such groups. Politicians act as brokers between these two groups and as entrepreneurs.

The politicians’ goal is to maximise their own political support, most obviously by passing legislation designed to appeal to particular groups. The politicians, however, can also use more creative tools. They can, for example, define issues around which newly formed groups can coalesce, and they can devise laws that overcome the organisational obstacles of high information costs and transaction costs (such as free rider problems) among interest group members that plague wealth-transfer seeking interest groups.

While interest groups compete in a political marketplace, legislative institutions behave like private-sector companies whose output is law. As such, the theory of the company, rather than the theory of market exchange, guides the public choice analysis of legislative institutions such as Congress. Like all companies, Congress organises its internal affairs to minimise the costs of assuring contractual performance. Perhaps the most acute bargaining problems facing politicians are those arising from non-simultaneity of performance. Politicians generally will attempt to obtain political support from interest groups before an election, but will not be able to ‘pay for’ this political support by supporting legislation favoured by the interest group until after the election. Interest groups, therefore, have reason to worry that politicians will renege on their promises of support. Conversely, incumbent politicians often will be called upon to provide political support for a particular bill favoured by an interest group in exchange for a promise of future support by the interest group.

Politicians, therefore, have reason to worry that the groups will renege on their promises. The most serious simultaneity of performance problem results from the fact that, even after an interest group has succeeded in achieving enactment of a particular statute, there can be no promise that future legislators will not renege on the previously agreed upon legislative deal, particularly if the original configuration of interest groups loses power. Much of the ‘industrial organisation’ of Congress is designed to deal with this simultaneity of performance problem. Congress has a strong incentive to resolve this problem because it would otherwise be difficult for Congress to make a credible commitment to an interest
group that a particular legislative scheme will have the crucial characteristic of durability. More durable statutes and regulations will be worth more to politicians than less durable statutes and regulations because interests groups are willing to pay for durability. The basic way that Congress and regulators deal with this non-simultaneity of performance problem is by making it difficult to pass legislation in the first place. The more difficult it is to pass legislation, the more difficult it will be to repeal legislation. In addition, the committee system was arguably developed in order to concentrate legislative power in the hands of a small number of people who would be closely linked to the interest groups associated with particular legislation. Thus, congressmen from farm states (and, more recently, from districts heavily populated by food stamp recipients) are disproportionately represented on the agricultural committees in Congress. This, in turn, means that as long as the farm lobby controls the congressmen from the farm states and the natural alliance between farmers and food stamp recipients holds, these interests will be able to block the introduction of legislation adverse to them.

These types of rules permit congressmen to make credible commitments to interest groups that laws passed now will survive in future legislative sessions. Similarly, the bicameral legislature, executive veto and independent judiciary all address this non-simultaneity of performance problem by moving the voting rules for passing new laws closer to a unanimity requirement, thus making laws difficult to modify once enacted. Politicians can also establish administrative agencies to make legislation more durable. Administrative agencies make legislation durable by creating a stable of professional bureaucrats whose own futures are inextricably linked to the maintenance of a particular regulatory regime. Thus, for example, the securities laws and the banking laws benefit certain industry participants both by creating barriers to new entry, and by creating other rigidities favourable to incumbents. These laws also created significant bureaucratic structures that have made reform and repeal of these laws surprisingly difficult.

The preferences of the majority of voters are virtually irrelevant in determining legislative outcomes when viewed within the public choice framework. Instead, laws are made by legislators whose primary goal is to aggregate the political support they receive from competing special interest group coalitions, and legal rules are generated by a process that conforms to standard microeconomic models of rationally self-interested behaviour.

Political scientists have used the terms ‘coalitional drift’ and ‘bureaucratic drift’ to describe with more precision how non-simultaneity of performance problems between politicians and interest groups can arise. Bureaucratic drift describes the problem where the high costs of monitoring and controlling the behaviour of bureaucracies leads to situations in which bureaucrats will act in ways inconsistent with the original deal or ‘coalitional arrangement’ struck between interest groups and politicians. Coalitional drift describes the contracting problem caused by fluctuations in the preferences of the electorate and shifts in the preferences of politicians. The problem of coalitional drift manifests itself when Congress wishes to undo deals struck between interest groups and prior legislatures.

Generally speaking, the cure for bureaucratic drift is ex-post control over bureaucratic behaviour by congressional subcommittees, oversight by specialised agencies such as the Congressional Budget Office and the General Accounting Office, and reliance on interest group notification. This oversight is supplemented by legal requirements that agencies provide
information about themselves to their political watchdogs. Similarly, micro rules (like the prohibition on ex-parte communication that enables politicians, but not interest groups, to gain direct one-way access to administrators) and macro rules (like congressional control over agency funding) permit politicians to control bureaucratic drift. The bottom line is that Congress and interest groups structure the administrative process in order to permit interest groups to preserve the benefits of the prior deals they have struck in the face of recalcitrant bureaucrats. The lesson of bureaucratic drift is that when bureaucrats misbehave, Congress has the ability to respond in a wide variety of ways.

But what happens when Congress lacks the will to respond? This is the problem of coalitional drift, which describes what happens when Congress’s own preferences change. The danger to legislator-interest group deal-making that this will occur is much more serious than the danger that Congress will be willing, but unable, to control behaviour by bureaucrats that they find undesirable. As Shepsle has observed, over some range of administrative outcomes, efforts to control bureaucratic drift by empowering Congress to micro-manage agencies will exacerbate the problem of coalitional drift. The more that Congress and congressional staff are free to interject themselves in the bureaucratic decision process, the smaller the problem of bureaucratic drift, but the greater the problem of coalitional drift. The more independent the administrative agency, the smaller the problem of coalitional drift, but the greater the problem of bureaucratic drift.

To some extent, both of these problems can be ameliorated simultaneously by structural decisions about the administrative agencies themselves. Initial agency structure and design can reduce coalitional drift and bureaucratic drift simultaneously by ensuring that particular interest groups will receive the funding needed to remain strong, thereby enabling such coalitions to maintain their control over Congress, thus preventing coalitional drift. Initial agency design can likewise prevent bureaucratic drift by creating a specialised bureaucratic structure in which bureaucrats have a single clientele like the savings and loan industry. This, in turn, creates a situation in which bureaucratic drift is unlikely since interest groups will capture the administrative agency, fill it with their own nominees, and then run it as a personal fiefdom. This not only helps ensure the fidelity of the administrative agency, but also the longevity and continued vitality of the interest group as well.

Of course, this happy outcome (from the interest group’s perspective) is not inevitable. Where, as in the savings and loan situation, the perverse incentives created by a particular regulatory structure are simply overwhelming, the system can collapse under its own weight, ruining both the agency and the industry it was designed to protect. The story of the savings and loan industry and its capture of the thrift regulators represents a textbook case of the way Congress can exert ex-ante control over the outcomes generated by an administrative agency. Congress accomplishes this by controlling the ability of other outside groups to exert political pressure on the agency and by reducing the incentive of other groups to form opposition to the agency’s behaviour.

Congress did not have to create a single regulator for the thrift industry. It could have consolidated thrift regulation with banking regulation generally. In creating a single-interest-group agency, Congress ensured the capture of the agency by the group. By contrast, where the original legislative enactment involves an administrative agency which allows access to a broad cross-section of groups, capture by a single group is much less likely. Thus, the
creation of single-interest administrative agencies is one mechanism legislators use to promote deals that have a significant degree of durability.

In other words, regulation is the product of a negotiation between special interest groups and the politicians and regulators who promulgate such regulations. In other words, regulation is not an exogenous phenomenon that occurs outside of the observational field of the companies being regulated. Regulation is not something that emerges suddenly fully formed from the regulators.

Consistent with one of the principal insights of this school of thought, politicians, bureaucrats and others involved in the policy-making process should be perceived as rational economic actors who, subject to a variety of constraints, are expected to act in their own self-interest, rather than some vaguely defined conception of the private interest. Students of public policy and others interested in improving the quality of regulation and policy formation should understand the incentive structure under which policy-makers and regulators operate.

In other words, regulation does not come as a surprise to anybody. It is the product of a bargaining process. These basic insights of public choice theory should cause a fundamental change in the way that we view the form of regulatory risk that posits that companies are somehow ‘at risk’ of the effects of unforeseen regulatory actions. Because regulation is a commodity that is bargained for, it is inaccurate to view it as a source of risk for at least two distinct reasons. First and most obviously perhaps, because regulation affecting financial institutions and other companies that operate in regulated industries get precisely the level or regulation that they bargain for, subject to certain foreseeable constraints. The ubiquitous legal rules requiring that lawmakers and regulators: (i) operate in the open (so-called sunshine provisions); (ii) give notice before regulating; (iii) permit affected companies to comment on proposed regulations; and (iv) routinely hold congressional hearings allowing those affected by regulation to make their views known, means that as a matter of fact new regulation does not come as a surprise to companies that are paying the slightest bit of attention to what is going on around them.

In fact there is an entire industry, known as the political intelligence industry, which is devoted to ferreting out even the most closely guarded information about government regulations. For example, on 1 April 2013 the Centers for Medicare and Medicaid Service announced that it was reversing a previous decision to cut funding for health care services. Before the change of policy was announced an email alert predicting the new approach was sent by a Washington-based investment-research company, Height Securities, which sent insurers’ stocks sharply higher. Height Securities is a political intelligence company. It profits by ‘connecting investors (and companies affected by regulation) with Washington insiders’ who have such information.15

Second, to the extent that regulation leads to high levels of industry consolidation, it generally increases profits. Thus, because regulation generally provides benefits to regulated entities that outweigh their costs, it generally turns out to be a net benefit rather than a net cost to the companies it affects. In this respect, regulatory risk is different from other risks. Regulation increases profits for the largest companies by raising fixed costs. This forces smaller companies to grow through mergers, leading in turn to consolidation and anti-competitive pricing opportunities. The effects of regulation, whether in the form of new or revised regulation that imposes marginal costs on existing companies does not come
as a surprise to the companies that are subject to the regulations. In fact, these companies generally have significantly influenced the content of any regulations that are promulgated that might affect their business.

In other words, regulation is epiphenomenal in the sense that it results from and accompanies extensive commentary and lobbying and influence by the groups that will be the subjects of the regulation. In fact, regulation is often demanded by the companies subject to the regulation, it is never merely foisted on an industry by regulators who have not first registered the concerns of various interest groups and adjusted their regulations accordingly. And, as with traditional regulation, regulatory arbitrage is not properly characterised as a risk to the extent that it is either: (i) part of the underlying ‘interest group bargain’ with lawmakers that generated the initial regulation in the first place; or (ii) accepted by regulators after they become aware of it ex-post.

Regulatory arbitrage is a core specialty of US banking and financial services lawyers. Regulatory arbitrage involves structuring or characterising a particular transaction or location decision in order to avoid the regulatory consequences of alternative structuring or location decisions. Another term for regulatory arbitrage is ‘loophole lawyering’, which involves exploiting the loopholes contained in particular regulatory regimes as well as exploiting the loopholes created by differences among regulatory regimes in different jurisdictions. An important aspect of regulatory arbitrage is that regulatory arbitrageurs do not only search for jurisdictions with less intrusive regulation, they also sometimes search for jurisdictions with better reputations for investor protection or less corruption or more regulatory stability.

**Regulatory causation risk**

While the risks associated with regulation and regulatory arbitrage are overstated, there is one form of regulatory risk that is clear and tangible, and this sort of risk is in all likelihood the most significant form of risk there is. The most significant and acute form of regulatory risk is the risk that regulators will make errors when they regulate, that these errors will be reflected in regulation, and that these errors will manifest themselves not just in a few companies with idiosyncratic business practices, but in all of the companies in a particular segment of the economy (such as finance). We call this form of risk regulatory causation risk.

This, in fact, is the story of the financial crisis that began in 2007. A number of regulatory errors led to the financial industry operating in lock-step, which in turn led to systemic failure in the financial sector. Here, we identify two examples of mistaken regulatory action that generated significant macro-economic risk; the first is the regulation of credit ratings and the second is the regulation of VaR models.

If these mistakes had not been made, the financial collapse of 2007 would not have occurred. Globally, bank regulators operating under the auspices of the Bank for International Settlements (BIS) generated flawed capital rules that led all banks to be severely undercapitalised. In the US, the Securities and Exchange Commission (SEC) promulgated regulations that led to vast over-reliance on the credit ratings generated by two credit rating agencies, Standard & Poor’s and Moody’s. And bank regulators further exacerbated systemic risk by assimilating flawed VaR models that led to macroeconomic risk in the form of excessive risk taking.
Credit ratings and systemic risk

In a series of seminal articles, Frank Partnoy explains the way regulation has transformed the useful but rather narrowly focused credit rating industry into a regulatory-enhanced juggernaut. As Partnoy points out, the regulatory environment created by the SEC and perpetuated by a host of other state and federal regulatory agencies conveys significant regulatory benefits on those credit rating agencies that have been fortunate enough to receive the SEC’s designation as nationally recognised statistical rating organisations (NRSROs). Partnoy explains that credit rating agencies are distinctly characterised by an ‘oligopoly market structure that is reinforced by regulations that depend exclusively on credit ratings issued by nationally recognised statistical rating organisations’.16

Prior to the intervention by the SEC, the credit rating agencies enjoyed a significant boom during the 1920s and then fell into a period of decline in the wake of the 1929 stock market crash.17 That decline persisted until the early 1970s, when the SEC began to promulgate a series of highly technical and obscure regulations that transformed the credit rating agencies into powerful monoliths in the classic sense: rating agencies with the NRSRO designation that are massive, unchanging and difficult to deal with on a human scale.18 In other words, the SEC’s regulations led directly to the dominance of a small number of credit ratings agencies, and this in turn created massive systemic risk as investors, like lemmings, gravitated towards the rated investments with the highest returns.

The rating agency morass developed in the following way. In 1975, the SEC imposed a uniform ‘net capital rule’ on broker dealer companies. The purpose of this rule was to insure that broker-dealer companies regulated by the SEC would have sufficient resources (capital) to meet their financial obligations to customers, counterparties and creditors.19 The SEC’s strategy for ensuring that brokers had sufficient capital to meet their obligations was to require ‘every broker-dealer to maintain at all times specified minimum levels of liquid assets, or net capital, sufficient to enable a company that falls below its minimum requirement to liquidate in an orderly fashion’.20

It is important to note here that regulating capital is notoriously difficult to do. One of the bigger problems facing regulators is that not all assets are alike, particularly with respect to characteristics such as liquidity and risk. This problem is compounded by the fact that, unlike returns, which are easy to measure accurately in real time, risk and liquidity are extremely difficult to evaluate at all, and accurate measurements are virtually impossible, particularly in real time.21

On a continuum measuring risk and liquidity, at one extreme would be short-term debt obligations of the US government, which are highly liquid and virtually riskless. At the other extreme would be such things as unique individually negotiated commercial loans from individual banks to small, obscure companies and complex, untraded derivative instruments. It would make no sense whatsoever to require companies to maintain the same amount of capital to support assets such as cash and US government notes as they must maintain to support assets such as consumer loans, commercial loans and investments in complex derivatives.

The SEC’s solution to this problem was to transform an existing private sector institution into a new quasi government institution. The old institution was the credit rating agency. The new institution into which the old credit rating agencies magically were transformed

www.rasabourse.com
was the NRSRO. In an apparently well intended effort to inject some subtle gradations into its net capital rules, the SEC decided that bonds and other debt obligations held by broker dealers that had high ratings from an NRSRO were safer (more liquid and stable) than other unrated obligations. As such, the SEC reasoned, broker-dealer companies were allowed to hold less capital for the highly rated assets on their balance sheets than what was required to offset the unrated (or poorly rated) assets on their balance sheets.22

And, because capital is very expensive for financial companies, especially relative to debt,23 it was less costly for companies to hold qualifying NRSRO rated securities than other assets, all else equal. The new net capital rules thus created an incentive for banks to invest in highly rated NRSRO securities, thereby raising the importance of credit ratings to issuers.

The use of the NRSRO designation to determine how much capital a broker-dealer was required to have in order to comply with the net capital rules was followed by an even more profound regulatory co-option of the private sector role of credit rating agencies. Here, the obscure regulation is SEC Rule 2a-7, which pertains to what are arguably the most important and fragile financial institutions in the SEC’s domain: money market mutual funds.24

Mutual funds are investment companies whose assets consist of investments in securities issued by other corporations. Money market mutual funds are a particular subset of mutual funds that compete with commercial banks by holding themselves out to the public as offering stable asset prices that feature safe, stable repositories for liquidity.25 Money market mutual funds compete with banks’ current accounts by maintaining a stable net asset value of US$1.00 per share. This, in turn, permits investors in these funds to enjoy check-writing privileges while still obtaining more competitive rates of return than are often available from banks’ current accounts.26

Money market funds are widely used by individual consumers and by corporations. Money market funds’ assets are extremely important sources of short-term liquidity for investors and borrowers around the world.27 Institutional investors use them as a cash-management tool.28 Money market mutual funds are by far the largest customers for commercial paper and repurchase agreements (repos) in the world.29

In observing the emergence of money market mutual funds onto the mutual fund landscape, the SEC felt compelled to devise regulations that would limit the ability of mutual funds to deceive investors by calling themselves money market mutual funds but not investing in the very high quality and highly liquid assets that would enable the funds to be able to withstand large-scale efforts by investors to obtain liquidity by cashing in (redeeming) their investments simultaneously.

The SEC ‘solved’ this problem by promulgating SEC Rule 2a-7. Rule 2a-7 restricts the kinds of securities that funds calling themselves money market mutual funds can invest in. The purpose of the rule is to make sure that the mutual fund invests in assets of sufficient quality and liquidity that the fund will be able to maintain a stable net asset value of one dollar even in times of significant stress and turmoil in the markets. Rule 2a-7, at the time of the financial crisis, provided that money market funds would be limited to investing in securities that were rated by an NRSRO in one of its two highest short-term rating categories (unless the board of directors determined that an unrated security was of comparable quality). Rule 2a-7 also required money market funds to continuously monitor the ratings of the securities in their portfolios and to respond appropriately in case of a downgrade.30
Over time, the reliance by regulatory agencies on the SEC’s NRSRO designation metastasised into the thousands, even defying scholars’ efforts to quantify all of the regulations at the federal, state and local levels that relied on the NRSRO designation. The invention of the NRSRO designation was very good for credit rating agencies. This regulation-driven demand for ratings motivated Thomas Friedman’s witticism that there are only two superpowers in the world – the US and Moody’s – and that sometimes it was not clear which was more powerful.

Astonishingly, as Partnoy observed, when Thomas Friedman made his famous quip, the credit rating agencies had not even begun their meteoric rise. Moody’s, the only publicly traded NRSRO, had operating revenues, profitability, capitalisation and market share consistent with that of a participant in a government-protected cartel. The value of Moody’s stock increased more than 300% in the five-year period prior to the 2007 market crash as the demand for rating agencies’ services blossomed as more and more exotic credit derivatives were issued (and rated).

Also, significantly, around this period, the largest credit rating agencies – Moody’s, Fitch and Standard and Poor’s (S&P) – began to charge the issuers for ratings. Previously, these credit rating agencies generated revenue by selling subscriptions to publications that contained, among other material, the ratings they generated. The credit rating agencies designated as NRSROs ‘have benefited from an oligopoly market structure that is reinforced by regulations that depend exclusively on credit ratings issued by NRSROs. These regulatory benefits... generate economic rents for NRSROs that persist even when they perform poorly’.

And perform poorly they do. Empirical studies indicate that credit ratings contain little, if any, timely or accurate information about issuers. While many have observed the poor performance of the credit rating agencies and lamented the distortions caused by the NRSRO designation, none have suggested, as we do, that there is a causal link between the NRSRO designation and the rating agencies’ poor performance. Professor Partnoy, for example, takes the view that credit rating agencies never performed well. But if credit rating agencies always performed poorly, it is unclear why they were, at least at one time, of value to investors and issuers. Rather it appears that credit rating agencies played a very modest role in corporate finance until the NRSRO designation uncoupled their profits from their performance.

Historically, ‘the only reason that rating agencies [were] able to charge fees at all [was] because the public ha[d] enough confidence in the integrity of these ratings to find them of value in evaluating the riskiness of investments’. Before companies were required to obtain ratings for their newly issued debt (so that their customers would be permitted by regulators to buy such debt), the only rationale for paying for a credit rating was that the cost of obtaining the rating was lower than the benefit, which came in the form of lower borrowing costs for debt that was subjected to the scrutiny of the credit rating agencies.

Nevertheless, it is not at all clear from the historical evidence that credit rating agencies ever were particularly good at generating accurate ratings. As Martin Fridson has observed, the historical evidence shows correlation between some massive mistakes (like Enron) and a tendency toward ratings inflation, but there also has been a correlation between ratings, defaults and losses, and net returns, suggesting that ratings did historically generate some information of value to investors. From this perspective, it appears that a major part of the problem that government regulation created in the credit rating context was that the
NRSRO designation caused credit ratings to be taken too seriously. Credit ratings, which used to be a mere sideshow in American corporate finance, became the main attraction in the capital markets’ biggest tents.

Moreover, to a large extent, credit ratings are a product of market inefficiency. Credit ratings are necessary in order to compensate for a lack of market-generated information. Over time, as information technology improves and competition among market participants increases, information asymmetry problems lessened, and one would expect that the natural evolution of the capital markets would be toward less reliance on credit ratings. Instead, of course, because of the NRSRO designation, credit ratings have become more, rather than less, important.

The phenomenon of rising demand for credit ratings over time does not appear to be the result of either improved credit ratings or deteriorating capital markets. In fact, the data suggests the opposite. Empirical studies have documented that yield spreads of corporate bonds start to expand as credit quality deteriorates but before a rating downgrade. These results cast doubt on the informational value of credit ratings because they indicate that prices in the capital markets generally anticipate future downgrades by the credit rating agencies. These data also suggest that differences in yields among securities (credit spreads) – the varying rates of return that investors can expect when buying securities with higher or lower yields – reflects the increases (or decreases) in risk associated with these various investments.

Once credit ratings were co-opted by the government’s NRSRO designations, not only did their businesses explode but their basic business model changed as well. Quality became less important because the NRSRO regulatory framework decoupled the quality of the ratings generated by the credit rating agencies from the demand for their services. Thus, the rational response from the credit rating agencies was to lower costs. As Partnoy has suggested, the growth in size and profitability in credit ratings likely is attributable to cost savings: ‘Both S&P and Moody’s have high levels of staff turnover, modest salary levels and limited upward mobility; moreover, investment banks poach the best rating agency employees. These factors limit the ability of rating agencies to generate valuable information. In addition, the process agencies use today to generate ratings does not obtain any obvious advantages over competing information providers and analysts. Credit rating agencies do not independently verify information supplied to them by issuers, and all rating agencies get the same data. Both Moody’s and S&P make rating determinations in secret. The agencies never describe their terms or analysis precisely or say, for example, that a particular rating has a particular probability of default, and they stress that the ratings are qualitative and judgmental. This secretive, qualitative process is not the type of process one would expect if the agencies had survived based on their ability to accumulate reputational capital. On the other hand, such processes make it more likely that an agency would be able to survive in a non-competitive market; if the rating process had been public or quantitative (rather than qualitative), other market entrants easily could have duplicated the rating agencies’ technology and methodology.’

The consequences of the misguided decision to incorporate credit ratings into securities and capital markets regulations were severe. The evidence shows that, whatever the quality and reliability of credit ratings might have been prior to the end of the 20th century, the rating agencies failed dismally between 2001 and 2008. As Thomas Gorman and many others have observed, ‘[c]redit rating agencies, and in particular, NRSRO, have been thought by
many to be at the centre of much of what went on with the market crisis, particularly in the area of structured products. The agencies have come under significant criticism for their methodologies, lack of procedures and conflicts of interest.46

The evidence strongly suggests that credit rating agencies lowered their standards between 2001 and 2008, especially with respect to their ratings of structured financial instruments.47 In particular, as Fridson points out, many credit default obligations received arguably ‘unde-
served Triple A ratings’48 during this period, making it simply ‘impossible to defend the agencies’ … ratings of mortgage-related collateralised debt obligations’49 because fully ‘89% of the investment grade mortgage-backed securities ratings that Moody’s awarded in 2007 were subsequently reduced to speculative grade’.50

Rating structured financial obligations such as collateralised debt obligations (CDOs) is particularly difficult for credit rating agencies it seems: ‘[W]hen a rating agency rates a mortgage-related CDO, it may have greater difficulty controlling the conflict that arises from the issuer-pay model. To begin with, the issuer is not an existing company with a new need for capital. Rather, the prospective offering has come about because an underwriter has structured a financing around a pool of mortgages. The deal is contingent on selling the senior tranche to investors who will accept a comparatively low yield in exchange for a very high level of perceived safety. Therefore, if the bankers are not fairly confident of being able to obtain a Triple-A rating on the senior tranche, they will not even bother to commence work on the deal. In that case, the CDO will not be created and the rating agencies will receive no revenue.’51

The NRSRO regulatory framework ultimately created a ‘competitive’ environment in which the oligopolistic credit rating agencies issuing NRSRO ratings would inevitably come to view the most rational business model to be that of supplying ratings not for the purpose of conveying information but for the purpose of providing prophylactic protection against various risks, including litigation risk and the risk of underperforming against one’s rivals.52

In response to the simultaneous increased reliance on and deteriorating quality of ratings, regulators have recently begun to try to fix the problem they caused in inventing the NRSRO designation. The Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank) created a new office of credit ratings that is responsible for administering the processes by which NRSROs calculate credit ratings. This office is tasked with conducting an annual audit of each NRSRO and issuing a public report on the NRSRO’s performance.53

In order to improve the flawed credit ratings, the SEC, which, in our view actually caused the poor quality of the ratings by inventing the NRSRO concept in the first place, is now responsible for the corporate governance of the credit rating agencies. Each NRSRO is required by Dodd-Frank to have a board of directors that is responsible for overseeing a system of internal controls over the policies and procedures used to determine ratings and other internal issues, such as promotion and compensation.54 But the systemic risk problem remains.

Dodd-Frank requires the SEC to regulate credit rating agencies in a number of additional ways. New rules must be promulgated to: (i) preclude ratings from being influenced by sales and marketing;55 (ii) define the meaning of rating symbols and require that they be used consistently;56 (iii) require each NRSRO to establish, maintain and enforce policies and procedures that ‘clearly define and disclose the meaning of any symbol used by the [NRSRO] to denote a credit rating’ and apply any such symbol ‘in a manner that is consistent for all
types of securities and money market instruments for which the symbol is used;\(^{57}\) (iv) require that each NRSRO assess and disclose the probability that an issuer will default or otherwise not make payments in accord with the terms of the instrument;\(^{58}\) (v) establish the criteria for the qualifications, knowledge, experience and training of persons who perform ratings;\(^{59}\) (vi) require the disclosure of information that will permit the accuracy of ratings and foster comparability among the agencies to be evaluated;\(^{60}\) and (vii) require the disclosure, on a form which will accompany each rating issued, of information about the underlying assumptions, procedures and methodologies employed as well as the data used in establishing the rating.\(^{61}\) These provisions are all based on the idea that SEC regulation can improve the quality of the ratings generated by the NRSROs.

Dodd-Frank also makes it easier for investors to sue credit rating agencies.\(^{62}\) This provision, of course, is based on the notion that the threat of liability can improve quality of the ratings generated by the NRSROs.\(^{63}\)

The sequence of events culminating in the provisions of Dodd-Frank related to credit rating agencies is strongly supportive of the hypothesis of this article, which is that government co-option of private institutions tends to cause the deterioration of such institutions. First, the government co-opted credit rating agencies by relying on NRSRO’s credit ratings to regulate financial institutions. This regulation ultimately distorted credit rating agencies’ incentives and removed their prior market-driven incentives to produce high quality ratings. But the regulation also created more dependence on credit ratings than ever before. This in turn created an acute need for more regulation (which came in the form of Dodd-Frank) aimed at improving the poor performance of the credit rating agencies that the prior regulations caused in the first place.

Congress’s recognition in Dodd-Frank of the failures resulting from regulators’ co-option of credit rating agencies went even further than these new regulations of credit rating agencies. Section 939A of Dodd-Frank requires the SEC to review its myriad regulations referencing credit ratings and to modify those regulations so as to ‘remove any reference to or requirement of reliance on credit ratings’.\(^{64}\) In place of these credit ratings, the SEC must promulgate ‘standard[s] of credit-worthiness as each respective agency shall determine as appropriate for such regulations’.\(^{65}\)

The implications of this rule change are manifold. In particular, the investment advisers who manage money market funds, at least in theory, should no longer rely automatically on the credit ratings generated by NRSROs when making investment decisions; money market fund managers will have to analyse their investments on their own.\(^{66}\) Removing reliance on the credit rating agencies will reduce systemic risk by decreasing the tendency of mutual funds to have investments that are highly correlated with (that is, the same as) the investments of other mutual funds.\(^{57}\) Prior to Dodd-Frank, mutual funds advisers and other money managers were tempted to pick the highest-yielding assets within any particular ratings category in order to maximise the risk-adjusted returns associated with their investments.\(^{68}\) Requiring money managers to make decisions on the basis of their own risk analysis rather than relying on credit rating agencies will, therefore, at least in part, reduce the systemic risk caused by the ‘herd behaviour’ of money managers.\(^{69}\)

To the extent that regulators must now remove the regulatory co-option of the credit rating designation, credit rating agencies will have to once again compete on the basis of
the quality of their ratings as the artificial demand for their services created by regulation subsides. Ironically, this means that the quality of the ratings generated by the agencies should improve just as professional money managers become less reliant on their use. In retrospect, the creation and expansion of the NRSRO designation can be viewed as an experiment with the regulatory co-option of privately generated credit ratings. Clearly this experiment failed; the very act of incorporating credit ratings into the regulatory framework changed the incentives of the companies issuing such ratings and ultimately corrupted the quality of the ratings themselves, and led to significant, non-diversifiable macroeconomic risk.

**Flawed reliance on VaR models**

No private sector innovation has captured the regulatory imagination more than the use of internal risk assessments by financial institutions. In the late 1980s, following the stock market crash of 1987, financial institutions – particularly the largest and most sophisticated and complex banking and financial services companies – began to utilise state-of-the-art financial mathematics to measure and control their companies’ financial risk. The methodology for doing this involves the utilisation of a VaR model or similar algorithm, which measures the maximum loss that a financial company might experience under various sets of assumptions.

VaR models generate risk measures on the specific portfolios of financial assets that financial institutions hold or may be considering. Starting with a particular portfolio, and using various measures of probability and time horizons, a portfolio’s VaR is the measure of the probability that the actual (marked-to-market rather than accounting) loss on the measured portfolio during the given time horizon will exceed a particular threshold value. Thus, for example, assigning to a particular portfolio of stocks a one-day 98.5% VaR of $100 million means that the financial institution assigning this VaR believes that there is a 1.5% chance that the portfolio will fall in value by more than $100 million over a one day period.

Following the financial crisis of 1987, VaR models became widely used in measuring risk, in implementing financial controls, and in determining appropriate levels of capital. Stress-testing is used as part of VaR calculations to measure the stability of a particular institution’s balance sheet. Stress-testing involves measuring the ability of a financial institution to withstand pressures in excess of normal operational levels in order to ascertain the point at which the entity will fail. Variations on the VaR modelling process such as ‘tail value at risk’ (TVaR, also known as tail conditional expectation) permit companies to measure the severity of possible failures as well as the probability that failure will occur.

As long as individual companies were using their own proprietary risk models that differed significantly from one to another, these models did not generate systemic risk. Unfortunately, however, regulators quickly incorporated VaR models into various regulatory frameworks for financial institutions. Beginning in 1997, the SEC pushed investment banks to provide VaR information in the notes to their financial statements in its ruling requiring US public corporations to ‘disclose quantitative information about the risk’ of their derivatives activity. Major banks and dealers implemented the rule by presenting approved VaR models in their financial statements. This created predictable systemic risk as VaR methodologies necessarily became standardised and homogeneous.
Even more significantly, the Basel II capital guidelines, which would, if adopted, dictate the capital requirements for virtually every bank in the world, have adopted VaR throughout the regulations. VaR is the preferred measure of market risk. Significantly, the Basel capital guidelines permit VaR and other so-called ‘internal ratings-based’ approaches to evaluate credit risk, thus permitting banks to rely on their own measures of a counterparty’s risk. These guidelines also envision that banks can effectively determine their own regulatory capital requirements first by creating, and then adjusting, their own internal models. Under the version of the Basel II capital guidelines adopted in the United States by various federal regulators, including the Office of the Comptroller of the Currency (OCC), the Board of Governors of the Federal Reserve System (Board), the Federal Deposit Insurance Corporation (FDIC) and the (now abolished) Office of Thrift Supervision (OTS), a risk-based capital adequacy framework requires certain banks, bank holding companies and savings and loan associations (S&Ls) to ‘use an internal ratings-based approach to calculate regulatory credit risk capital requirements and advanced measurement approaches to calculate regulatory operational risk capital requirements’.

Interestingly, regulators fully understand that the use of financial institutions’ own internal risk assessment tools will occur in an environment in which banks’ models are constantly evolving. Unfortunately, the regulators do not seem to have grasped the notion, central to the analysis in this article, that banks’ incentives regarding how to utilise a market process, such as an internal risk assessment tool, changes when that process becomes incorporated into regulation.

Internal risk assessment tools are particularly subject to distortion by regulatory co-option. Absent any regulatory overlay, internal risk assessment tools are useful for banks to see what sorts of potential problems they are facing. As soon as internal risk assessment tools become internalised into regulation, the consequences of using these tools change dramatically. Regulators and banks, after all, want to use internal risk assessments in order to calculate the amount of capital that banks must maintain to offset particular assets. Because capital is very expensive for companies relative to debt, once internal risk assessments become incorporated into regulation, the use of such assessments can significantly impact the costs of financial institutions. Banks understand that once internal risk assessments can be used to justify reserve capital requirements, they will have incentives to adjust their models, and to modify the way that their models are used, in order to mitigate the costs of using these models.

Besides the odd fact that regulators do not seem to realise that their regulations might actually have an influence on the development of banks’ internal risk assessment tools, a second striking thing about the incorporation of banks’ internal risk assessment processes into the capital adequacy rules is the amount of trust that the regulators apparently have in the banks.

The regulators understand that ‘a system is only as good as the inputs that go into it’. They also assume that financial institutions permitted to use their own internal risk assessments to measure risk will be able ‘to measure the key statistical drivers of credit risk’. Strangely, though, there does not appear to be any theoretical or statistical basis for assuming that banks can do this, particularly in the absence of independent oversight. Moreover, there is no generally accepted paradigm or protocol for determining which of
the myriad technical approaches to internal risk assessment represents best practices or even acceptable industry standards. Whatever banks come up with themselves appears to be fine with the regulators. As one regulator observed, ‘the Basel Committee clearly recognises that there is more than one way to [measure credit risk].’ As such, the applicable capital rules should ‘provide banks with the flexibility to rely on data derived from experience, or from external sources, as long as the bank can demonstrate the relevance of the external data to its own exposures’.

Investors have a strong incentive in accurately measuring the risks of the banks in which they invest. And properly motivated and incentivised bankers have an incentive to measure risk accurately in order to improve loan performance by avoiding lemons and identifying cherries in loan applications and other investments. On the other hand, the necessity of disclosing such risk measurements to powerful and potentially intrusive regulators distorts this incentive and may provide incentives for banks to understate the levels of risk faced.

The well-publicised multibillion-dollar loss suffered in early 2012 by JPMorgan provides an illuminating example of the potential results of a financial institution deriving its own VaR model that qualifies as sufficient for regulatory purposes. Jamie Dimon, CEO of JPMorgan acknowledged, while announcing JPMorgan’s losses, that the bank had altered its VaR model for its Chief Investment Office so that potential losses calculated by the VaR were approximately halved, from US$129 million to US$67 million. The reasons for this change in the VaR model ‘appear certain to be at the centre of regulatory and shareholder inquiries into the losses’. Yet, JPMorgan may not have disobeyed existing regulations. The SEC requires that financial institutions disclose changes in their risk models. However, the institutions do not need to disclose this information until the end of the fiscal year. JPMorgan’s initial first quarter report on 13 April revealed the US$67 million risk but did not note that the VaR had changed. Dimon attributed the change in models to the natural experimental process: ‘There are constant changes and updates to models – always trying to get them better than they were before... That is an ongoing procedure’. Many have expressed doubt regarding Dimon’s statements, noting that, while banks do alter their VaR models, rarely do the changes produce such different results.

The JPMorgan case has resulted in congressional investigations of the bank’s losses and its regulators. Furthermore, the already tenuous credibility of VaR models has come under further scrutiny. The losses show the downside of co-opting the VaRs for regulatory purposes. Ironically, JPMorgan led the development of VaR models in the early 1990s. The financial crisis, however, showed that many of the VaR models used faulty historical data and assumptions. The requirement to disclose changes to VaR models solely at the end of the fiscal year seemingly created this loophole through which JPMorgan was able to change its VaR model for the CIO, but not in other offices, and accomplish its highly risky actions. The Basel Committee, as noted above, acknowledged multiple ways to measure risk and wished to ‘provide banks with the flexibility to rely on data derived from experience’. By making VaRs part of the financial regulatory apparatus but allowing banks to continue to alter the models and only disclose these changes at set intervals, financial regulators appear to have created the conditions under which JPMorgan seemingly was able to alter its own VaR models to lessen the apparent risk in what was actually an extremely risky series of actions.
In other words, banks are now permitted to design the very examinations that will be used to evaluate their financial condition.

Financial institutions subject to minimum capital requirements were not merely allowed to determine for themselves whether they were in compliance with such requirements; they were also allowed to design for themselves the test that they would use for making this determination. As Partnoy observed, financial institutions used faulty VaR models due to legal rules. Regulators required companies to use and to disclose their VaR measures but, in doing so, they were ‘inadvertently encouraging companies to misstate their own risks’.

Partnoy suggests that there was a problem with VaR models themselves, calling them ‘faulty’ on the grounds that ‘all these models really did was compare historical measures of risk and return’. This observation, while true, is describing the first generation of VaR models developed in the mid to late-1990s. The models, such as the ones used by failed companies including Bankers Trust and Long Term Capital Management, were what regulators ultimately assimilated into their own regulations. Once these models were adopted by regulators, however, they became ossified. If these models had not been incorporated into regulation, the companies using them would have had strong incentives to refine and improve them in order to make them more accurate and useful as internal evaluation mechanisms.

Regulatory assimilation distorted the incentives of the companies who developed VaR models. The process of assimilation subverted the integrity of the very algorithms being co-opted. Unfortunately, like credit ratings, internal risk models, which were once a promising means for banks to improve their internal controls, have become largely discredited. For example, Sheila Bair, the former chairman of the FDIC, observed that ‘in the years preceding the financial crisis, under the advanced approaches of Basel II, regulators allowed the largest banks to use their own internal models to set individual risk-based capital requirements’. Regulators permitted this on the premise that ‘the largest banks did not need as much capital as smaller banks due to their sophisticated internal risk models and diversification’. According to Ms Bair, ‘[t]he crisis demonstrated the fallacy of this thinking’. In other words, regulation of bank capital, although done with the best of intentions, created the systemic risk that it was supposed to.

Concluding observations

The reason why regulatory risk is such an acute problem is that regulatory mistakes are common and such mistakes have systemic consequences, while the consequences of mistakes by individual market participants are significantly more localised. In the most straightforward way, by definition, when financial companies in a particular jurisdiction comply with a certain regulation, then by definition, they are acting in harmony. In other words, often regulation transforms company-specific decisions into industry-wide decisions. In so doing, regulation transforms company-specific risk, which can be diversified away, into systemic risk, which cannot be avoided through diversification. To the extent that the regulatory choice is misguided or erroneous, the consequences will be much more widespread in the economy than when the same bad decision is made by individual companies on an ad hoc basis.

For this reason, regulatory globalisation is a particular problem, because it moves systemic risk from a national scale to a global scale. Regulatory Globalisation is pursued by
entrepreneurial administrative agencies as one of a number of possible responses to exogenous market forces that threaten to make the agency, or its mission, irrelevant. Other responses by regulatory agencies to threatened obsolescence that have previously been identified in the literature include: (i) agency ‘imperialism’ in the form of ‘turf grabbing’; (ii) greater proclivity towards agency ‘capture’ by the agency’s special interest group constituents; and (iii) manufactured or fabricated crises to make it appear that there is something for the agency to do.

In the simplest possible terms, threatened irrelevance provides incentives for bureaucrats within agencies to press for regulatory globalisation. Because regulators would prefer to retain power, and do not want to share their power with their counterparts in other countries, they have to be given an incentive to make the sacrifice in autonomy that is represented by regulatory globalisation. Threatened irrelevance provides such an incentive. The downside is that such centralisation increases the already-significant problem of systemic risk attributable to regulation.

---

17. See id. at pp. 62–63.
Regulatory and enterprise risk

See id. at pp. 64–65; Pinto, supra note 8, at pp. 343–44, 348.

See 17 C.F.R. § 240.15c3-1 (2011); Sinclair, supra note 2, at p. 42.

Net capital requirements on introducing brokers’ proprietary accounts assets held by clearing brokers, NYSE Interp. Memo 98-10, 10 December 1998.

Risk is difficult to measure: when we observe in retrospect the performance of a financial asset, it is simple to measure the returns to investors by looking at capital gains and periodic payments such as dividends or interest payments; for risk, there is no such tangible measure, even after the fact. See Accenture, Navigating the complexities of liquidity risk 2, 2010, available at: www.accenture.com/SiteCollectionDocuments/PDF/Navigating_the_Complexities_of_Liquidity_Risk_Final_Low_Res.pdf; noting that ‘liquidity risk [is] extraordinarily difficult to measure’.

See 17 C.F.R. § 240.15c3-1(c)(2)(vi)(E)–(H), 2011, allowing deductions from the net capital requirements if a security obtains a certain rating from NRSROs; Sinclair, supra note 2, at p. 42; cf. Partnoy, supra note 10, at p. 64, discussing the SEC’s decision to begin relying on the ratings of the NRSROs.

‘[D]ebt is cheap and equity is expensive,’ Case, B, ‘Should banks hold more capital? It worked out great for REITs,’ Seeking Alpha, 6 February 2011.

See 17 C.F.R. § 270.2a-7, 2011.

See Partnoy, F, ‘How and why credit ratings agencies are not like other gatekeepers’, in Fuchita and Litan (eds), Financial Gatekeepers: can they protect investors? 2006, Brookings Institute. Stating that money market mutual funds are ‘supposed to be ultra-safe, basically a substitute for cash or a current account’.


Interview by David Gergen with Thomas L. Friedman, 13 February 1996.


Theories of regulatory risk


As markets become more efficient, ratings become less relevant because in an efficient market the information gleaned from ratings will be contained in securities prices themselves. The role of credit ratings is to help close the information gap between lenders and borrowers by providing independent opinions of creditworthiness. See ‘The role of credit ratings in the financial system’, 17 May 2012, 2.24 pm; www.standardandpoors.com/ratings/articles/en/us/?articleType=HTML&assetID=124533793833; see also Gilson, RJ and Kraakman, RH, ‘The mechanisms of market efficiency’, Virginia Law Review 79, 1984, pp. 609–12, discussing various mechanisms through which financial markets become more efficient.


Gorman, TO, ‘Dodd-Frank impact on credit rating agencies’, LexisNexis Communities, 24 August 2010.

Structured finance is the catchall term for financial transactions that create new, often complex, legal entities (special purpose vehicles) whose sole purpose is to issue debt securities on a stand-alone basis (meaning the entities had no business of their own apart from issuing securities) in which the repayment of principal and interest on the securities created is based on the cash flows generated by assets such as mortgages, credit card receivables and car loans. See Jobst, AA, ‘A primer on structured finance’, Journal of Derivatives & Hedge Funds 13(3), 2007, pp. 199–213. Structured financial instruments include a wide variety of securities issued by specialised entities, primarily: asset-backed securities, mortgage-backed securities, collateralised mortgage obligations, collateralised debt obligations, collateralised bond obligations and collateralised obligations of hedge funds and private equity funds. See id. at pp. 201–202. In technical terms, structured investments typically: (i) combine traditional asset classes with contingent claims, such as risk transfer derivatives or derivative claims on commodities, currencies, or receivables from other reference assets; or (ii) replicate traditional asset classes through synthetization or new financial instruments. Structured finance is invoked by financial and nonfinancial institutions in both banking systems and capital markets if either: (i) established forms of external finance are unavailable (or depleted) for a particular financing need; or (ii) traditional sources of funds are too expensive for issuers to mobilise sufficient funds for what would otherwise be an unattractive investment based on the issuer's desired cost of capital. Structured finance offers issuers enormous flexibility to create securities with distinct risk-return profiles in terms of maturity structure, security design and asset type, providing enhanced return at a customised degree of diversification commensurate to an individual investor's appetite for risk. Hence, structured finance contributes to a more complete capital market by offering any mean-variance trade-off along the efficient frontier of optimal diversification at lower transaction cost. The increasing complexity of the structured finance market and the ever-growing range of products being made available to investors, however, invariably create challenges in terms of efficient management and dissemination of information. Id. at pp. 200–201.

Fridson, supra at p. 60.

Id. at p. 56.

Id.

Id. at p. 58


Regulatory and enterprise risk

See Dodd-Frank Act § 932(a)(2)(B)(3), requiring establishment of internal control structure; id. § 932(a)(8)(t), ‘Each nationally recognised statistical rating organisation shall have a board of directors.’

§ 932(a)(4)(3).
§ 932(a)(8)(q)(1).
§ 938.
§ 936.
§ 932(a)(8)(q)(1).
§ 932(a)(8)(s)(1)–(3). The Act mandates that the SEC require NRSROs to ‘prescribe a form to accompany the publication of each credit rating that discloses’ assumptions used, data, and use of servicer or remittance reports, as well as ‘information that can be used by investors and other users of credit ratings to better understand credit ratings in each class of credit rating issued by the nationally recognised statistical rating organisation’. § 932(a)(8)(s)(1).

See Gorman, supra note 80.
Dodd-Frank Act § 931(3).
§ 939A(b).

See Zimmerman, CJ, ‘Over- or under-rated? The SEC proposes credit rating amendments impacting money market funds’, Investment Lawyer 18(5), 2011. Discussing the impact of the new rules on money market funds, observing that the purpose of these rules is to diminish reliance on credit ratings, and predicting that, while companies will continue to rely on ratings in the short run, this reliance may diminish over time.

See Bikhchandani, S and Sharma, S, ‘Herd behaviours in financial markets’, IMF Staff Papers 47(3), 2000, pp. 279–310; ‘Intentional herding may be inefficient and...can lead to excess volatility and systemic risk’, (footnote omitted).

Lynch, TE, ‘Deeply and persistently conflicted: credit rating agencies in the current regulatory environment’, Case Western Reserve Law Review 59, 2009, pp. 284–85, explaining that investors ‘herding behaviour’ was in part caused by the supposed authority of credit rating agencies in evaluating risk; see Dallas, LL, ‘Short-termism, the financial crisis and corporate governance’, Journal of Corporate Law 37(2), 2012, pp. 314–264 (exploring the various incentives for asset managers to engage in herd behaviour); cf. Partnoy, supra note 14, at p. 132, noting that in the 1990s ‘fund managers all seemed to be buying the same financial instruments’.


See id.

See id. at pp. 24–25.


Id.


See id. at 69,294; see also supra notes 3–4 and accompanying text.

Theories of regulatory risk

Id. at 69,291, ‘Because bank risk management practices are both continually evolving and subject to uncertainty, the framework should be viewed as an effort to improve the risk sensitivity of the risk-based capital requirements for banks, rather than as an effort to produce a statistically precise measurement of risk.’

Partnoy, supra at pp. 262–63.

See Beder, supra, at 12, showing how VaR calculations are ‘extremely dependent on parameters, data, assumptions and methodology’ – that is, different methods of calculating VaRs can yield significantly different results – and observing that VaR calculations ‘are not sufficient to control risk’.

See supra note 55.

Rutledge, supra.

Id.

See Wolcott, R, ‘Beyond the numbers: do banks manage risk?’, Reuters, 14 June 2012, observing that ‘[i]t is a myth that banks are in possession of fancy gadgetry that allows them to measure risk on a minute by minute basis from a specialised risk-control tower and react to it effectively, thus averting catastrophe’; see also Magas, I, ‘Financial innovation: regulator’s nightmare’, International Relations Quarterly 2, 2011, pp. 1, 2; available at www.southeast-europe.org/pdf/07/DKE_07_A_V_Magas-Istvan.pdf. ‘When the bulk of a company’s assets were physical and its markets were relatively stable, valuation was more straightforward. Now growing proportions of a company’s assets – brands, ideas, human capital – are intangible and often hard to identify, let alone value... This new, very innovative partly IT-related, complex market development has increased the difficulties of assessing risk and value, especially in a global context.’

The losses, originally announced on 10 May 2012, stemmed from JPMorgan’s Chief Investment Office (CIO). The media initially reported in April 2012 on a trader in London named Bruno Iksil, known to many as the ‘London Whale’. Ruhle, S, Keoun, B and Childs, M, ‘JPMorgan trader’s positions said to distort credit indexes’, Bloomberg, 6 April 2012, discussing Bruno Iksil, the trader involved in the credit-derivative indexes. Iksil, who specialised in credit-derivative indexes, and the CIO built up an extremely complicated, high-stakes bet over a period of years. According to the New York Times, ‘[i]n its simplest form... the complex position assembled by the bank included a bullish bet on an index of investment-grade corporate debt, later paired with a bearish bet in high-yield securities, achieved by selling insurance contracts known as credit-default swaps’. Schwartz, ND and Silver-Greenberg, J, ‘JPMorgan’s trading loss is said to rise at least 50%’, DealB%k, 16 May 2012, a ‘big move’ in the interest rate spread between risk free government bonds and investment grade securities, paired with a lack of ‘equally large moves in the price of the insurance on the high yield bonds’, caused the bet to lead to huge losses. Id. Total losses could eventually total US$9 billion. See Silver-Greenberg, J and Craig, S, ‘JPMorgan trading loss may reach $9 billion’, DealB%k, 28 June 2012, see also Farrell, M, ‘JPMorgan’s losses keep adding up’, CmfnMoney, 12 June 2012; Silver Greenberg, J and Eavis, P, ‘JPMorgan discloses $2 Billion in trading losses’, DealB%k, 10 May 2012, discussing JPMorgan’s failed hedging strategy and its potential effect on regulatory reforms.

Id.

Id.

‘Banks sometimes refine their value-at-risk, or VaR, models but those commonplace changes do not by themselves produce such dramatically different results’, said Christopher Finger, one of the founders of RiskMetrics Group, which pioneered VaR models and is now a unit of MSCI Inc.

See Patterson, S, ‘Comptroller sees risk-management breakdown at JPMorgan’, Wall Street Journal, 5 June 2012, ‘The OCC has been taking heat from members of Congress for failing to pick up on JPMorgan’s trading activities quickly enough’; Protess, B, ‘Regulator concedes oversight lapse in JPMorgan loss’, DealB%k, 6 June 2012, discussing the Senate Banking Committee hearing regarding JPMorgan’s loss, where the Comptroller of
the Currency ‘faced the brunt of the scrutiny’, and noting previous hearings where officials from JPMorgan’s other regulators, the CFTC and SEC, testified.

99 Id; see Henry, supra.
100 The VaR was only changed for the CIO; the other divisions of the bank continued to use more ‘conservative’ models. Henry, supra.
101 Rutledge, supra.
103 Rutledge, supra.
104 Id. at pp. 261–62
106 Id.
107 Id.
Chapter 9

Key principles of enterprise risk management for insurance groups

David Ingram
Willis Re

Overview

Insurance companies, like other businesses, have a need for sales, profits, security and value growth. The way that an insurance company approaches risk management will depend on where they stand with respect to those corporate needs. The approach to risk management will also depend upon the risk environment in which their business operates. But for all insurers, their risk management will be built with seven key principles of enterprise risk management (ERM). Those principles are: diversification, underwriting, control cycle, consideration, provisioning, portfolio and future risks. Insurers will each stress different combinations of these principles in their risk management programs. The most recent developments in risk management are: (i) the portfolio principle; (ii) the risk capital part of provisioning; and (iii) the application of the control cycle to aggregate risk tying together (i) and (ii). This chapter will go into detail explaining these seven principles and how insurers apply and misapply them.

Introduction

Even for insurers, ERM needs to be seen in perspective of the other management priorities. Like the human needs articulated by Maslow, there is a hierarchy of corporate needs and ERM is not first on that list.

The first need of a corporation is to have a product or service that someone will buy from them. Simply put, a company must have sales. That this is the first and most basic need of all corporations is often lost from sight, perhaps because the corporation has had sales for many, many years. But any person who founds a new company will very clearly remember when they were first sure that they had found something that their customers would buy. No company can exist for any length of time without sales. In fact, Peter Drucker insists rather strongly that ‘there is only one valid definition of business purpose: to create a customer’.  

The second need is for the business to have the capability to create that product or service at a cost that is lower than the price that customers will pay, a profit. As products and services go through cycles of demand ebbs and flows, companies that had achieved their profit may be challenged to stay profitable. Of course, a company cannot continue in the long run without a fundamentally profitable business. Drucker says that ‘the first test
of any business is not the maximisation of profit but the achievement of sufficient profit to cover the risks of economic activity and thus avoid loss’.

Once a company has sales and profits, that company has the potential to be valuable. But one more thing is needed for that company to actually be valuable – the ability to stay in business in the future and repeat the process of making sales at a profit. These first two needs of companies and the ability to continually meet those needs should be of the highest concern to the executives and boards of companies.

The third corporate need is that repeatability of sales and profits. That is where ERM comes into the hierarchy of need; in third place. When the business has the ability to repeat the process of profitable sales, then the business has a value. ERM is the comprehensive process for defending that value. Once the decisions have been made regarding how to go about making sales and profits, the role of ERM is defensive. Corporate value is a function that combines all of the first three needs – amount of sales, profit per unit of sales and likelihood of achieving the sales and profit. To satisfy the third level corporate need, ERM usually focuses on maintaining or improving the likelihood of achieving a profit.

The defence of sales is often completely separated from the defence of profits. Some risk managers have named that process ‘strategic risk’ and seek to include it within ERM. In many companies, defence of sales is a more longstanding and better staffed function than ERM.

When a company has a value, a fourth need comes into play; that is the need to increase that value (Exhibit 9.1). A company can increase value in three ways: by increasing sales, by improving profits per unit of sales, or by increasing the likelihood of achieving the future sales and profits. In some circumstances, there are options that exist for a company to improve one of these three elements at minimal cost to the other two, but those options are rare. In most cases, companies must make a trade-off. It is in deciding on those trade-offs that ERM can provide unique insights about how decisions will impact upon the likelihood of achieving future profits.

This ability to systematically and quantitatively evaluate the trade-offs ‘is’ an exciting new addition to management’s capabilities that ERM can bring to the table. It is the capability that drives the process that Standard & Poors (S&P) refers to as ‘strategic risk management’. To many practitioners, this capability is ERM.

The term ERM grew to be used by risk managers in financial institutions in the 1990s. ERM, like many developments of financial economics in the post-World War II period, promises very much. It purports to deliver an enterprise-wide approach to risk that can be used to align the risks and rewards for the activities of a financial company with the ultimate objectives of the company.

At banks, ERM grew in popularity, budget and influence throughout the 1990s. As problems were experienced, mostly with derivatives, either the industry or the regulators, or both, together would study the problem and create additions to ERM practices to forestall recurrences. When the Dot Com bubble crashed and the New York World Trade Center buildings were destroyed by terrorists, there were very few banks that suffered losses that they were not able to bear. The story of 2001 was very different for insurers. Many insurers had large exposures to equities, including the internet stocks that lost 90% or more of their value. And over 100 reinsurers and insurers paid a total of about US$32 billion in claims for the World Trade Center disaster.
A handful of insurers had started to develop ERM programs prior to 2001. The contrasting experiences of banks and insurers led many of the largest insurers and reinsurers to start new ERM programs.

Insurers have always been in the risk management business. Unlike banks, whose business models in the last 30 years had developed to be more of risk intermediaries, insurers were historically, and remain, risk aggregators. Risk aggregators need a very different type of risk management. Some insurers readily recognised that and built their ERM systems out from their existing risk management systems. Others created an almost totally separate group level risk function where none had existed.

As ERM gained in popularity, many companies added ERM programs that were a copy of what was popular elsewhere. But that resulted in a churning of interest from management groups. It was quite common for an insurer to fund ERM for several years and then to sharply reduce funding. Ultimately, you could say that this was the result of ERM over-promising and under-delivering.

That happened because ERM was not clearly aligned with the hierarchy of corporate needs. ERM seemed to be out of step with the corporate culture, based upon its own internally defined purpose rather than with the actual perceived needs of the insurer’s management. ERM programs would be built up to support a process to enhance value growth, the fourth level corporate need when management wanted a profit assurance function to support second and third level needs.
In addition, few of the early adopters of ERM may have looked carefully at the existing risk management capabilities of the insurer and designed an ERM program that made sense, and that was in addition to those existing capabilities, but later adopters who copied those pioneers may not have done that homework to find out exactly what was needed.

A full program of risk management and enterprise risk management needs to address seven major capabilities of an insurer.

- **Diversification**: risks must be diversified. There is no risk management if a company is just taking one big bet.
- **Underwriting**: there must be a process for risk acceptance that includes an assessment of risk quality. Companies need to be sure of the quality of the risks that they take. This implies that multiple ways of evaluating risks are needed to maintain quality, or to be aware of changes in quality. There is no single source of information about quality that is adequate.
- **Control cycle**: there must be a control cycle to manage the amount of risk taken. This implies measurements, appetites, limits, treatment actions, reporting and feedback.
- **Consideration**: there must be a process for assuring that the consideration received for accepting risk is adequate. For risks that are not traded, such as operational risks, the benefit of the risk needs to exceed the cost in terms of potential losses.
- **Provisioning**: there must be appropriate provisions held for retained risks, in terms of set asides (reserves) for expected losses and capital for excess losses.
- **Portfolio**: there must be an awareness of the interdependencies within the portfolio of risks that are retained by the insurer. This would include awareness of both risk concentrations and diversification effects. An insurer can use this information to take advantage of the opportunities that are often associated with its risks through a risk-reward management process.
- **Future risks**: there must be a process for identifying and preparing for potential future emerging risks. This would include identification of risks that are not included in the processes above, assessment of the potential losses, development of leading indicators of emergence and contingent preparation of mitigation actions.

**Enterprise risk cycle**

In practice, these seven principles do not exist as a list that can be completed, but as a cycle that is repeated indefinitely (Exhibit 9.2).

ERM touches and is touched by the actions that are triggered by each and every one of these principles. The execution of some or all of the seven ERM principles may be charged to the ERM function, or the ERM function has an oversight role for some. However, since most of these functions existed before ERM, there is confusion and conflict in some companies where some of the connections between ERM and these key functions do not exist.

A full description of each of the key principles, how they work, how they can go wrong, and what it takes for them to go right, follows. After those descriptions, we will return to the ‘enterprise risk control cycle’ and discuss how insurance companies go about co-ordinating their implementation of these principles.
Shirreff advises: ‘The first rule of risk management is to identify your risk; the second is to diversify it.’ There are records showing that the power of diversification of risks was known to the ancients. Investors who financed trading ships clearly favoured taking fractions of a number of ships to owning all of a single ship.

The benefits of diversification are clear. The maths are highly compelling. A portfolio of \( n \) risks of the same size \( A \) that are truly independent have a volatility that is a fraction of the volatility of totally dependent risks.

Here is a simple example. There is a 1 in 200 chance that a house will be totally destroyed by fire. Company A writes an insurance policy on one $500,000 house that would pay for replacement in the event of a total loss. That means that company A has a 1 in 200 chance of paying a $500,000 claim. Company B decides to write insurance that pays a maximum of $50,000 in the event of a total loss. How many policies do you think that Company B would need to write to be as safe as Company A?
needs to write to have a 1 in 200 chance of paying $500,000 of claims if the risks are all totally independent and exactly as prone to claims as the $500,000 house? The answer is an amazing 900 policies or 90 times as much insurance!

When an insurer is able to write insurance on independent risks, then with each additional risk, the relative volatility of the book of insurance decreases. Optimal diversification occurs when the independent risks are all of the same size. For insurers, the market is competitive enough that the company writing the 900 policies is not able to get a profit margin that is proportionate to the individual risks. The laws of micro economics work in insurance to drive the profit margins down to a level that is at or below the level that makes sense for the actual risk retained. This provides the most compelling argument for the price for insurance for consumers, they are getting most of the benefit of diversification through the competitive mechanism described above. Because of this, things are even worse for the first insurer with the one policy. To the extent that there is a competitive market for insurance for that one $500,000 house, that insurer will only be able to get a profit margin that is commensurate with the risk of a diversified portfolio of risks.

Real insurance situations are much more complex than this example. ‘Diversification benefits vary substantially from company to company, and depend on an enormous variety of factors, including business mix, geographical mix, corporate structure (for example, branch versus subsidiaries) and capital management philosophy and use of risk transfer instruments, to name but a few.’

It is curious to note than in many situations, both insurers and individuals do not diversify. This may be explained by imagining that they either forget about diversification when making individual risk decisions, which suggests some degree of irrationality, or that they are acting rationally and believe that the returns for the concentrated risk that they undertake are sufficiently large to justify the added risk.

Exhibit 9.3 shows the degree to which individuals in various large companies are acting against the principle of diversification in their 401K retirement investment decisions. From a diversification point of view, the Procter & Gamble employees are most similar to the insurer above that writes the one $500,000 policy. They may believe that Procter & Gamble is less risky than a diversified portfolio of stocks. Unlike the insurer, where the constraint on the amount of business that they can write is the 1/200 loss potential, the investor in this case is constrained by the amount of funds to be invested. So if a $500,000 401K account with Procter & Gamble stock has a likelihood of losing 100% of value of 1/200, then a portfolio of 20 $25,000 positions in similarly risky companies would have a likelihood of losing 15% of value of 1/1000. Larger losses would have much lower likelihood.
Exhibit 9.3

401K allocations to stocks of the company in which the employees work

<table>
<thead>
<tr>
<th>Company</th>
<th>Percentage of pension in company stock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbott Laboratories</td>
<td>82</td>
</tr>
<tr>
<td>Anheuser-Busch</td>
<td>83</td>
</tr>
<tr>
<td>Bank of America</td>
<td>43</td>
</tr>
<tr>
<td>Ford Motor Company</td>
<td>50</td>
</tr>
<tr>
<td>Marsh &amp; McLennan</td>
<td>61</td>
</tr>
<tr>
<td>McDonalds</td>
<td>74</td>
</tr>
<tr>
<td>Pfizer</td>
<td>82</td>
</tr>
<tr>
<td>Procter &amp; Gamble</td>
<td>92</td>
</tr>
<tr>
<td>SBC</td>
<td>38</td>
</tr>
<tr>
<td>Verizon</td>
<td>51</td>
</tr>
</tbody>
</table>

*Source: Mitchell, O and Utkus, S, ‘Company stock and retirement plan diversification’, Pension Research Council, 2002*

With that kind of maths in its favour, it is hard to imagine that the holdings in employer stock in the 401Ks represents a rational estimation of higher returns, especially not on a risk-adjusted basis. People must just not be at all aware of how diversification benefits them.

There could be another explanation, in the case of stock investments. It can be most easily framed in terms of the capital asset pricing theory (CAPM) terms. CAPM suggests that stock market returns can be represented by a market or systematic component (beta) and a company specific component (alpha). Most stocks have a significantly positive beta. In work that the author has done replicating mutual fund portfolios with market index portfolios, it is not uncommon for mutual fund returns to be 90% explained by total market returns. People may be of the opinion that since the index represents the fund, that everything is highly correlated to the index and therefore not really independent.

The simplest way to refute that thought is to show the variety of returns that can be found in the returns of the stocks in the major sectors (Exhibit 9.4). The S&P 500 return for 2012 was 16%. Clearly, all sectors do not have returns that are closely related to the index, either in 2012 or for any other period shown in Exhibit 9.4. Both insurance companies and investors can have a large number of different risks but not be as well diversified as they would think. That is because of the statement above, that optimal diversification results when all risks are equal. Investors such as the 401K participants with half or more of their portfolio in one stock may have the other half of their money in a diversified mutual fund. But the large size of the single position is difficult to overcome. The same thing happens to insurers who are tempted to write just one, or a few risks that are much larger than their usual business. The diversification benefit of their large portfolio of smaller risks disappears quickly when they add just a few much larger risks.
Exhibit 9.4

Returns in major sectors

<table>
<thead>
<tr>
<th>Sector</th>
<th>1-year</th>
<th>3-year</th>
<th>5-year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic materials</td>
<td>12.23</td>
<td>3.49</td>
<td>2.85</td>
</tr>
<tr>
<td>Communication services</td>
<td>17.53</td>
<td>11.43</td>
<td>2.77</td>
</tr>
<tr>
<td>Consumer cyclical</td>
<td>32.99</td>
<td>20.21</td>
<td>11.05</td>
</tr>
<tr>
<td>Consumer defensive</td>
<td>25.2</td>
<td>15.78</td>
<td>9.78</td>
</tr>
<tr>
<td>Energy</td>
<td>6.84</td>
<td>7.06</td>
<td>1.6</td>
</tr>
<tr>
<td>Financial services</td>
<td>30.95</td>
<td>5.7</td>
<td>−2.45</td>
</tr>
<tr>
<td>Healthcare</td>
<td>26.8</td>
<td>74.85</td>
<td>6.71</td>
</tr>
<tr>
<td>Industrials</td>
<td>19.4</td>
<td>11.86</td>
<td>5.58</td>
</tr>
<tr>
<td>Real estate</td>
<td>27.82</td>
<td>16.09</td>
<td>5.75</td>
</tr>
<tr>
<td>Technology</td>
<td>18.58</td>
<td>13.19</td>
<td>9.5</td>
</tr>
<tr>
<td>Utilities</td>
<td>6.22</td>
<td>2.34</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Source: Morningstar

Diversification is the universal power tool of risk management. But like any other tool, it must be used properly to be effective.

The bond insurers diversified out of their niche of municipal bonds into real estate-backed securities and suddenly these two markets that previously seemed to have low correlation were highly correlated as the sub-prime crisis brought down the bond insurers and their problems rippled into the municipality market. They seemed uncorrelated, but of course they are highly interdependent since a high fraction of municipal incomes comes from taxes relating to real estate values. That is a major problem with the statistical idea of correlation – statistical approaches must never be used uncritically.

But the point of the bond insurer story above is that interdependencies do not have to come from the fundamentals of two markets – that is to come from common drivers of risk. Interdependencies, especially of market prices, can and often do come from common ownership of securities from different markets. The practice of holding risks from seemingly unrelated risks or markets is generally thought to create better risk-adjusted results because of diversification.

But the perverse truth is that as with many things in real economics (not book economics) the more people that use this rule, the less likely it is that it will work.

There are several reasons for this.

- When a particularly large organisation diversifies, their positions in every market will be large. For anyone to get the most benefit from diversification, they need to have positions in each diversifying risk that are similar in size. Since even the largest companies had to have started somewhere, they will have a primary business that is very large and so
will seek to take very large positions in the diversifying markets to get that diversifying benefit. So there ends up being some very significant specific risk of a sudden change in correlation if that large company runs into trouble. These events only ever happen once to a company, so there is never, ever, any historical correlations to be found. But if you want to avoid this diversification pitfall, it pays to pay attention to where the largest companies operate and be cautious in assuming diversification benefits where ‘they’ are the correlating factor.

• When large numbers of companies use the same correlation factors (think Solvency II), then they will tend to all try to get into the same diversifying lines of business where they can get the best diversification benefits. This results in both the specific risk factor mentioned above and to a pricing pressure on those markets. Those risks with ‘good’ diversification will tend to price down to their marginal cost, which will be net of the diversification benefit. The customers will end up getting the advantage of diversification.

• Diversification is commonly believed to eliminate risk. This is decidedly ‘not true’. No risk is destroyed via diversification. All of the losses that were going to happen do happen, unaffected by diversification. What diversification hopes to accomplish is to make these losses relatively less important and more affordable because some risk taking activity is likely to be showing gains while others are showing losses. So people who thought that because they were diversified, they had less risk, were willing to go out and take more risk. This effect causes more of the stampede for the exits behaviour when times get tough and the losses that were ‘not’ destroyed by diversification occur.

• The theory of a free lunch with diversification encourages companies who are inexperienced with managing a risk to take on that risk because their diversification analysis says that it is ‘free’. These companies will often help to drive down prices for everyone, sometimes to the point that they do not make money from their ‘diversification play’ even in good years. Guess what? All that fancy correlation maths does not work as advertised if the expected earnings from a ‘diversifying risk’ is negative. There is no diversification from a losing operation because it has no gains to offset the losses of other risks.

**Underwriting of risks**

Underwriting is the process of reviewing and selecting risks that an insurer might accept, under what terms, and assigning those an expected cost and level of riskiness.

• Some underwriting processes are driven by statistics. A few insurers who developed a highly statistical approach to underwriting personal auto cover have experienced high degrees of success. With a careful mining of the data from their own claims experience, these insurers have been able to carefully subdivide rating classes into many finer classes with reliable claims expectations at different levels. This allows them to concentrate their business on the better risks in each of the larger classes of their competitors while the competitors end up with a concentration of below average drivers in each larger class. This statistical underwriting process is becoming a required tool to survive in personal auto and is being copied in other insurance lines.
Many underwriting processes are highly reliant on judgment of an experienced underwriter, especially commercial business or other types of cover where there is very little close commonality between one case and another. Many insurers consider underwriting expertise to be their key corporate competency.

Usually the underwriting process concludes with a decision on whether to make an offer to accept a risk under certain terms and at a determined price.

**How underwriting can go wrong.**

- Insurers are often asked to ‘give away the pen’ and allow third parties to underwrite risks on their paper. Sometimes there is a very sad ending to this.
- Statistical underwriting can spin out of control due to anti-selection if not overseen by experienced people. The bubble of US home mortgage securities can be seen as an extreme example of statistical underwriting gone bad. Statistics from prior periods suggested that sub-prime mortgages would default at a certain low rate. Over time, the US mortgage market went from one with a high degree of underwriting of applicants by skilled and experienced reviewers to a process dictated by scores on credit reports and eventually the collection of data to perform underwriting stopped entirely with the ‘no doc’ loans. The theory was that the interest rate charged for the mortgages could be adjusted upwards to the point where the extra interest collected could pay for the excess default claims from low credit borrowers.
- Volume incentives can work against the primary goals of underwriting.
- Insurance can be easily undone by underwriting decisions that are good risks, but much too large for the pool of other risks held by the insurer.

To get underwriting right you need to have a clear idea of the risks that you are willing to accept, your risk preferences. And be clear that you are going to be saying no to risks that are outside of those preferences. You should not let the pen get entirely out of the hand of an experienced underwriter that is trusted to make decisions in the interest of the company, either to a computer or to a third party. Oversight of underwriting decisions needs to be an expectation at all levels. The primary objective of this oversight should be to continually perfect the underwriting process and knowledge base. Finally, underwriters need to be fully aware of the results of their prior decisions by regular communication with claims and reserving people.

**Controlling risks with a cycle**

This is a ‘risk control cycle’ (Exhibit 9.5). It includes thinking/observing steps and action steps. The only reason a sane organisation would spend the time on the ‘assessing’, ‘planning’ and ‘monitoring’ steps is so that they could be more effective with the ‘risk taking’, ‘mitigating’ and ‘responding’ steps.
A process capable of limiting losses can be referred to as a complete risk control process, which would usually include the following.

- **Identification of risks:** with a process that seeks to find all risks inherent in an insurance product, investment instrument, or other situation, rather than simply automatically targeting ‘the usual suspects’.
- **Assess risks:** this is both the beginning and the end of the cycle. At the end, this step is looking back and determining whether your judgment about the risk and your ability to select and manage risks is as good as you thought that it would be. At the beginning, you look forward to form a new opinion about the prospects for risk and rewards for the next year. For newly identified risks/opportunities this is the due diligence phase.
- **Plan risk taking and risk management:** based upon the risk assessment, management will make plans for how much of each risk the organisation will plan to accept and then how much of that risk will be transferred, offset and retained. These plans will also include the determination of limits.
• **Take risks:** organisations will often have two teams of individuals involved in risk taking. One set will identify potential opportunities based upon broad guidelines that are either carried over from a prior year or modified by the accepted risk plan (sales). The other set will do a more detailed review of the acceptability of the risk and often the appropriate price for accepting the risk (underwriting).

• **Measuring and monitoring of risk:** with metrics that are adapted to the complexity and the characteristics of the risk as well as ‘regular reporting of positions’ versus limits/checkpoints, where the timing needs to be effective depends on the volatility of the risk and the rate at which the insurer changes their risk positions. Insurers may report at a granular level that supports all specific decision making and actions on a regular schedule.

• **Regular risk assessment and dissemination of risk positions and loss experience:** with a standard set of risk and loss metrics and distribution of risk position reports, and with clear attention from persons with significant standing and authority in the organisation.

• **Risk limits and standards:** directly linked to objectives. Terminology varies widely, but many insurers have both hard ‘limits’ that they seek to never exceed and softer ‘checkpoints’ that are sometimes exceeded. Limits will often be extended to individuals within the organisation with escalating authority for individuals higher in the organisational hierarchy.

• **Response:** enforcement of limits and policing of checkpoints, with documented consequences for limit breaches and standard resolution processes for exceeding checkpoints. ‘Risk management processes’ such as risk avoidance for risks where the insurer has zero tolerance. These processes will ensure that constant management attention is not needed to assure compliance. However, occasional assessment of compliance is often practiced. ‘Loss control processes’ to reduce the avoidable excess frequency and severity of claims and to assure that when losses occur, the extent of the losses is contained to the extent possible. ‘Risk transfer processes’, which are used when an insurer takes more risk than they wish to retain and where there is a third party who can take the risk at a price that is sensible after accounting for any counterparty risk that is created by the risk transfer process. ‘Risk offset processes’, which are used when insurer risks can be offset by taking additional risks that are found to have opposite characteristics. These processes usually entail the potential for basis risk because the offset is not exact at any time or because the degree of offset varies as time passes and conditions change, which is overcome in whole or in part by frequent adjustment to the offsetting positions. ‘Risk diversification’, which can be used when risks can be pooled with other risks with relatively low correlation. ‘Risk costing/pricing’, which involves maintaining the capability to develop appropriate views of the cost of holding a risk in terms of expected losses and provision for risk. This view will influence the risks that an insurer will take and the provisioning for losses from risks that the insurer has taken (reserves). This applies to all risks but especially to insurance risk management. Co-ordination of insurance profit/loss analysis with pricing with loss control (claims) with underwriting (risk selection), risk costing and reserving, so that all parties within the insurer are aware of the relationship between emerging experience of the risks that the insurer has chosen to retain and the expectations that the insurer held when it chose to write and retain the risks.
Key principles of enterprise risk management for insurance groups

• *Assess risks:* and the cycle starts again.

In practice there are a few common problems with the execution of the control cycle that may lessen or even completely negate the efficacy of this process.

Many underwriting processes are set up with a hierarchy of authority limits for people up the chain of command in the underwriting hierarchy. The larger the case, the more senior the person required to approve. A control cycle may be set up with the same sort of hierarchy for approving limit breaches. However, an expectation that limit breaches will be approved makes it more and more likely that the people whose activities are targeted for controlling the cycle will decide that when a particularly good opportunity comes along that would cause a limit breach, that they might as well go ahead and add the risk with the expectation that it will be approved. Frequently allowing limit breaches undermines the whole purpose of the control cycle. However, some organisations have an objective to fully utilise their limits. That objective will necessarily cause great tension within the control cycle process.

In other organisations, the control cycle is not particularly effective because the limits are more for show than for use. In these organisations, limits are commonly set at levels that are twice the level of planned activity. The rationale for limits that are so far away from activity levels is that they provide flexibility for management discretion. Also, in companies with these generous limits, there is often not a clear sense of any overall risk tolerance. These broad limits are usually not set in terms of a risk-adjusted value. For example, many investment policies of insurers will have these very broad limits. The limits are stated in terms of the investment values without adjustment for risk. If the investment managers shift from lower risk to higher risk investments, this type of system will usually only deal with the difference in risk levels with lower limits for the higher risk assets, but often without any recognition of a wide range of riskiness within a class of riskier assets. The result is that investment managers are free to make major changes to the level of risk of the investment portfolio without oversight that is triggered by the limit system.

Other companies may have board approved risk limits, but management never takes those limits and creates the control cycle that would enforce the limits. Few employees who are involved in risk taking are aware of the limits. Reports that might inform a control cycle monitoring process do not show the risk limits. Actual compliance with the limits is just not known at any point in time.

The control cycle, and especially the risk appetite, tolerance, limit setting process can be the basis for a healthy discussion between management and the board. But in many insurance companies, management does not explain to the board the potential ramifications of higher or lower limits.

Ultimately, to get the most risk management benefit out of a risk control cycle, management must set limits at a level that matters and are tied to good measures of risk, make sure that everyone knows what the limits are, and frequently and publicly review the status of compliance with limits. But in addition to a policing function, the control cycle needs to include a learning element. With each pass through the cycle, management should gain some insight into the characteristics of their potential risks, their risk mitigation alternatives, as well as the reactions of both to changes in the risk environment.
Consideration for risks

Consideration for accepting a risk needs to be at a level that will sustain the business and produce a return that is satisfactory to investors.

Investors usually want additional return for extra risk. This is one of the most misunderstood ideas in investing. Baker et al. state: ‘In an efficient market, investors realise above-average returns only by taking above-average risks. Risky stocks have high returns, on average, and safe stocks do not.’ But their study found that stocks in the top quintile of trailing volatility had real return of –90% versus a real return of 1,000% for the stocks in the bottom quintile. But the thinking is wrong. Excess risk does not produce excess return. The cause and effect are wrong in the conventional wisdom. The original statement of this principle may have been: ‘In all undertakings in which there are risks of great losses, there must also be hopes of great gains.”

Marshal has it right. There are only ‘hopes’ of great gains. There is no invisible hand that forces higher risks to return higher gains. Some of the higher risk investment choices are simply bad choices.

Insurers’ opportunities to make ‘great gains’ out of ‘risks of great losses’ is when they are determining what consideration, or price, that they will require to accept a risk. Most insurers operate in competitive markets that are not completely efficient. Individual insurers do not usually set the price in the market, but there is a range of prices at which insurance is purchased in any time period. Certainly the process that an insurer uses to determine the price that makes a risk acceptable to accept is a primary determinant in the profits of the insurer. If that price contains a sufficient load for the extreme risks that might threaten the existence of the insurer, then over time, the insurer has the ability to hold and maintain sufficient resources to survive some large loss situations.

One common goal conflict that leads to problems with pricing is the conflict between sales and profits. In insurance as in many businesses, it is quite easy to increase sales by lowering prices. In most businesses, it is very difficult to keep up that strategy for very long as the realisation of lower profits or losses from inadequate prices is quickly realised. In insurance, the premiums are paid in advance, sometimes many years in advance of when the insurer must provide the promised insurance benefits. If provisioning is tilted towards the point of view that supports the consideration, the pricing deficiencies will not be apparent for years. So insurance is particularly susceptible to the tension between volume of business and margins for risk and profits, and since sales is a more fundamental need than profits, the margins often suffer.

As just mentioned, insurers simply do not know for certain what the actual cost of providing an insurance benefit will be, certainly not with the degree of certainty that businesses in other sectors can know their cost of goods sold. The appropriateness of pricing will often be validated in the market. Follow-the-leader pricing can lead a herd of insurers over the cliff. The whole sector can get pricing wrong for a time. Until, sometimes years later, the benefits are collected and their true cost is known. According to a Wall Street Journal article of 24 June 2002, ‘A decade of short sighted price slashing led to industry losses of nearly US$3 billion last year.’
Pricing can also go wrong on an individual case level. The ‘Winners Curse’ sends business to the insurer that most under-imagines the riskiness of a particular risk.\(^{12}\)

There are two steps to reflecting risk in pricing. The first step is to capture the expected loss properly. Most of the discussions above relate to this step and the major part of pricing risk comes from the possibility of missing that step. But the second step is to appropriately reflect all aspects of the risk that the actual losses will be different from expected. There are many ways that such deviations can manifest themselves.

The following is a partial listing of the risks that might be examined.\(^{13}\)

- **Type A risk**: short-term volatility of cash flows in one year.
- **Type B risk**: short-term tail risk of cash flows in one year.
- **Type C risk**: uncertainty risk (also known as parameter risk).
- **Type D risk**: inexperience risk relative to full multiple market cycles.
- **Type E risk**: correlation to a top 10.
- **Type F risk**: market value volatility in one year.
- **Type G risk**: execution risk regarding difficulty of controlling operational losses.
- **Type H risk**: long-term volatility of cash flows over five or more years.
- **Type J risk**: long-term tail risk of cash flows over five years or more.
- **Type K risk**: pricing risk (cycle risk).
- **Type L risk**: market liquidity risk.
- **Type M risk**: instability risk regarding the degree that the risk parameters are stable.

There are also many different ways that risk loads are specifically applied to insurance pricing. Three examples are as follows.

- **Capital allocation**: capital is allocated to a product (based upon the provisioning) and the pricing then needs to reflect the cost of holding the capital. The cost of holding capital may be calculated as the difference between the risk free rate (after tax) and the hurdle rate for the insurer. Some companies alternately use the difference between the investment return on the assets backing surplus (after tax) and the hurdle rate. This process assures that the pricing will support achieving the hurdle rate on the capital that the insurer needs to hold for the risks of the business. It does not reflect any margin for the volatility in earnings that the risks assumed might create, nor does it necessarily include any recognition of parameter risk or general uncertainty.

- **Provision for adverse deviation**: each assumption is adjusted to provide for a worse experience than the mean or median loss. The amount of stress may be at a predetermined confidence interval (such as 65%, 80%, or 90%). Higher confidence intervals would be used for assumptions with higher degrees of parameter risk. Similarly, some companies use a multiple (or fraction) of the standard deviation of the loss distribution as the provision. More commonly, the degree of adversity is set based upon historical provisions or upon judgment of the person setting the price. Provision for ‘adverse deviation’ usually does not reflect anything specific for the extra risk of insolvency.
• **Risk-adjusted profit target:** using either or both of the above techniques, a profit target is determined and then that target is translated into a percentage of premium of assets to make for a simple risk charge when constructing a price indication.

The consequences of failing to recognise an aspect of risk in pricing will likely be that the company will accumulate larger than expected concentrations of business with higher amounts of that neglected risk aspect.\(^{14}\)

To get consideration right you need to: (i) regularly get a second opinion on price adequacy either from the market or from a reliable experienced person; (ii) constantly update your view of your risks in the light of emerging experience and market feedback; and (iii) recognise that high sales is a possible market signal of underpricing.

**Provisioning for risks**

There are two levels of provisioning for an insurer, reserves and risk capital. The two are intimately related. In fact, in some cases, insurers will spend more time and care in determining the correct number for the sum of the two, called total asset requirement (TAR) by some.

Insurers need a realistic picture of future obligations long before the future is completely clear. This is a key part of the feedback mechanism. The results of the first year of business are the most important indication of business success for non-life insurance. That view of results depends largely upon the integrity of the reserve value. This feedback information affects performance evaluation, pricing for the next year, risk analysis and capital adequacy analysis, and capital allocation.

The other part of provisioning is risk capital. Insurers also need to hold capital for less likely swings in potential losses. This risk capital is the buffer that provides for the payment of policyholder claims in a very high proportion of imagined circumstances. The insurance marketplace, the rating agencies and insurance regulatory bodies all insist that the insurer holds a high buffer for this purpose. In addition, many valuable insights into the insurance business can be gained from careful analysis of the data that is input to the provisioning process for both levels of provisioning.

However, reserves are most often set to be consistent with considerations. Swings of adequate and inadequate pricing are tightly linked to swings in reserves. When reserves are optimistically set capital levels may reflect the same bias. This means that inadequate prices can ripple through to cause deferred recognition of actual claims costs as well as under-provisioning at both levels. This is more evidence that consideration is key to risk management.

There is often pressure for small and smooth changes to reserves and risk capital but information flows and analysis provide jumps in insights both as to expectations for emerging losses as well as in terms of methodologies for estimation of reserves and capital. The business pressures may threaten to overwhelm the best analysis efforts here. The analytical team that prepares the reserves and capital estimates needs to be aware of and be prepared for this eventuality. One good way to prepare for this is to make sure that management and the board are fully aware of the weaknesses of the modelling approach, and so are more prepared for the inevitable model corrections.
Insurers need to have a validation process to make sure that the sum of reserves and capital is an amount that provides the degree of security that is sought. Modellers must allow for variations in risk environment as well as the impact of risk profile, financial security and risk management systems of the insurer in considering the risk capital amount. Changes in any of those elements may cause abrupt shifts in the amount of capital needed.

The TAR should be determined without regard to where the reserves have been set so that the risk capital level does not double up on redundancy or implicitly affirm inadequacy of reserves.

The capital determined through the provisioning process will usually be the key element to the risk portfolio process. That means that accuracy in the subtotals within the models is just as important as the overall total. The common practice of tolerating offsetting inadequacies in the models may totally distort company strategic decision-making.

**Portfolio of risks**

In 1952, Harry Markowitz wrote the article ‘Portfolio selection’, which became the foundation for the theory called modern portfolio theory (MPT). MPT promises a path to follow to achieve the maximum return for a given level of risk for an investment portfolio.

It is not clear who first thought to apply the MPT ideas to a portfolio of risks in an insurer. In 1974, Gustav Hamilton of Sweden’s Statsforetag proposed the ‘risk management circle’ to describe the interaction of all the elements in the risk management process, including assessment, control, financing and communication. In 1979, Randell Brubaker wrote about ‘Profit maximisation for a multi-line property/ liability company’. Since then, the idea of risk and reward optimisation has become to many the actual definition of ERM.

Standard & Poor’s calls the process ‘strategic risk management’. Strategic risk management (SRM) is the Standard & Poor’s term for the part of ERM that focuses on both the risks and returns of the entire company. Although other aspects of ERM mainly focus on limiting downside, SRM is the process that will produce the upside, which is where the real value added of ERM lies.

The ‘risk portfolio management’ process is nothing more or less than looking at the expected reward and loss potential for each major profit making activity of an insurer and applying the MPT ideas of portfolio optimisation to that risk and reward information.

At the strategic level, insurers will leverage the risk and reward knowledge that comes from their years of experience in the insurance markets as well as from their ERM systems to find the risks where their company’s ability to execute can produce better average risk-adjusted returns. They then seek to optimise the risk/reward mix of the entire portfolio of insurance and investment risks that they hold. There are two aspects of this optimisation process. First is the identification of the opportunities of the insurer in terms of expected return for the amount of risk. The second aspect is the interdependence of the risks. A risk with low interdependency with other risks may produce a better portfolio result than another risk with a higher standalone return on risk but higher interdependence.

Proposals to grow or shrink parts of the business and choices to offset or transfer different major portions of the total risk positions can be viewed in terms of risk-adjusted return. This can be done as part of a capital budgeting/strategic resource allocation exercise.
and can be incorporated into regular decision-making. Some companies bring this approach into consideration only for major ad-hoc decisions on acquisitions or divestitures and some use it all the time.

There are several common activities that may support the macro-level risk exploitation.

**Economic capital (EC):** EC flows from the provisioning principle. EC is often calculated with a comprehensive risk model consistently for all of the actual risks of the company. Adjustments are made for the imperfect correlation of the risks. Identification of the highest-concentration risks, as well as the risks with lower correlation to the highest-concentration risks, is risk information that can be exploited. Insurers may find that they have an advantage when adding risks to those areas with lower correlation to their largest risks, if they have the expertise to manage those risks as well as they manage their largest risks.

**Risk-adjusted product pricing:** the portfolio principle can inform the consideration principle when product pricing is ‘risk-adjusted’ using one of several methods. One such method is to look at expected profits as a percentage of EC resulting in an expected return-to-risk capital ratio. Another method reflects the cost of capital associated with the economic capital of the product as well as volatility of expected income. The cost of capital is determined as the difference between the price to obtain capital and the rate of investment earnings on capital held by the insurer. Product profit projections then will show the pure profit as well as the return for risk of the product. Risk-adjusted value added is another way of approaching risk-adjusted pricing.

**Capital budgeting:** the capital needed to fulfil proposed business plans is projected based on the EC associated with the plans. Acceptance of strategic plans includes consideration of these capital needs and the returns associated with the capital that will be used. Risk exploitation as described above is one of the ways to optimise the use of capital over the planning period. The allocation of risk capital is a key step in this process.

**Risk-adjusted performance measurement (RAPM):** financial results of business plans are measured on a risk-adjusted basis. This includes recognition of the cost of holding the economic capital that is necessary to support each business as reflected in risk-adjusted pricing as well as the risk premiums and loss reserves for multi-period risks, such as credit losses or casualty cover. This should tie directly to the expectations of risk-adjusted profits that are used for product pricing and capital budgeting. Product pricing and capital budgeting form the expectations of performance. RAPM means actually creating a system that reports on the degree to which those expectations are or are not met. For non-life insurers, ‘risk portfolio management’ involves making strategic trade-offs between insurance, credit (on reinsurance ceded) and all aspects of investment risk based on a long-term view of risk-adjusted return for all of their choices.

Insurers that do not practice portfolio risk management usually fail to do so because they do not have a common measurement basis across all of their risks. The recent move by many insurers to develop economic capital models provides a powerful tool that can be used as the common risk measure for this process. EC is most often the metric used to define risk in the risk/reward equation of insurers.

Some insurers choose not to develop an EC model and instead rely upon rating agency or regulatory capital formulas. The regulatory and rating agency capital formulas are by their nature broad market estimates of the risk capital of the insurer. These formulae will
Key principles of enterprise risk management for insurance groups

overstate the capital needs for some of the insurer’s activity and understate the needs for others. The insurer has the specific data about their own risks and can do a better job of assessing their risks than any outsider could ever do. In some cases, insurers took high amounts of catastrophe exposure or embedded guarantee and option risks, which were not penalised in the generic capital formulas. In the end, some insurers found that they had taken much more risk than their actual loss tolerance or capacity.

Risk portfolio management (RPM) provides insurers with the framework to take full advantage of the power of diversification in their risk selection. They will look at their insurance and investment choices based on the impact, after diversification, on their total risk/reward profile. These insurers will also react to the cycles in risk premium that exist for all of their different insurance risks and for all of their investment risks in the context of their total portfolio.

Sales of most insurance company products result in an increase in the amount of capital needed by the business due to low or negative initial profits and the need to support the new business with ‘economic capital’. After the year of issue, most insurance company products will show annual releases of capital both due to the earnings of the product as well as the release of supporting capital that is no longer needed due to terminations of prior cover. The net capital needs of a business arise when growth (new sales less terminations) is high and/or profits are low and capital is released when growth is low and/or profits are high.

The definition of the capital needs for a product is the same as the definition of distributable earnings for an entire business: projected earnings less the increase in economic capital. The capital budgeting process will then focus on obtaining the right mix of short and long-term returns for the capital that is needed for each set of business plans.

Both new and existing products can be subjected to this capital budgeting discipline. A forecast of capital usage by a new product can be developed and used as a factor in deciding which of several new products to develop. In considering new and existing products, capital budgeting may involve examining historic and projected financial returns.

Pitfalls of RPM

In theory, optimisation processes can be shown to produce the best results for practitioners. And for periods of time when fluctuations of experience are moderate and fall comfortably within the model parameters, continual fine-tuning and higher reliance on the modelled optimisation recommendations produce ever growing rewards for the expert practitioner. However, model errors and uncertainties are magnified when management relies upon the risk model to lever up the business. And at some point, the user of complex risk models will see that levering up their business seems to be a safe and profitable way to operate. When volatility shifts into a less predictable and/or higher level, the highly levered company can find itself quickly in major trouble.

Even without major deviations of experience, the RPM principles can lead to major business disruptions. When an insurer makes a major change in its risk profile through an acquisition or divestiture of a large part of their business, the capital allocation of all other activities may shift drastically. Strict adherence to theory can whipsaw businesses as the insurer makes large changes in business.
Insurers need to be careful to use the risk model information to inform strategic decisions without overreliance and abdication of management judgment. Management should also push usage of risk and reward thinking throughout the organisation. The one assumption that seems to cause the most trouble is correlation. The saying goes that ‘in a crisis, all correlations go to one’. If the justification for a major strategic decision is that correlations are far from one, management should take note of the above saying and prepare accordingly. In addition, management should study the variability of correlations over time. They will find that correlations are often highly unreliable and this should have a major impact on the way that they are used in the RPM process.

**Future risks**

Often called emerging risks. Going back to Knight’s definitions of risk and uncertainty, there is very little risk contained in these potential situations. Emerging risks are often pure uncertainty. Humans are good at finding patterns. Emerging risks are breaks in patterns.

Emerging risks are defined by Easop et al. as ‘new or evolving risks that are difficult to manage because their identification, likelihood of occurrence, potential impacts, timing of occurrence or impact, or correlation with other risks, are highly uncertain’. An example from the past is asbestos; other current examples could be problems deriving from nanotechnology, genetically modified food, climate change and so on. Lloyd’s, a major sufferer from the former emerging risk of asbestos, takes emerging risks very seriously. They think of emerging risks as ‘an issue that is perceived to be potentially significant but which may not be fully understood or allowed for in insurance terms and conditions, pricing, reserving or capital setting’.

Easop et al. suggest that insurers need ‘sound risk management practices relative to its risk profile and considering the risks inherent in the liabilities it writes, the assets it acquires and the market(s) in which it operates, and takes into consideration new and emerging risks’. In 2013, AM Best added a question asking insurers to identify emerging risks to the ERM section of the supplemental rating questionnaire (SRQ). ERM has been one of the five major pillars of the Standard & Poor’s Insurance ERM ratings criteria since 2006.

The Global Risks 2012 report of the World Economic Forum is based on a survey of 469 experts from industry, government, academia and civil society that examines 50 global risks. Those experts identified 8 of those 50 risks as having the most significance over the next 10 years: chronic fiscal imbalances; cyber attacks; extreme volatility in energy and agriculture prices; food shortage crises; major systemic financial failure; rising greenhouse gas emissions; severe income disparity; and water supply crises.

This survey method for identifying or prioritising risks is called the Delphi method and can be used by any insurer. Another popular method is called environmental scanning, which includes simply reading and paying attention for unusual information about situations that could evolve into future major risks.

Many companies do not have any process to consider emerging risks. At those companies, managers usually dismiss many possible emerging risks as impossible. It may be the company culture to scoff at the sci-fi thinking of the emerging risks process. The process Taleb (2008) describes of finding ex-post explanation for emerging Black Swan risks is
often the undoing of careful plans to manage emerging risk. In addition, lack of imagination causes some managers to conclude that the past worst case is the outer limit for future losses.

The objectives for emerging risks management are just the same as for other more well-known risks: to reduce the frequency and severity of future losses. The uncertain nature of emerging risks makes that much more difficult to do cost effectively. Insurers can use scenario testing to examine potential impact of emerging risks and to see what actions taken in advance of their emergence might lessen exposures to losses. This scenario testing can also help to identify what actions might lessen the impact of an unexpected loss event that comes from a very rapidly emerging risk. Finally, insurers seek to identify and track leading indicators of impending new risk emergence.

Reinsurance is one of the most effective ways to protect against emerging risks, second only to careful drafting of insurance contract terms and conditions. Many of the largest insurers and reinsurers have developed very robust practices to identify and to prepare for emerging risks. Other companies can learn from the insurers who practice emerging risk management and adapt the same processes to their emerging risks.

Normal risk control processes focus on everyday risk management, including the management of identifiable risks and/or risks where uncertainty and unpredictability are mitigated by historical data that allow insurers to estimate loss distribution with reasonable confidence. Emerging risk management processes take over for risks that do not currently exist but that might emerge at some point due to changes in the environment. Emerging risks may appear abruptly or slowly and gradually, are difficult to identify, and may for some time represent an ill-formed idea more than factual circumstances. They often result from changes in the political, legal, market or physical environment, but the link between cause and effect is fully known in advance. As mentioned above, an example from the past is asbestos; other examples could be problems deriving from nanotechnology, genetically modified food, climate change, and so on. For these risks, normal risk identification and monitoring will not work because the likelihood is usually completely unknown. Nevertheless, past experience shows that when they materialise, they have a significant impact on the insurers and therefore cannot be excluded from a solid risk management program. Insurers have, therefore, implemented unique specific strategies and approaches to cope with them properly.

Emerging risks have not yet materialised or are not yet clearly defined and can appear abruptly or very slowly. Therefore, having some sort of early warning system in place, methodically identified either through internal or external sources, is very important. To minimise the uncertainty surrounding these risks, insurers will consistently gather all existing relevant information to amass preliminary evidence of emerging risks, which would allow the insurer to reduce or limit growth of exposure as the evidence becomes more and more certain. However, insurers practicing this discipline will need to be aware of the cost of false alarms.

Assess the relevance (that is, potential losses) of the emerging risks linked to a company’s commitment – which classes of business and existing policies would be affected by the materialisation of the risk – and continue with the assessment of the potential financial impact, taking into account potential correlation with other risks already present in the company. For an insurer, the degree of concentration and correlation of the risks that they have taken on from their customers are two important parameters to be considered; the risk in question could be subject to very low frequency/high intensity manifestations, but if exposure to that
particular risk is limited, then the impact on the company may not be as important. On the other hand, unexpected risk correlations should not be underestimated; small individual exposures can coalesce into an extreme risk if underlying risks are highly interdependent. When developing extreme scenarios, some degree of imagination to think of unthinkable interdependencies could be beneficial.

A further practice of insurers is to sometimes work backwards from concentrations to risks. Insurers might envision risks that could apply to their concentrations and then track for signs of risk emergence in those areas. Some insurers set risk limits for insurance concentrations that are very similar to investment portfolio credit limits, with maximum concentrations in specific industries in geographic or political regions. In addition, just as investment limits might restrict an insurer’s debt or equity position as a percentage of a company’s total outstanding securities, some insurers limit the percentage of cover they might offer in any of the sectors described above.

Responses to emerging risks might be part of the normal risk control process, that is, risk mitigation or transfer, either through reinsurance (or retrocession) in case of insurance risks, through the financial markets for financial risks, or through general limit reduction or hedging. When these options are not available or the insurer decides not to use them, it must be prepared to shoulder significant losses, which can strain a company’s liquidity. Planning access to liquidity is a basic part of emerging risk management. Asset-selling priorities, credit facilities with banks and notes programs are possible ways of managing a liquidity crisis.

Apart from liquidity crisis management, other issues exist for which a contingency plan should be identified in advance. The company should be able to quickly estimate and identify total losses and the payments due. It should also have a clear plan for settling the claims in due time so as to avoid reputation issues. Availability of reinsurance is also an important consideration. If a reinsurer were exposed to the same risks, it would be a sound practice for the primary insurer to evaluate the risk that the reinsurer might delay payments.

For the risks that have been identified as most significant and where the insurer has developed coherent contingency plans, the next step is to create and install an advanced warning process. To do that, the insurer identifies key risk indicators that provide an indication of increasing likelihood of a particular emerging risk.

Finally, sound practices for managing emerging risks include establishing procedures for learning from past events. The company will identify problems that appeared during the last extreme event and identify improvements to be added to the risk controls. In addition, expect to get better at each step of the emerging risk process with time and experience.

But emerging risk management costs money, and the costs that are most difficult to defend are the emerging risks that never emerge. A good emerging risk process will have many more misses than hits. Real emerged risks are rare. A company that is really taking emerging risks seriously will be taking actions on occasion that cost money to perform and possibly include a reduction in the risks accepted and the attendant profits. Management needs to have a tolerance for these costs; but not too much tolerance.

These seven principles of ERM for insurers can be seen as forming an ‘enterprise risk control cycle’ (see Exhibit 9.2). The cycle starts with assessing and planning for risk taking. That process may include the diversification principle and/or the portfolio principle, next to the steps of setting considerations and underwriting the risks. These steps are sometimes
operated together and sometimes separate, usually depending upon the degree to which the risks are small and homogeneous or large and unique.

The risk control cycle is then applied to the risks that have been accepted. That step is needed because even if a risk is properly priced and appropriately accepted, the insurer will want to manage the aggregate amount of such risks. Within the risk control cycle, there is a risk mitigation step and within that step an insurer may choose to reduce their total risk or to increase their risk taking capacity. Risks that have been accepted through the underwriting process and that the insurer is retaining after the risk control cycle process must be assessed for provisioning, both for reserve and capital.

Finally, for this discussion of the ERM cycle, the insurer needs to consider whether there are additional risks that have been unknowingly accepted that may emerge in the future. The future risk principle provides a path for that step.

For the ‘enterprise risk cycle’, there is actually no such thing as ‘finally’. As a cycle, it repeats infinitely. Exhibit 9.2 has many two headed arrows in addition to the one way arrows that represent a single circular process. The ERM idea sits in the middle of these seven principles. The ERM idea is the idea that an insurer will follow a cycle like this for all of its risks and in addition for the aggregation of all risks. This will be done to protect all of the stakeholders of the insurers – policyholders, stockholders, bondholders, management, employees and communities – to the greatest extent that their sometimes contradictory interests allow.

Most companies will put different degrees of emphasis on different elements. Some will have very faint arrows between ERM and some of the other principles. Some insurers will neglect some of these principles completely. It may be that the choice of which principles to emphasise is tightly linked with their view of the risk environment.

**Risk environment**

There are an infinite number of possible risk environments, but those possibilities can be grouped into four groupings.24

1. A benign environment when risk taking is generally rewarded. Companies that do not expand during this environment are often left behind. This is also called a ‘boom’ environment.
2. An adverse environment when risk taking is generally penalised. Companies that do not quickly trim their risks are usually devastated. Also called a ‘bust’ environment.
3. A moderate environment when there may well be significant ups and downs but a knowledgeable and disciplined company can do well. Careful risk taking is rewarded while reckless risk takers are penalised in the ‘moderate’ environment.
4. An ‘uncertain’ environment where it is simply unclear what to expect. Things are generally not predictable and those who expect predictability and act upon that belief are usually disappointed.

Exhibit 9.6 looks at the US housing prices and provides a very clear real world picture of these four risk environments. During the period through roughly January 2001, the housing market followed the moderate environment pattern. There was definitely real volatility of
housing prices, but that volatility was constrained. Sometime in 2001, the prices took off, heading in a steep upward path with little volatility. Those who completely disregarded risk of price decline, like the house ‘flippers’ who bought houses for the sole purpose of reselling them in a few weeks or months seemed to be earning totally riskless profits. That is the boom environment. Sometime in 2007, that boom came to an end and a bust started. That is when mortgage-linked securities went bad and flippers lost out. Mortgages became ‘underwater’ as prices dropped below mortgage amounts. But since 2009, the market has floundered, with spikes up and down but without movement in any particular direction for long. The uncertain environment is like that, mostly unpredictable.

Exhibit 9.6

Performance of US house prices

In each risk environment, different ERM principles will seem more important.

- In the moderate environment, provisioning, control cycle and portfolio will seem to be the most important of the seven ERM principles. In fact, many observers would define
ERM with just those three principles. That misunderstanding is a result of the lengthy moderate environment experienced in many of the developed economies for a period of 20 years or more. That moderate environment came to be thought of as the ‘normal’ environment and the techniques that work in that environment as the ‘correct’ way to manage risks.

- In the boom environment, the only bad risks are the mispriced risks. So in that environment, underwriting and consideration are the important ERM principles. Insurers can flourish and grow safely with just those two principles during the boom.
- But in the bust environment, priorities shift to provisioning, control cycle and future risks. However, during the bust environment, the approach to provisioning and control cycle is very different to when those principles are applied in the moderate environment. During the bust, provisioning will tend to be conservative. Loss reserves are set to a level that is sure to exceed potential losses, rather than a realistic middle of the road pick. Capital models may shift over to stress tests and a fixed stress scenario seems to be safer than a stochastic model where it is unclear why it gets the answer that it does. Control cycles are run with risk appetites that are also conservative estimates of a loss that the insurer almost expects to experience.
- Uncertain environments shift the focus to diversification, underwriting and future risks. Management will want to be sure that they are not subject to any one concentrated risk that will put an end to the company. They will also want to be very hands on with underwriting. And future risks are considered to be real live possibilities that management takes seriously.

**Conclusion**

Insurance companies have always existed at the intersection of risk and opportunity. For the entire history of an organised insurance industry, risk management has been fundamental to the reliable assurance of payment of claims when due to their customers. When ERM became a ‘new’ management practice in the mid-1990s, the ERM practitioners added the ‘portfolio’ principle. To accomplish what they wanted with the portfolio principle, they needed to expand the ‘provisioning’ principle to incorporate risk capital modelling. This was finally practical with the advances in computer power. Finally, the ‘control cycle’ principle was used to further link together provisioning with portfolio via a capital budgeting process and a control cycle of aggregate risk.

This definition of ERM was made without necessarily an admission that ERM actually required a company had in place practices based upon the principles of underwriting, consideration and provisioning for reserves.

Finally, as the extreme volatility in almost every area through the first decade of the new century taught us that the risk models did not have all of the answers, future risks became the final piece of a full ERM system.

Insurers can look forward with recognition that they must pay careful attention to the risk environment to make sure that they are emphasising the right principles just as we all hope that we wear the right coat for the weather when we set out each morning.
Bibliography


19 Knight, FH, Risk, Uncertainty and Profit, 1921, Signalman Publishing.
Overview

This chapter builds upon Greenbaum,1 a first-pass at a conceptual framework for enterprise risk management (ERM). Because a good bit of confusion is owing to the definition of primitives, we begin by deconstructing the concepts of risk, risk management and ERM. Essentially exegetical, we address the aetiology and etymology of these building blocks. Following is the fundamental distinction between ‘core’ and ‘ancillary’ risks that we use to explain risk management. We distinguish between the risks the organisation processes in pursuit of profit versus those it avoids as a nuisance. Both are managed but, as explained, quite differently. Finally, we offer normative ideas about how to organise the ERM effort. Because ERM was motivated by glaring errors of omission, process is central and hence best practices are overarching.

Introduction

We begin this chapter by deconstructing the concepts of risk, risk management and ERM. These are all unsettled ideas, but they have been adapted and deployed in the effort to advance management techniques in public companies and other types of organisations concerned with their mortality. We next explore the distinction between core risks, those that businesses process profitably and ancillary risks, all others that constitute a nuisance in that the company possesses no special skills in dealing with them. In the final section, we share some normative thoughts about ERM as a process. The perspective is unabashedly that of the board member. At the end of the day, ERM is little more than formality designed to reduce the threat of terminal hazards, hazards that tend to be ignored in the normal course of events.

Risk, risk management and ERM

ERM was born out of 25 years of destabilising organisational failures widely attributed to errors of omission in risk management. These derelictions were in some cases venal, in others merely naïve. In public company settings, they commonly represented violations of accepted standards of governance with respect to loyalty, care and good faith. The modifier
‘enterprise’ in ERM is advised because risk management has a long and chequered history, widely judged to have failed to adequately sensitise management to major business risks. Gaps and voids were identified, indicting traditional practices as flawed and incomplete. ERM was designed to correct the noted shortcomings.

However, best practices of ERM are still in their infancy. Codification began only in the past decade with a major reorientation in the interim. Large financial institutions are among the leaders in refining ERM practices, but these also are the institutions that have experienced the most egregious failures. Financial institutions are among the most intensively regulated and the public regulators have been a major spur encouraging the reform of risk management.

Originally, ERM was seen as correcting errors of omission by perfecting the communication of a comprehensive and integrated set of risk measures that the board and management of public companies could use in developing and vetting business strategy. Full integration proved overly ambitious and quickly gave way to a more workable focus on those risks that threaten organisational sustainability. These were the most amorphous, ill-defined and intimidating risks, for the most part resident in the negative tail of the applicable probability distribution.

Risk

Kahneman explains decision theorists focus on choice among gambles in universal terms. All decisions are thought to have consequences that are known only to a first approximation. Apart from contrived closed systems, decisions result in uncertain outcomes and are therefore probabilistic. Taleb’s admonitions to the contrary notwithstanding, (in)decisions necessitate forecasts, if only implicitly, and the distillate of the consequent uncertainty is what we commonly comprehend as risk. Explicit representations of risk vary according to application and typically involve selected parameters (moments) of a conjoined notional probability distribution. For some purposes range or variance will do, whereas other applications focus on extreme values, and in some cases only negative tail parameters.

Knight defines a special case of risk where the underlying probability distribution is ‘unknown’. He refers to this special case as uncertainty. This is more or less what Taleb means by black swans and these, he argues, confound forecasting. Tail events that occur with minuscule probabilities are poorly defined, but can be particularly impactful. That is, even small errors in assessing the probabilities of extreme events can produce large deviations in outcomes. In economic terms, the elasticity of outcomes with respect to associated probabilities tends to be large in the tails of probability distributions.

But it is the negative tail of the probability distribution that excites the greatest interest among students of enterprise risk, and unsurprisingly it is in this neighbourhood that the greatest controversy seems to reside. For example, Kahneman and Tversky argue that loss aversion leads to overweighting negative tail events. Both positive and negative tail events are over-weighted relative to subjective probabilities according to Kahneman and Tversky, but since losses are more painful than equivalent gains are satisfying, the negative tail is over-weighted relative to the positive tail. In contrast, Taleb describes unprecedented events, summary ignorance and hence under-weighting of their probability. There is a related behavioural paradigm of denial: some outcomes are so odious and potentially paralytic that
they induce neglect. In addition, the probability of tail events is so small and ill-defined that we heuristically set it to zero. This was the case at TEPCO in considering the prospect of a tsunami disabling their Fukushima-Daiichi nuclear plant.10

In any case, the Kahneman and Tversky, and Taleb stories are conflicting explanations of decision making in the negative tail. Taleb invokes a kind of Knightian uncertainty and infers underpricing of negative tail risks. Kahneman and Tversky, on the other hand, posit a hypersensitivity to large losses and a willingness to insure and ensure against horrific outcomes.11 In his earlier analysis of capital markets, Mandelbrot12 presents a story similar to Taleb’s. He observes that the widespread use of the Gaussian distribution, owing to its seductive two-parameter simplicity, has led to systematic underestimation of equity price volatility. Mandelbrot introduces the notion of ‘chubby’ tails with the implication that this kind of capital market risk is underpriced.

We conclude that risk is a term used to describe uncertain prospects, but since risk metrics are situational, the term takes on a variety of meanings as varied as the problems being analysed. Precision, therefore, is available only in application. Metrics that focus on extreme events are especially fraught owing to paucity of experiential data and associated horrific, often terminal, outcomes. The latter quality, in particular, makes these risks both unimaginable and unthinkable. Not surprising then that the best contemporary thinking on the subject is disparate.

**Risk management**

The management of risk, however defined, is as time-honoured as the business organisation, even if formal risk management is a contemporary development.13 Weatherall14 traces the use of financial derivatives back to Mesopotamia. In the Italian city states of the 15th and 16th centuries, letters of credit were used to insure the safe return of commercial ships dispatched on dangerous trading forays. Limited liability corporations and public companies are likewise instruments of risk diversification and redistribution.15

Risk management, until recently, focused for the most part on individual exposures within organisations. Insurable risks, such as ‘errors and omissions’ or ‘directors and officers liability’ or capital market risks (liquidity, interest rate, foreign exchange, or counterparty issues), were addressed with financial derivatives and/or insurance policies. Other risks were addressed with operational measures such as slack and sourcing diversification to ensure the integrity of supply chains.

Risk management remained essentially selective and fragmented and its hallmark was the independent treatment of more or less well-defined hazards within the business. The approach was notably ‘siloed’ and the principal domain of the risk manager was insurance and capital markets. As a consequence, risk management tended to ignore less obvious hazards, especially tail risks and covariation among risks within the organisation. Spectacular public company failures of the past two decades drew attention to these omissions. The savings and loan collapse of the 1990s saw a whole industry obliterated by a protracted yield curve inversion. Following were the millennial scandals of Enron, WorldCom, HealthSouth and Arthur Andersen, attributed to financial control and governance derelictions widely interpreted as risk management failures.

www.rasabourse.com
The shortcomings of traditional risk management practices were ultimately confirmed by the Great Recession of 2008, the BP Gulf of Mexico oil spill and the TEPCO nuclear disaster at Fukushima-Daiichi. Each displayed a striking neglect or dismissal of a major tail risk. In the case of the Great Recession, it was a collective failure to appreciate numerous warnings of an ominous housing bubble together with the unregulated and expansive growth of a ‘shadow’ banking system.\textsuperscript{16,17} In the case of BP, it was failure to deploy a recommended second blowout protector, a US$50 million investment that almost certainly would have prevented a subsequent US$50 billion loss. In the case of Fukushima-Daiichi, the protective seawall was built to a height of 15 foot on the argument that a 14-foot tsunami was the largest experienced at the time, notwithstanding the scientific community’s recommendation of a 30-foot barrier on the argument that global warming was exacerbating climatic convulsions. TEPCO internal documents described the risk of a tsunami threat to Fukushima-Daiichi as ‘unthinkable’. The TEPCO disaster may have been the most poignant indictment, if indeed one was needed, of traditional risk management techniques, a clarion call for the rethinking of public company risk management practices.\textsuperscript{18}

**ERM**

The upshot of this post-millennial rethinking of risk management practices, ERM is a process designed to mitigate the tendency to underestimate, and in the extreme, ignore certain types of catastrophic risks confronting the public company. In its early conception, ERM focused on the summation of all risks.\textsuperscript{19} It was comprehensive in seeking to encompass, aggregate and confront all risks. The concept of economic capital as commonly employed in the insurance industry captures this idea in that it seeks to identify all risks and assigns to each an appropriate amount of financial capital, allowing for covariation among risk categories. But practitioners are quick to acknowledge the practical limitations of economic capital, which is now seen as one among a kit of tools for triangulating on risk measurement.

The intellectual journey from traditional risk management to ERM is little more than a decade in the making, and yet in that brief time it has undergone a major directional reorientation. Originally motivated by the idea of correcting all errors of omission by integrating a disjoint process, more sophisticated practitioners quickly realised that the summation of all risks was something of an *ignis fatuus*, if not beside the point.\textsuperscript{20,21} This led to the reinterpretation of ERM as a process to sensitise the organisation to those hazards that threaten strategic intent or continuity. ERM became the guardian of organisational sustainability and thereby narrowed its focus to those hazards described in the negative tail of the probability distribution.\textsuperscript{22} Whereas this redefinition simplifies by excluding the more mundane, well-defined risks as well as windfalls, what remains are the most fraught and ill-defined hazards, those that tend to be ignored or denied while simultaneously feared and despised.

**The framework**

Risk management is commonly confused with risk mitigation because the risk management epiphany is commonly prompted by a recognition that the organisation is somehow displaced
from its risk-return efficiency frontier. Hazards are discovered because an embarrassing loss is sustained or covariation among risks is brought to the fore because an untoward event produces cascading secondary effects. Discovery of such avoidable losses typically prompt enhanced monitoring, transparency and process refinement. In short, an unintended consequence is revealed and remedied and risk management is risk mitigation. Much of ERM is of this type, inefficiency correction via process augmentation. A chief risk officer (CRO) is appointed, a board committee is established, transparency and awareness is heightened, and perforce operating costs increase. This process is important but pedestrian; more on this to follow in our discussion of best practices. However, there is a second, more strategic aspect of ERM that involves risk acquisition as well as mitigation.

Core risk

Financial institutions are in the business of processing risks. For example, commercial banks are defined by their special competency in processing credit risk whereas life insurance companies are identified with mortality risk. Perhaps less obvious, non-financial businesses are identified by a core risk deriving from a core competence. Consider Boeing: each airplane that they produce poses a risk to stakeholders. This risk is readily mitigated by reducing sales, but this is clearly not a strategy for success. Rather, Boeing is eager to assume the kind of risk that accompanies the sale of its airplanes. Similarly, Toyota suffered costly embarrassment owing to accidents traced to the braking and accelerator mechanisms of its automobiles. The discovery of these risks did not prompt a desire to curtail the sale of automobiles. Quite the contrary, Toyota sought to expeditiously fix the mechanical problems, address the attendant lawsuits, and loss of goodwill and then struggled mightily to reclaim its erstwhile position as the largest automobile producer in the world. These risks that companies seek to acquire and process are their stock in trade. They are core risks that they can absorb profitably because they possess a competitive advantage in processing them.

Commercial banks extend credit because they are able to manage credit risk with less Sturm und Drang than other agents in society. Hence, counterparties are willing to pay banks a premium to do so. The banks have a core competence relating to credit risk that others lack. Moreover, provided that these special skills, which are usually embedded in human and financial capital inputs, are scalable, the appetite of the bank for credit risk is potentially limitless. It is only the fixity, however temporary, of the key inputs together with possible spillover effects from ancillary risks that tempers banks’ desire to process credit risk indefinitely. Therefore, managing core risk centres around managing input scalability and ancillary risk. Input scalability is the domain of resiliency that Sheffi and Senge inform.

Core risks are fundamentally and qualitatively different from those the company acquires incidental to doing business, but for which it possesses no core competence. These we refer to as ancillary, and ancillary risks do not produce a surplus for the company. Indeed, these are risks that others may possess a core competence in processing and are therefore candidates for risk shifting. However, the company also may confront ancillary risks that are not tradable because agents with a core competence in managing them are unavailable or simply do not exist.
Resiliency is all about controlling or reducing the adjustment cost of inputs. These include the cost of acquiring incremental financial capital for which the company may face a rising supply curve or similarly restraining adjustment costs. In the case of human capital, similar considerations prevail. Specialised skills are not available to the individual bank or industry under conditions of infinitely elastic supply. The commercial bank seeking to increase its lending will need readily available originating, monitoring, servicing, renegotiating and legal capabilities. These skills must either be internally diverted, acquired in the marketplace, or trained. Adjustment costs are inescapable even with far-sighted investment in resilience. It is these adjustment costs that create friction and bound the appetite for core risks.

Of course, the bank may expand its lending with fixed inputs, but layering more and more throughput on fixed inputs will eventually dissipate the premium to the bank’s core competence. With infinitely elastic supply of the critical inputs and absent adjustment costs, one can imagine the bank profiting indefinitely from the expansion of lending. Under ideal, but hardly relevant, circumstances the appetite for core risk becomes unbounded.

With positive adjustment costs, the appetite for core risk is bounded even if adjustment costs also give rise to opportunities for investment in resiliency. Investments in resiliency might include training for technical skills or relationship building in capital markets to improve access in anticipation of demand growth. These initiatives inflate current operating costs in exchange for future flexibility and therefore constitute the purchase of real options.

Adjusting inputs, either internally or by outsourcing, may be viewed as managing a given core competence, a kind of vertical adaptation. But, companies also may expand horizontally by acquiring a wider variety of core competencies. Thus when Travelers Insurance Company combined with Citibank to form Citigroup, the company expanded its core competencies. The credit risk processor became an insurer as well, arguably a distinct albeit related competence.

Core competence and associated risks are not immutable. Although substantially a legacy, core competencies can be augmented horizontally as well as vertically, but typically at a cost that is convex increasing. We used the example of Travelers acquiring a peripheral core competence, but we might have chosen a congeneric such as General Electric, Tyco, or ITT, which are (were) agglomerations of orthogonal core competencies, but nevertheless subject to similar principles of growth. In the short run, the company’s core competencies are fixed, but in the fullness of time core competencies are adjustable and define a central aspect of business strategy.

Ancillary risks

According to our taxonomy, there are but two kinds of risks, core and all others. All other risks we refer to as ancillary. As explained earlier, core risks define the business. It is these risks that companies profit by processing provided that they have access to those critical inputs in which their core competencies are embodied, typically specialised human and financial capital. It is these core risks for which the company displays a risk appetite, an appetite limited in the first instance by their access to key inputs.

Ancillary risks are quite another matter in that the company possesses no special competence in processing these risks even if others may. Therefore, managing ancillary risks is always directed toward mitigation. These risks absorb critical inputs without directly producing a
profit. They, therefore, reduce profit by reducing inputs available for the processing of core risks. Reinsuring, hedging, or otherwise shifting ancillary risks to others frees resources permitting the company to increase profit by processing additional core risks.

The relationship between core and ancillary risks makes the latter inescapable. The bank seeking to process credit risk must finance its loans. In doing so, it confronts interest rate and liquidity risks that typically are not core to the bank. The life insurance company eager to assume mortality risk finds itself with investable assets that again pose an interest rate risk as well as an asset selection challenge. Neither of these has any direct link to mortality risk and arguably is not core to the life insurer. Take Boeing as a non-financial example, a venerable company with deep engineering and manufacturing experience toward providing commercial airframes that provide swift and safe transportation over long distances. Their implied core risk is illustrated by the lithium-ion battery problems that grounded the Dreamliner fleet. But risks related to management succession, labour relations, fuel price volatility and macroeconomic conditions all represent hazards that are ancillary and unavoidable if Boeing is to continue selling commercial airframes.

Were core risks available without accompanying ancillary risks, the risk management problem would be simplified immensely. But separation is possible only as an abstraction. In the world of practice, core and ancillary risks arrive together, and extricability is achieved under conditions of convex increasing costs, so the company is almost always grudgingly left holding a residuum of nuisance ancillary risk in order to pursue its business of profitably processing what is core.

History is replete with stories of companies confusing ancillary with core risks and also badly estimating the parameters of ancillary risks. Banks often think they can profit by speculating on the yield curve. Others speculate in currencies and international interest rate spreads. The broadly construed ‘carry trade’ seems virtually irresistible. Insurance companies often pursue earnings by pushing the limits of prudence in their asset selection. Confusing ancillary risk with core can be a form of hubris wherein the company erroneously believes it possesses a competence. The gods do not treat kindly such arrogance. Examples abound of misguided willingness to accept a small probability of a large future loss in exchange for a highly probable, small and regular payment (premium). The implosive failure of Long-Term Capital Management might be offered as exhibit one.29 But describing others could become a popular parlour game. The systematic under-estimation of probabilities and/or outcomes for negative-tail ancillary risk is essentially a Mandelbrot-Taleb phenomenon. Difficult to gauge small probabilities are routinely set to zero and terminal losses are often ignored as inevitable.30

Dwelling on the possibility of a nuclear meltdown or major oil spill can be paralytic. Further, some ancillary risks may be neither core nor acceptable to anyone, so the Hobson’s choice is to accept the risk or exit the industry. Ancillary risks for which there is no market are either managed at a net cost or they become the predicate for termination of the business. Disparate attitudes toward nuclear power are suggestive of this sort of risk. Thus, even if critical inputs are readily available, the accumulation of ancillary risks not amenable to mitigation may limit the organisation’s appetite for risk.

It falls to the ERM process to catalogue, assess and address the ancillary risks so as to prevent their mishandling. There are but three ways to deal with these risks. They can be
shifted to others where a market exists. When one company’s ancillary risks are another’s core risks, markets tend to emerge for financial claims – derivatives, futures and forward contracts, swaps, letters of credit, loan commitments, (re)insurance policies – and the company can, at an often tolerable cost, divest itself of ancillary risks. In cases where there are no markets for particular risks, producers must either accept the risk and seek to price it or leave the industry. Neglecting ancillary risks leads to underpricing them, the bane of management.

**Plasticity of risks**

Risks are manageable and malleable by the adroit. Core risk capacity is augmented, both vertically and horizontally, and ancillary risks can be conflated with core risks. Outsourcing offers yet another example of how risks are transformed. Banks commonly outsource loan origination and servicing. Back office activities of banks are often performed by non-bank specialists. Boeing was notable in outsourcing component manufacture for the Dreamliner. These examples can be viewed as transformations of core into ancillary risks or as shifting ancillary risks. If the latter, then the activity is routine and likely benign. If the former, the company is presumably circumventing a capacity constraint and may be problematic. This latter case is interesting because the company would not ordinarily offload core risk (syndicated lending and reinsurance of oversized insurance policies are counter examples), especially if the reciprocal flow is ancillary risk.

The Boeing Dreamliner engendered an extraordinary amount of outsourcing, perhaps partly spurred by anticipated offshore airplane sales. However, at the time, the outsourcing was explained in terms of reducing costs and risks associated with development and engineering of a radically new design. This is an activity in which Boeing presumably possessed a core competence, even if the Dreamliner project was outsized. Boeing was, in a sense, shedding a core risk and accepting new ancillary risks which would ultimately delay completion of the project and elevate costs beyond plan.

The new ancillary risks were that subcontractors would fail to perform as contracts stipulated, a counterparty risk. Managing sub-contractors was not an activity foreign to Boeing, but the scale of the activity was new and Boeing proved to be unprepared to effectively supervise and monitor their suppliers. In retrospect, Boeing underestimated the counterparty challenge and found a number of sub-contractors unable to deliver components as promised. A core competence in monitoring subcontractors might have discovered these shortcomings at an earlier stage in the manufacturing process enabling mitigation.

Arguably, managing an extended supply chain requires a distinct set of skills, not identical to manufacturing. Boeing had already established bona fides in manufacturing airframes, but not nearly so convincingly in the realm of supply chain management, especially when the supply chain became abnormally elongated. Managing the supply chain involves establishing and communicating standards, contract design, monitoring and enforcing standards. Credit risk, the defining characteristic of commercial banks, is a form of counterparty risk. Nike, Apple and others fell victim to failures in supply chain management, but these companies were built on offshore outsourcing suggesting failures of core risk management. Boeing differed in having a rich history of internal research and development, engineering and manufacture. To be sure, they had previously outsourced components, namely engines and avionics, but
never had they carried the outsourcing as far as with the Dreamliner, where sections of the
airframe were manufactured by offshore companies. This suggests a transformation of core
into ancillary risk that should have given management and the Board of Directors pause.
At the time, the decision was explained in terms of dissipating risks, given that materials
and technological advances were so numerous in this new aircraft. But we are left with the
question of whether risk was dissipated or merely reconfigured.

Best practices in formalising the ERM process

Process is central to ERM because so many of the motivating organisational failures have
been ascribed to errors of omission. What was previously an informal aspect of management
became a formally structured set of procedures. This formalisation is well documented.32,33,34,35
Accountability is defined, best practices are specified, and the implementation is monitored.
Many of the worst decisions had previously simply slipped past appointed monitors, auditors,
compliance executives and supervisory personnel. In other cases, the monitoring structure was
non-existent. A well designed and implemented risk monitoring system could be expected
to reduce the probability of untoward outcomes. Formality added cost to be sure, but also
held promise of improving the quality of decision making. Much of the early work on ERM
addresses details of the process appropriate for managing major organisational risks. The
process forces major risks onto the agenda for regular deliberation by senior management
and boards of directors.

Errors of omission may be venal as seemed likely when Enron’s board waived their
corporate code of ethics to permit the chief finance officer (CFO) to own off-balance sheet
special purpose entities sponsored by Enron.36 The CFO was therefore engaged in self-dealing
with the permission of Enron’s board and in explicit violation of Enron’s corporate code
do ethics.37

Errors of omission may be the result of decision biases and heuristics of the type essayed
in Kahneman38 and discussed earlier with regard to TEPCO, where management wittingly
chose to ignore a major hazard because of a diminutive probability of its occurrence. Indeed,
the contemplated contingency was unprecedented, as in the case of Taleb’s Black Swan.39 Yet
a third possibility is that omission errors arise from unwitting neglect owing to incompetence
or negligence. All three sources of omission errors are amenable to mitigation via process
formality. Institutions, protocols and rules that mandate attention to risk and that sustain
candour and transparency reduce the probability of omission errors.40

ERM protocols

Localisation of responsibility for risk management at the board level may be the most
fundamental of risk management protocols. Even before the advent of ERM, the board’s
responsibility for risk management was implied in their vetting of business strategy. Risk
is a facet of strategy and the board’s time-honoured responsibility for probing, testing and
sceptically questioning management’s proffered business strategy should be seen as the most
basic mandate for risk management. ERM is then a restatement and formalisation of a
pre-existing responsibility.
It is rapidly becoming best practice for boards to have a standing committee dedicated to ERM. This is new. Virtually all large financial institutions follow the counsel of the Walker Report and have such a fourth standing committee. Board ERM committees are less common in non-financial industries where the responsibility for ERM is typically lodged in the board audit committee. However, it seems to be increasingly common for larger non-financial companies to appoint standing board committees focused on ERM. In the UK, this is now considered best practice among larger non-financial public companies.

Public companies have long had mandated audit, compensation and governance committees. Some boards have additional standing committees growing out of special needs of the company or industry. These board committees typically have three notable features. They are for the most part advisory, their mandates tend to be non-intersecting, and they do not typically address broad issues of business strategy as this topic is consciously reserved to the committee of the whole.

The ERM committee is unique in that its responsibilities overlap with virtually all the board’s committees. In addition, since ERM is an aspect of business strategy, the convention of reserving the discussion of business strategy to the board as a whole is unavoidably violated. With regard to the mandated committees, note that integrity of financial reporting, internal controls and compliance are all major business risks. Likewise, succession, managerial performance and board performance are similarly risks that have undone major public companies, hence the overlap of an ERM committee with the audit, compensation and governance committees.

This explains why a standing ERM board committee may be controversial, especially in the collegial and consensual culture of boardrooms. Many companies resist the creation of an ERM committee because it may challenge accepted board practices, even when the board’s committees are principally advisory. The ERM committee challenges the culture of corporate governance by encroaching on the responsibilities of other standing committees as well as the domain of strategy normally reserved for the committee of the whole. The ERM committee is, therefore, fundamentally disruptive, perhaps even abrasive, even if its redundancy is prudent.

The territorial overlap among the board’s committees resulting from the addition of a standing ERM committee has both divisive and positive aspects. Exclusivity among committee responsibilities almost surely fosters a live-and-let-live collegiality among board members that is considered desirable in corporate governance circles. But, the overlap introduced with an ERM committee may promote accountability. For example, common weaknesses in governance are often found in succession planning and board performance assessment. The former is typically a responsibility of the Compensation Committee, whereas the latter is commonly the purview of the Governance Committee. CEOs often resist succession planning for egocentric reasons. Succession planning also may be seen as disruptive among senior leadership aspirants to higher office. Board performance assessments are historically superficial because they are self-examinations, dubious in general, but especially questionable among the strong personalities commonly found on visible public company boards.

These areas of functional weakness benefit especially from the overlap of responsibility introduced with a standing ERM committee. Because board dysfunction and managerial shortcomings are major causes of organisational failure, the ERM committee will keep these
issues on the board’s agenda. The governance committee, therefore, will not find it easy to deal with board and director performance superficially nor will the compensation committee be able to blithely dismiss the management succession issue. The ERM committee, therefore, serves as a source of constructive tension, a catalyst heightening the general effectiveness of governance and management.

**Organisational design for ERM**

The creation of a standing ERM board committee has immediate implications for organisation design. An officer’s risk committee will be brought into existence if one is not already extant. This committee of senior officers exists to process the information that will normally become the grist for deliberations of the board committee. Indeed, the officer’s risk committee has the obvious job of digesting, interpreting or framing risk information for delivery to the board. Risk dialogue at the board level is normally tripartite, and the third party is the CRO, together with his/her staff. It is not uncommon to observe the creation of the CRO position almost simultaneously with the inception of the board and officer’s risk committee. These three organisational nodes then become the key elements in sustaining the risk dialogue within the organisation. The primary and unique responsibility of the CRO is the collection, organisation and dissemination of risk information. Given the wide variety and opacity of most tail risks, the collection and presentation of risk data is no small chore and the CRO will therefore require both staff and rapporteurs at every consequential operating and staff node throughout the organisation. Risk information flows from the bottom up and laterally as well, so the CRO will need reliable sources of information throughout the organisation and it will be his/her unit’s responsibility to maintain dialogue with all the organisational units capable of accepting risks on the organisation’s behalf. The London Whale’s US$6 billion loss illustrates how the failure of communication regarding risk can result in outsized financial and reputational losses.

Whereas the CRO is responsible for the timely assembly and dissemination of risk information, interpretation is the responsibility of all three risk management nodes. Acting upon this information falls to the board and senior management. Precisely how these organisational entities fit together in discharging their responsibilities is typically memorialised in the charters of board’s and officers’ ERM committees, the charge of the CRO, and a document that sets forth the risk culture of the organisation. The organisation’s risk appetite and tolerances are usually codified in subordinate documents.

**External versus internal information flows**

The organisational design for ERM provides for enhanced informational flows and structures that force attention at the highest organisational levels upon threats to sustainability. By facilitating the unencumbered flow of information and regular meetings with disciplined agendas for both the officers’ and directors’ risk committees, the ERM process ensures consideration of risk-based information relevant to major and physical hazards. But these hazards are in large part company-specific, originating with operations within the organisation. This follows from the ERM process playing out largely within the company with professionals
focused largely on internal operations. Consequently, there may be a tendency to slight those hazards originating from without. Who within the company is attentive to the butterfly’s wings flapping halfway around the world? Climate change, over-fishing of sharks, ageing of population, urbanisation pathologies, fresh water underpricing – who within the representative public company can essay the organisational implications of such hazards? The typical electric generating company is apt to have limited competence in social media, cyber-security and threats of expropriation. Yet, such external forces may well impinge upon the enterprise risk of the company. The ERM process must find a way to correct for the all-too-common tendency to focus unduly on internal hazards. Consultants and continuing education help, but these often prove less than totally satisfactory.

Institutionalising the flow of external information may prove as valuable as any other element in the risk management infrastructure. BP failed to install a second blowout protector and TEPCO failed to build the Fukushima-Daichi seawall to 30 foot despite expert advice to the contrary. External expertise is too easily ignored. Here too, process may be lacking. Had these external voices been empanelled as a board of advisers to the risk management community within the company, such externally sourced risks might not have been so readily dismissed. Meetings of external advisers would generate discoverable minutes or proceedings that the officers’ and directors’ risk committees could not readily ignore. Internalising the voices from without should mitigate the all-too-natural tendency for risk committees to be biased toward internal exigencies.

**Conclusion**

In large part, the motivation for ERM derives from underestimating the probability and/or outcomes of catastrophic hazards among public companies. In the argot of professionals, systematic underpricing of negative tail events served as the catalyst for the post-millennial development of ERM.

Not surprising then that the focus of ERM was a formalisation of monitoring of those risks that threatened organisational sustainability. By institutionalising what was formerly an informal process, ERM could at the very least lower the probability of ignoring a class of hazards that had wreaked havoc on private and collective wealth. A robust and resilient ERM process might even correct the estimation biases growing out of moral hazard, incompetence, or lethargy. Rules of thumb and similar heuristics might be replaced by more informed and less biased decisions. If the process accomplished little more than improved transparency and communication, risk management benefits could be expected to ensue. Business decisions would become less subject to distortions and short-termism, and the posterior allegations of litigants and public regulators could be rebutted more effectively.

The substantial costs of ERM are justified by a widely observed bias in decision-making with regard to negative tail events. Public companies and others as well, seem too eager to accept remote and improbable risks of large negative payoffs in exchange for seemingly regular and relatively modest but more immediate payoffs. This can be described as a generalised ‘carry trade’ problem or underpricing Black Swans, namely BP, AIG, TEPCO, LTCM, London Whale, and so on. The underlying behavioural drivers of this phenomenon are varied and conjectural. In public companies, moral hazards may be operative in that the beneficiaries of
the positive immediate cash flow may differ from the inheritor of the remote liability. Present management and the board may have interests that diverge. Similarly, present management and boards can be expected to have interests that diverge from future management and boards. The disclosures promoted by a robust and resilient ERM process will not eliminate these incentive problems, but it will illuminate them and risk mitigation can be predicted.

Even in the absence of moral hazard problems, decision making is distorted and impaired when negative outcomes are large, remote, and improbable. Probabilities are poorly defined owing to infrequency or myopia. Similarly, outcomes may be paralytic leading to denial. The records suggest that the TEPCO disaster at Fukushima-Daiichi may have been owing to this kind of impairment. In Kahneman’s terms, this syndrome leads to ‘thinking too fast’.

As indicated, ERM has cost entailments that must be weighed against improved decision making, especially with regard to negative tail hazards. Still another way to think about ERM is as a corrective for the excesses of total quality management (TQM), including Kaizen and Six Sigma. Their emphasis on expanded spans of control, just-in-time production and the elimination of virtually all buffer stocks, effectively reduced slack in production and distribution, and led to the elevation of all sorts of operating risks. But this again may be viewed as a form of carry trade whereby reduced current operating costs are achieved in exchange for elevated risks of remote breakdowns. Of course, the advocates of TQM did not frame their program in terms of this trade-off. Rejecting the possibility of disingenuousness, we again come to a myopia relating to remote, improbable, or outsized outcomes. Again, we confront the insights of Taleb, Mandebrout and Kahneman.

The author would like to acknowledge helpful suggestions from ME Greenbaum, H Novack, and D Ingram without implicating them for remaining errors of fact or judgment.

2 Notably, the impugned credit rating agencies, most especially Standard and Poors, have encouraged the development of ERM by making their examinations of risk management a routine part of the credit rating process.
5 Knight, F, Risk, Uncertainty and Profit, 1921, Cornell University Library.
11 Kahneman refers to reconciling discussions with Taleb, but we are left with the less than totally satisfying conjecture that unthinkable hazards, for whatever reason, are underpriced whereas those that can be conjured
Regulatory and enterprise risk

are overpriced. Unthinkable events may be unprecedented, as in Taleb’s Black Swan, or the simple product of a failure of imagination, or a psychological blockage as in denial. The distinction is murky at best (see Kahneman, D, *Thinking Fast and Slow*, 2012, Farrar, Straus and Giroux, especially page 333). But note that in either case, conjured or not, tail events are mispriced, a predicate for ERM.


18 In addition to uncountable deaths and disease, Japanese nuclear power generation was designated for replacement.


22 ISO 31000, ‘A structured approach to enterprise risk management (ERM) and the requirements of ISO 31000’, 2009; http://theirm.org/documents/SARM_FINAL.pdf. This is among the more definitive ERM guides and puts forward a two-tailed definition of risk that is less readily reconcilable with a sustainability interpretation of ERM. By way of contrast, CARE (International Actuarial Association, ‘Comprehensive actuarial risk evaluation’, 2010) maintains ‘we can think of risk as the potential for an outcome with negative consequences’.

23 Some divide the control infrastructure into three parts: (i) internal controls; (i) risk management; and (iii) crisis management and disaster recovery. But others see items (i) and (iii) as parts of a fully articulated ERM program.


28 Differences among agents may arise from risk aversion embedded in utility functions and this may give rise to a trade in risks not rooted in core competence.


31 Even where risks are shifted to others, ‘basis’ risk typically remains to be managed, but basis risk is typically an order of magnitude smaller than the original risk being shifted. Basis risk arises from the almost universal disparities between the cash flows of the hedging instrument as against the cash flow being hedged.


37 The venality arises not from the violation of the organisation’s code of ethics, but rather from the failure to maintain basic fiduciary standards of care, loyalty and good faith.


Mining, energy and construction companies often have safety committees and companies with notable public exposure may have social responsibility committees. Standing committees dedicated to counterparty risk, finance or investments are common among financial intermediaries.


Deming, WE, Out of the Crisis, 1982, MIT Press.
Chapter 11

Re-engineering risks and the future of finance

Charles S Tapiero
NYU-Polytechnic Institute

Overview

Convergence of financial theory and practice heralded by the seminal and fundamental economic research by Arrow and Debreu in the early 1950s has led in the hands of financial engineers to extraordinary financial innovations in the 1970s and beyond. The theoretical ability to price future assets has led to an explosive growth of financial trading, liquidity and the growth of finance and its citadels. Options and credit derivative markets, and securitisation of non or partially-liquid assets have provided extraordinary opportunities to ‘unearth’ frozen assets for trade and profit.

The 2007–2009 financial crisis has, however, evoked increased awareness that traditional financial dogma is based on assumptions that have become more difficult to justify. Globalisation, the growth of ‘too big to fail’ (TBTF) companies, insider trading, information and power asymmetries, and the growth of regulation have, among other factors, conspired to render the assumption of complete markets to be unsustainable. Deviant behaviours in financial markets, non-transparency of transactions, complexity and dependence on a global scale have created a fragile and contagious global economy, where systemic risks are no longer an exception but a permanent threat.

By the same token, the explosive growth of information and data fuelled by the internet and social media, as well as developments in IT, has created greater dependence of financial systems and institutions on information technology (IT), emphasising information as assets they seek to use for both financial management and competitive advantage. Technology, engineering, and financial trends and developments have combined these factors to yield an extraordinary growth in complexity, regulation and globalisation, providing both new opportunities and risks that undermine the traditional model-based approaches to finance.

The purpose of this chapter is to outline a number of factors that underlie the future of finance and undermine finance’s fundamental theories, globalisation and regulation. Specifically, we emphasise a strategic and multi-polar finance, beset by complexity, chaos and countervailing forces that lead to a multi-agent environment, which is computational and financial data-analytics driven rather than based on simple risk models of uncertainty.

A number of pricing models based on strategic finance and a micro-macro-matching of economic and financial statistics are summarised and their practical implications are discussed. Topics such as ‘big-data’ finance and multi-agent financial modelling are outlined to provide elements for future theoretical and practical developments to be concordant with a changing
economic, financial and technological environment. This chapter is a work in progress, hence its intention is to attract greater attention to some elements (however selective and partial) that could contribute to potential transformations of a financial future.

**Introduction**

What is finance? In a broad sense, ‘finance is a derivative’ of social, political and needs engineered by economic theories, and computational and data driven technologies. Finance seeks to create and manage rationally and efficiently financial markets, liquidity, exchanges and risks based on the basic tenets of economic models. In this sense, finance is not an end, but a means which is needed for economic development and an efficient allocation of individual and society’s resources, dictated by individual and collective wants. It is engineered by financial models that seek to bridge theories and their practice and at the same time comply with financial regulation. Current trends indicate a future that differs from the past, fuelled by numerous major factors including essentially:

- the growth of A-National and globally networked financial citadels, that are TBTF, too big to bear (TBTB) and too big to jail (TBTJ);
- globalisation;
- an e-finance, fully integrated in complex IT networks;
- the growth of a comprehensive regulation and sovereign states concerned by systemic risks;
- the evolutions of the environment, politics and societies and their transformational consequences; and
- default financial models in turbulent and strategic financial times.

Recurrent financial crises have eroded the trust and the belief that financial institutions and markets left to their own will necessarily lead them to be fair and efficient. These beliefs are challenging conventional theories that underlie the tenets of free and complete markets based on assumptions of: (i) anonymity of traders and investors, with predictable future state preferences; and (ii) no arbitrage, with unique asset prices quantified by a pricing martingale, and so on. Rather, economic inequalities and the dominance of few agents, with their power and information asymmetries, have altered the assumptions that underlie fair prices. A mismatch between financial valuation models and macroeconomic variables, dictating market trends, are challenging conventional wisdom embedded in complete (and socially efficient) markets. Greater attention needs to be given to: ‘engineer a sustainable finance and systemic risks’; develop efficient regulations; increase awareness of models’ uncertainties; recognise default models and define their risks and the consequences of their risk externalities; have greater sensitivity to sovereign agendas and so on. These provide ample grounds for speculations on the need for a ‘finance’s future’.

A changing global environment, political realignments and IT are no less important. For example, globalisation, population growth, desertification and the migration of populations due to hunger and the loss of habitat, the supply of natural assets, data and computing technologies, and the complexity of societies increasingly threatened by a tyranny of minorities, and so on, are contributing towards redefining our economic future. A future that is
defined by more natural, macro-social, and economic events, IT, a cult for regulation, and the growth of gargantuan financial citadels and sovereign states inclined to be far more subservient to economic and social agendas.¹

Financial theories are based on risk models that in fact are models of uncertainty. They reduce future events to be known when they need not be known. ‘Theories’ are, therefore, sometimes right and, of course, sometimes wrong. Financial models derived from a ‘statistical’, ‘typical’, ‘averaged’, and ‘a-personal’ financial agent falter when finance is defined by relatively few and unequal agents, all of which may be endowed in wealth, information and power. Big-data, IT assisted ‘sentiments streams information’, algorithmic and high frequency trading, and so on, are expanding the technological advantages of the few agents that wield such capabilities, thereby increasing the information and power of these agents. These are some of the elements of ‘things to come’ in finance. Financial risk models, assets valuation and pricing, financial risks management and financial exchanges may thus be different in the future. They may be transparent on the one hand to some, and stealth to others; and accessible to the many on the one hand, but managed by a few intermediaries on the other. National regulatory agencies have awakened, and are now at the front and back ends of finance. They control adverse economic and social effects of information and power asymmetries, but at the same time contribute to a transformation of finance by the power they have and the way they exercise it. Regulation, while essential, may also lead to the growth of A-National, stealth and virtualised corporate entities, which evade controls and taxation in the pursuits of corporate agendas, and to competing and global regulatory agencies and systems.²,³,⁴,⁵,⁶,⁷,⁸,⁹,¹⁰,¹¹ These elements contribute to the growth and the complexity of financial institutions and exchange markets (whether electronic, global, or not) that are challenging financial theories and its many practices.

These developments contribute to strategic risks fuelled mainly by ‘what we do not know, or choose to ignore’, by ‘what, we and others do’, and by ‘what we do to each other’.¹²,¹³,¹⁴ In this context, the current state of financial analysis and financial risk management may be misleading. At a theoretical level, long standing principles of rational decision processes, which are based on profit-value maximisation, that lead to self-correcting (Pareto efficient) economic processes (presuming that inequality while socially inefficient are in fact economically irrelevant!) have also been questioned. For example, companies are observed to deliver individual returns that are neither efficient in an economic sense nor tied to individual performance. In the words of Herbert Simon, spoils accumulate to agents ‘at the locus of uncertainty absorption’.¹⁵,¹⁶,¹⁷,¹⁸ The strategic position of financial agents (large or small) at ‘loci of uncertainty absorption’ render finance far more strategic.

Fundamental theories of finance have served us well, however, when financial markets are predictable and stable.¹⁹,²⁰,²¹,²²,²³,²⁴ But, they fail during turbulent, complex, rare and unpredictable times. In such conditions, what can finance be? This is a question that we may have no answer for.

**Finance and theories**

Finance theory is based fundamentally on risk models. These are models of uncertainty, and are consequently only hypotheses – sometimes right, sometimes wrong, always in doubt and
never confirmed. As a result, financial models are merely partial models of uncertainty that do not predict or define prices uniquely.

For conventional finance, the ‘predictable future is now’, with the current price reflecting future risks and their price. As a result, theoretical constructs and risk management are based on managing the ‘predictable’ rather than managing the unpredictable and consequential risks. When financial dogma is based on assumptions taken for granted but do not always hold (such as, infinitely liquid markets, balancing those who seek a flight from risk and those who seek it for profit, and so on), booms and busts, systemic risks, contagious behaviours and so on, may recur unpredictably.

The following is a summary of some of its elements underlying the transformation of a future finance.

- Increased complexity and uncertainty.
- Default models due to incomplete financial modelling.
- An environment increasingly untamed by population growth, competitive demands for consumption and economic inequalities, both within and between countries.
- An increasingly strategic finance with dominant agents. In this world, a general equilibrium may no longer be possible, sustainable, or efficient. These require new approaches to financial risk management.
- Financial greed, with TBTF enterprises, information and power asymmetries; increasingly aggressive regulations that stifle financial innovation and cause capital to migrate; and a breakdown of the tenets of free markets.
- The prospects of failing regulations in a global world, contributing to a strategic competition between the regulated and the regulators across national borders.

These processes are changing tomorrow’s economic world, and therefore the future of finance. This may result in extreme behaviours fuelled by excessive unfiltered information, far more apt to generate contagious behaviours and therefore ‘financial runs’. Security risks, networks and IT also emerge as important candidates that will redefine operational risks and resulting in techniques of financial risk management applied ex-ante to be far more preventive, robust and anti-fragile.

These ‘futures’ provide both ‘opportunities and ills’. In their path, the future of financial investments, risk management and trading, liquidity, financial regulation and risk models will necessarily change.

For example, traditional financial tools to price financial assets, focused on ex-ante and predictable risks (whether market or credit risks), may be amended to counter the effects of strategic risks and (systemic) risk externalities in global and competing financial markets. In this environment, a plethora of conventional financial tools, such as Value at Risk (VaR) and market based hedging models for risk transfer, may also be of limited use (albeit necessary to assure the ‘skin in the game’ that financial institutions have to maintain to prevent excessive risk taking behaviours).
Big-data and the future of finance

Big-data is touted by the media as the future, replacing statistics as a model of what is likely to happen, to a future based on what has happened or happenings. While it may eventually be the case, right now, it is challenging. For theoretical finance, financial risk management and credit risks there are obvious possibilities. For example, big-data can be used to better define future state preferences, thereby reducing our future ‘Knightian’ uncertainty. By the same token, big-data can be used to better define aggregate and individual credit risks through greater ‘personalisation’ and detection of risk factors. For regulators it may be used to better define what are the factors contributing to systemic risks on the one hand and detect non-compliance of large financial companies on the other. Large companies, when subject to thousands of regulations requirements (as it is the case with the passing of the Dodd-Frank Act) can use big-data to test their compliance. The promises and the risks of big-data in finance are in their infancy. Some of the issues yet to be addressed are their integration in financial models, the power it provides to the financial institutions that wield it, their implications for individual liberties and security, as much as a fundamental re-evaluation of finance as an IT profession rather than a financial economic and engineering one.

Big-data is essentially a technological derivative of the extraordinary growth of information generated by the internet and social media, fuelling a plethora of data analytics techniques to make some sense of it all. An increasing number of software companies and start-ups are proposing ‘black boxes’ to interpret ‘consumers’ sentiments and intents using internet comments on stocks, on financial assets, and variables deemed pertinent to the financial environment. Algorithms and automatic learning machines are then created to seek and interpret images, to detect a ‘flow of sentiments’ that are claimed to be related to (and thereby be early predictors of) stock markets performance, and predict consumers’ choices and their implied preferences. A rising tide of data driven algorithms based on ‘pseudo-rationalities’ is thus emerging and engulfing large companies and business interests with an avid and information dependent finance. Big-data competes with, and at the same time complements, the traditional ‘statistical approach’ to data measurement and interpretation, providing a renewed interest to data analytic techniques.

Big-data finance may engender self-fulfilling prophecies and ‘hoped for discoveries’. For example, say that a company hires a big-data consulting company to determine the public’s attitude towards that company and its’ CEO? It would be like paying a psychiatrist to hear what one wants to hear. In a big data set the chances of finding what one wants to hear and what may be the facts, are equally high. Searching for meaning in large datasets, without theories, may be like seeking the North Pole without a compass. For these and other reasons, big-data based on the accumulation of private information is a growing source of risk that has already contributed to important security problems. Sadly, big-data has made it possible to steal national and military secrets, to divulge private financial and health information, and so on.

The traditional statistical approach, unlike the IT intensive big-data, is based on fundamental and structured hypotheses emanating from scientific statements or theories, to be refuted or not, based on statistical (data) evidence. As a result, the ‘statistical/
scientific’ approach reveals ‘uncertainty’ from a given and tested knowledge base. Data driven algorithms, on the other hand, are based on a ‘robotised artificial intelligence’ that has the capacity to treat large datasets to reveal co-variations, associations, patterns, and so on, based on their search (in some cases improved by built-in learning algorithms) rather than based on theory. While the statistical/scientific approach is an evolutionary and evolving process, based on a cycle to hypothesise, measure, test, and confirm-or-not, the data driven approach is a short cut, seeking to know even though it may not know why. It is an interpretation of data at a point in time with a predictive artificial intelligence that decision makers might not understand, as data revelations can be as complex and as stealth as artificial intelligence may be to its users. It is a statement of a current fact – a presumed certainty, rather than a recognition that all knowledge is partial – embedded in a greater uncertainty.

Exhibit 11.1

**Statistical versus data analytic approaches**

<table>
<thead>
<tr>
<th>Statistical approach</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Factual</td>
</tr>
<tr>
<td></td>
<td>Complex/stealth</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data analytic approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientific</td>
</tr>
<tr>
<td>Hypotheses</td>
</tr>
<tr>
<td>Experimentations</td>
</tr>
<tr>
<td>Bank of knowledge</td>
</tr>
</tbody>
</table>

*Source: Author’s own*

The British astrophysicist Arthur S Eddington once wrote: ‘No experiment should be believed until it has been confirmed by theory!’ In this sense, data driven measurements can be misleading and reveal anything that one may choose to discover. The digitalisation of the Bible with a ‘robotised artificial intelligence’ has revealed to ‘digital mystiques’ sequences of statements revealing new prophecies to the end of days. To disprove its validity, a digitalisation of Tolstoy’s *War and Peace* and a search engine revealed ‘secret intents’ as well. Other books have done so as well, predicting even political stalemates in the Middle East and their resolution. In this sense, data driven measurement models can make sense out of non-sense even though it is in fact nonsense. Similarly, in biology, where big-data has already found its footing, Werner Callebaut raises important and philosophical issues. Can
data speak for themselves? Can a society that permits biology to become an engineering
discipline, and allows science to slip into the role of changing the living world without
trying to understand it, be a danger to itself? Can biology (as well as finance) become an
IT management profession?

Big-data has its perils, with type II statistical errors (that is, making a false discovery),
enslaving decision makers to rationalities they do not comprehend and provides a new
twist to the misuse of data measurement by seeking and finding the confirmation of
one’s certainty.

Finally, big-data can turn out to be an unwieldy process victimised by the belief that
a larger hay stack may help find a needle in that hay stack. Yet, big-data banks, if tamed,
can complement the ‘statistical/scientific’ approach to measurements by providing an oppor-
tunity to reveal new hypothesis and new directions that can set such approaches on a surer
footing. Scale analysis techniques developed by Louis Guttman have provided approaches to
the measurement and the reduction of data dimensions into a scale (used profusely in the
social sciences and psychology, called the Guttman scale). A summary of his posthumous
contribution summarised in Wikipedia is given below:

Scale analysis, Facet theory and his mathematical and philosophical treatments of Factor
analysis are among the important parts of his scientific legacy. Notably, Guttman first
proved several fundamental theorems in matrix algebra, as discussed in papers by
Hubert, Meulman and Heiser (2000) and Takane and Yanai (2005) (that have allowed
the development of algorithms of matrix dimensions reduction). Several of Guttman’s
contributions have been incorporated into computer packages, among them Smallest
Space Analysis (SSA) and Partial Order Scalogram Analysis (POSAC).

Exhibit 11.2

The Guttman approach

Source: Author’s own
Future

For banks, traders and suppliers of financial information and advice, data and information are becoming primary assets. The Economist reports that between 1990 and 2005 more than 1 billion people worldwide entered the middle class, and by 2013 the amount of data transferred over the internet will reach 667 exabytes annually. According to Cisco the quantity of data continues to grow faster than the ability of the network to carry it. Companies like Amazon’s Web Services, AT&T’s Synaptic Hosting, AppNexus, GoGrid, Rackspace Cloud Hosting, the HP/Yahoo/Intel Cloud Computing Test bed, the IBM/Google and Micro Strategy BI Cloud, are providing various types of clouds services to ease these data storage problems.

Big-data implies not only large volumes of data, but a wide variety of data types and a high velocity of data streams, with internal and external data integrated. While technologies from social media have shown a way to handle vast amounts of unstructured data for analysis, our abilities to translate these complex, diverse and dynamic data sources into workable financial information remains unproven.

At the same time, the world and financial systems have become more digitised, allowing context-specific analyses. Information and/or knowledge extracted from digital records may render financial banks’ jobs easier in accurately diagnosing and detecting risks, and their clients’ propensity to assume risks, and hence improve the overall quality of their services. Similarly, digitised data may prevent cybercrimes more effectively and thus contribute to the increasingly complex systems of financial networks, e-financial markets and an increasing financial retailing dependence. Despite the potential for big-data and financial data analytics, it may be still immature, misunderstood and misused. In terms of security, Michael de Crespigny, CEO at ISF points out that: ‘Only half of organisations surveyed by the ISF are using some form of analytics for fraud prevention, forensics and network traffic analysis, while less than 20% are using it to identify information related to subject matter requests, predict hardware failures, ensure data integrity or check data classification... Few organisations currently recognise the benefits for information security, yet many are already using data analytics to support their core business.’

The business of big-data finance is also bright and challenging, however a number of questions remain unanswered.

- Is big-data looking for a pin in a haystack by adding hay?
- Is big-data the end of financial statistics or a complement to improve financial models?
- Does big-data, based on past facts, negate fundamental modelling of finance futures?
- Is big-data merely another IT data-driven tool to model complexity and justify what we are doing, or, rather, it defines what is and what we ought to do?
- Is big-data the end of privacy?
- Is big-data an algorithmic search model transparent to those who may want to use it?
- Is big-data a means to reveal outcomes a priori sought or does it reveal the unknown and the unexpected?
- Is big-data something new? Or the marketing of well-known data analytics tools up-ended with more information, expanded by new computational hardware?
- Is big-data a means of simplifying complexity or merely a means to ignore it and focus on the many outcomes it can be used to identify?
• Is the growth of data and our ability to deal with this data unsustainable?

**Global future finance challenges**

Sovereign states are increasingly assertive, prone to regulate or over-regulate, leading, therefore, to the migration of finance to where it is hindered least, and contributing to multi-polar financial markets. Although the financial crisis has led some sovereign states to retreat into national agendas, globalisation may already be irreversible. Questions and future challenges remain, however. Is globalisation sustainable? Is globalisation contributing to economic growth in all countries or only to some? Does growth favour the few or the many? Is globalisation contributing to a better world, to greater equality between countries? Is it contributing to greater inequalities within countries, making the rich, richer and leading to the disappearance of a middle economic class? Can the traditional valuation theory capture the complexity of pricing international assets and their exchange? What are the implications of strategic economies to the future?

Emerging facts are pointing to a greater ‘equalisation’ of the global economy and, therefore, the loss of traditional competitive advantages to developed countries. This trend, in its historical perspective, is countering a trend that has accelerated inter-nation inequalities ever since the beginnings of the first industrial revolution. Greater ‘equalisation’ assumes many aspects. It occurs in technology diffusion and in industrial might (through outsourcing), among others. A-National companies seeking greater yield, less regulation, as well as National Champions, have expanded their global might. The process of ‘global equalisation’ has thus been accelerated with open borders and ‘the flow of comparative advantages’ – whether in natural resources, in technology or know-how, or based on financially engineered and repression policies in trade and global finance and taxation. The outcome is a tsunami from West to East.

The predicaments of financial theory and practice are confronted with a world with increased competition for basic materials and commodities, and increased expectations and needs by a growing and awakened population. Greater inequalities within countries, financial contagions, ineffective regulation and technology growth also exist. Greater interventions by sovereign states seeking to augment their revenue results in a gradual loss of their power to manage their own financial affairs and their markets. Increased indebtedness, the growth of A-National companies and inefficient global financial regulation are depreciating nations’ ability to pursue their own economic agenda. These factors and the pursuit of global opportunities by investors, companies and sovereign states are ‘game changes’ in the future of finance. These trends raise new challenges to financial theories and the management of financial risks. A summary of future financial risks due to globalisation includes the following.

• Competing financial markets and economic regions (US, Asia, the Euro Zone, the BRICs, the nascent Latin American pole, Turkey, Indonesia and Africa) with increased interventions of sovereign states, large A-National companies, combined with financial repression (taxation and regulation hindering globalisation).
• A volatile world with currencies ‘at war’ and a retreat from single currency dominance.
• Statistical and contagious financial dependence across national boundaries.

www.rasabourse.com
• A persistent transformation, stealth and empowering information technology.
• The uncertainty pervading a globally rational financial regulation and its consequences.

These risks are amplified by political and social risks, external and natural risks, as well as by the limits of financial models based on assumptions that do not hold in a very complex and dependent global finance. In particular in a global setting:

• heterogeneous investors that differ by their values, their preferences, their sophistication and so on;
• differences in regulation, taxation, trade incentives and repression;
• ‘oligarchic’ financial agents rather than ‘atomic’ ones;
• managed currencies (explicitly and mostly implicitly); and
• political risks and political agendas.

These lead to heterogeneous and strategic financial risk models that invalidate basic financial risk models. There is, however, ample evidence that globalisation has contributed economically to both economic efficiency and to a process of financial and economic equalisation. Such processes are likely to reduce the dominance of financial sovereignty while increasing global wealth albeit distributed unequally.

**Alternative financial risks models**

Current financial risk models are a partial representation of financial and economic uncertainty, neglecting knowingly or unknowingly future and consequential events. These models cannot, therefore, claim nor fully represent the unknown or the neglected. In this sense, models are bounded by what we currently know, by our preferences, by what others know, by what we know about others and what they know about us, and what we intend and do to each other. Traditional financial models, neglecting these elements, and based on predictable (risk) elements and their consequences will falter. Financial risk models are at best a ‘rational subjectivity’ applied to provide a state of mind or a frame of reference to a bounded and tolerable view of future uncertainty. This frame of reference provides finance with an operational framework used to predict prices, trade, invest, speculate and manage risk. These models do not confront their residual uncertainty.

Psychologists and behavioural scientists such as Simon, Kahneman and Tversky, and Kahneman and Lovallo have sought to better understand and account for behaviours that are not predicted by rational models. Numerous studies have advanced the importance of ‘regrets’ and other ex-post judgmental aspects of decision making (rather than just ex-ante aspects) to provide an ex-post perspective to risk assessment and management. Kahneman and Tversky in particular have suggested an approach coined ‘prospect theory’ that questions the underlying assumptions of a universal ‘utilitarian-rational behaviour’.

Such approaches have ushered an ‘interpretation’ of events based on their observations and subjective value by their probabilities and their consequences. This assumes many forms, such as re-normalising probabilities and renormalising their consequences to account for
the important versus the less important, the known versus the unknown, the likely versus
the less likely, and so on. They reflect behavioural and psychological states emanating from
theoretical and empirical studies in the social and human sciences. Re-normalisations can
result in deformed probability distribution functions with parameters estimated based on an
empirical observation of decisions made by decision makers, their risk attitudes and their
behaviour. These test unconfirmed hypotheses, whether simple or complex, and frame a
financial future as a normalised risk. To do so, many approaches and techniques are used.
Fundamental finance, based on defining a (martingale) probability measure, is the most
important and broadly accepted approach. Other approaches may include, parametric ration-
alities embedded in utility functions (such as CCAPM models) or future probabilities based
on the known and its randomisation by applying some rationality.57 These approaches have
basically ‘done away with risk’ by developing models based on both countable and their
consequential future states, priced implicitly by current information and valued by financial
exchange risk models. These models define a ‘price’ when they are used as the exchange
price. In this sense, the ‘traded future’ is priced and embedded in an exchange now, and,
therefore, has neither risk nor uncertainty.

Fundamental competitive economic models58 have defined an economic equilibrium
expanded to future states by assigning a price to both current assets and to their known
future states. The equilibrium price is defined by a price that traders and investors use as their
exchange price (implying that markets are complete). This approach, validated in complete
markets, can mislead investors and traders when its underlying assumptions are violated in
fact. While it may be ‘complete’ in an aggregated financial market, it may be incomplete for
a finite number of individual and endowed agents or in a broad set of conditions where the
assumptions underlying complete markets are violated. In future finance these caveats have
to be confronted first by recognising markets uncertainty, which is in fact and commonly
understood by financial practitioners.

These elements transpire to financial advice, investment policies and financial regulation.
In particular, ever since Markowitz59 financial investments were based on the construction
of extensive and diversified portfolio neglecting the effect of macro-economic factors. An
alternative approach, initiated in the 1950s by John Kelly, suggests instead that investments
are far more like a racehorse, seeking assets geared to perform rather than diversify and then
use partial hedges to avoid a total loss. In a future strategic finance, investments may thus
be based far more on poker and liars pokers, emphasising the strategic aspects of financial
trading, and the dominance of financial companies and national agendas, each holding the
cards that make financial markets function and use them to their advantage.

Traditional finance theories, have assumed that financial markets left to their own devices
supply liquidity. In particular, by using securitisation, options and credit derivatives financial
markets can expand the supply of liquidity. However, ever since the recent financial crisis,
governments have acted as the last resort suppliers of financial liquidity. Whether one or the
other is the future supplier of liquidity is extremely important in predicting a finance future.
For example, in an increasingly regulated environment, banks will return to be retailers, or
‘marketers’, while central bankers become the suppliers of liquidity. In a future finance world,
its promises to be far more the case than it is today. Curtailing financial optional contracts
may only exacerbate this situation.
Fundamentally, current financial market models mismatch micro and macro financial statistics – separating micro-economic considerations from macro ones and negating their underlying effects in micro (pricing) financial models. Such a mismatch leads to financial markets becoming ‘incoherent’. For example, the mortgage-backed securities (MBS) crisis of 2008, was such a mismatch; combining the condition of ‘a home for everyone’, based on low initiation costs and interests, with long run (and unsustainable) individual and systemic risks. Future models in finance may, therefore, be concerned fundamentally with the issues stated above. Below, we consider a case problem to demonstrate how particular issues, if considered, may change an underlying pricing model.

A multi-agents’ strategic pricing model

Below, we outline a model that departs from CCAPM models and considers the effects of agents’ wealth endowment. For simplicity, we avoid all quantitative proofs and summarise some essential results. In a market with predictable states, CCAPM market prices (with inflation) are based on Euler’s equation, stating for an individual investor that:

$$1 = E\left(\tilde{M}_t^k \left(1 + \tilde{R}_t^k \right) \left(1 + \tilde{\eta}_t^k \right) \right), \quad \tilde{M}_t^k = \beta \frac{\partial u^k_t (\tilde{c}_t^k)}{\partial c_t^k} / \frac{\partial c_t^k}{\partial c_t^k}.$$ 

Where \((1 + \tilde{R}_t^k)\) is the return on savings and \(1 + \tilde{\eta}_t^k\) is the consumption price inflation rate and \(\tilde{M}_t^k\) a pricing kernel. If an investor invests in an inflation indexed bond whose rate of return is \(R_{t, ind}^k\), then

$$1 = (1 + R_{t, ind}^k) E\left(\tilde{M}_t^k \right)$$

and therefore, setting:

$$1 = E\left(\tilde{M}_t^k \left(1 + \tilde{R}_t^k \left(1 + R_{t, ind}^k \right) \right) \left(1 + \tilde{\eta}_t^k \left(1 + R_{t, ind}^k \right) \right) \right) = (1 + R_{t, ind}^k) E\left(\tilde{M}_t^k \left(1 + \tilde{R}_t^k \right) \left(1 + \tilde{\eta}_t^k \right) \right)$$

And replacing \((1 + R_{t, ind}^k)\) by \(E\left(\tilde{M}_t^k \right)^{−1}\), the following stochastic discount factor model is obtained:

$$1 = \frac{1}{1 + R_t} E^{\tilde{M}_t^k} \left(1 + \tilde{R}_t^k \right), \quad \text{with probability measure} \quad \frac{\tilde{M}_t^k}{E\left(\tilde{M}_t^k \right)}$$

Under this probability measure, the risk premium is null and the risk free rate equals the future rate of return, or

$$R_t = E^{\tilde{M}_t} \left(\tilde{R}_t \right).$$

With wealth significant agents, prices are a function of agents’ endowments (wealth, regulatory power, information and so on) and prices may not be unique, each price reflecting what the consumer will be willing to pay. For example, would the rich and the poor be willing to pay the same price for the same goods? By the same token, bid-ask prices do not reflect the market price but the prices that prospective investors and equity owners are willing to buy or sell a stock at. In this sense, the price one is willing to pay for and the market price define the premium (of the buyer or the seller). A transaction occurs when both have a positive premium; the buyer willing to buy, the seller willing to sell. Complete market (equilibrium) prices result from the exchanges between various endowed populations of buyers.
and sellers, potentially with different risk preferences, needs and abilities. The presumption that there is a ‘representative’ and a homogenous set of economic agents to define a market price is then, necessarily misleading. If all buyers and sellers of risk, real goods and so on, were all the same, there would be no exchange. Similarly, if pricing future returns were the same for all investors, there would be no future market. A multi-agent (endogenous) pricing model is, therefore, not an alternative to Arrow or Debreu’s pricing framework (which is based on no arbitrage, one price and so on) but a complement to the fundamental CCAPM approach, by defining additional factors that underlie both potential buyers and sellers. Unlike the exogenous (traditional) model, the multi-agent pricing model provides an opportunity to assess the effects of investors’ endowment (wealth), supply factors and consumers’ propensity to save (that is, consume less now and save more for future consumption). In addition, it will reconcile micro-factors defining the consumer and macro-economic factors that define the price of consumption. In these conditions, Tapiero\textsuperscript{64,65} has shown that Euler’s equation in a multi-agent framework is given by:

\[
1=E \left\{ \tilde{M}_k \frac{(1 + \tilde{R}_k)}{1 + \tilde{\eta}(\tilde{C}_i, \tilde{S}_i)} \left[ 1 + \sum_{j=k}^n \frac{p_j(1 + \tilde{R}_j)}{(1 + \tilde{R}_j)} \right] \right\}
\]

where \( p_k \) is an agent k share of the aggregate invested wealth. It is convenient to define all other agents by returns \( 1 + \tilde{R}_i = \sum_{j=k}^n \frac{p_j}{1 - p_k} (1 + \tilde{R}_j) \) in which case, an agent pricing model is reduced to a (John Kelly, racehorse) game between one and all other agents – the kth and a synthetic agent representing all other agents whose rate of return is \( 1 + \tilde{R}_i \). In other words:

\[
1=E \left\{ \tilde{M}_k \frac{(1 + \tilde{R}_k)}{1 + \tilde{\eta}(\tilde{C}_i, \tilde{S}_i)} \left[ 1 + \frac{p_k(1 + \tilde{R}_k)}{(1 - p_k)(1 + \tilde{R}_i)} \right] \right\} = 1=E \left\{ \tilde{M}_k \frac{(1 + \tilde{R}_i)}{1 + \tilde{\eta}(\tilde{C}_i, \tilde{S}_i)} \left[ 1 + \frac{(1 - p_k)(1 + \tilde{R}_i)}{p_k(1 + \tilde{R}_i)} \right] \right\}
\]

This means that financial pricing models are defined by agents-investors that ‘game’ all other agents, summarised by a ‘synthetic agent’. Wealth endowment has an effect on prices while a discount factor may be defined not only with respect to a risk free asset whose price is a priori known but also a function of the synthetic agent’s rate of return – with both returns depending on the portfolio these agents engineer to obtain rates of returns \( \tilde{R}_i \) and \( \tilde{R}_i \), respectively. Their dependence defines a game, where information and power are important. This game is expanded further by agents’ investments portfolios consisting often of common ETFs. These elements, of course, contradict the fundamental assumptions of pricing models that assume that investors returns are statistically independent of one another.
To obtain an implied pricing model, we first note that by definition:

\[
1 = E \left( \tilde{M}_k \left( 1 + \tilde{R}_k \right) \left( 1 + R_f \right) \left( 1 + \tilde{R}_k \right) \left( 1 + R_f \right) \left( 1 + \tilde{R}_k \right) \left( 1 + R_f \right) \right)
\]

Or assuming that the price of an inflation indexed bond is \( 1 + R_f^\text{ind} = \frac{1 + R_f}{1 + \tilde{\eta}(\tilde{C}, \tilde{S})} \)

we have:

\[
1 = \left( \frac{1 + R_f^\text{ind}}{1 + R_f} \right) E \tilde{M}_k \left( 1 + \tilde{R}_k \right) \left( 1 + \frac{p_k \left( 1 + \tilde{R}_k \right)}{\left( 1 - p_k \right) \left( 1 + \tilde{R}_k \right)} \right)
\]

Next assume that agent k invests in a risk free bond, thus:

\[
1 = \left( \frac{1 + R_f^\text{ind}}{1 + R_f} \right) \left( 1 + R_f \right) E \tilde{M}_k \left( 1 + \frac{p_k \left( 1 + R_f \right)}{\left( 1 - p_k \right) \left( 1 + \tilde{R}_k \right)} \right)
\]

which is reduced to a stochastic discount factor pricing model whose discount rate is:

\[
1 = \left( 1 + E^{\tilde{M}_k} \frac{p_k \left( 1 + R_f \right)}{\left( 1 - p_k \right) \left( 1 + \tilde{R}_k \right)} \right)^{-1}
\]

\[
1 = \frac{1 + R_f^\text{ind}}{1 + R_f} \left( 1 + R_f \right) \frac{p_k \left( 1 + \tilde{R}_k \right)}{1 - p_k} E^{\tilde{M}_k} \left( 1 + \tilde{R}_k \right) \left( 1 + \frac{p_k \left( 1 + \tilde{R}_k \right)}{\left( 1 - p_k \right) \left( 1 + \tilde{R}_k \right)} \right)
\]

where the ‘real risk free discount factor’ is now:

\[
1 + R_k^\text{real} = \left( 1 + R_f \right) \left( 1 + R_f \right) \frac{p_k \left( 1 + \tilde{R}_k \right)}{1 - p_k} E^{\tilde{M}_k} \left( 1 + \tilde{R}_k \right) \left( 1 + \frac{p_k \left( 1 + \tilde{R}_k \right)}{\left( 1 - p_k \right) \left( 1 + \tilde{R}_k \right)} \right)
\]

The simultaneous pricing model of the agent and its synthetic counterpart is then:

\[
1 = \left( 1 + R_f \right) \left( 1 + R_f \right) \frac{p_k \left( 1 + \tilde{R}_k \right)}{1 - p_k} E^{\tilde{M}_k} \left( 1 + \tilde{R}_k \right) \left( 1 + \frac{p_k \left( 1 + \tilde{R}_k \right)}{\left( 1 - p_k \right) \left( 1 + \tilde{R}_k \right)} \right)
\]

The intention of this development was to indicate that heterogeneity alters the risk pricing model. Of course, in a market when all agents have neither the wealth nor the power to affect market prices, we obtain the traditional Euler price since for \( p_k \approx 0 \), the pricing model is reduced to \( 1 = \frac{1}{1 + R_f} E^{\tilde{M}_k} \left( 1 + \tilde{R}_k \right) \). These multi-agents pricing models implied in each agent’s

www.rasabourse.com
stochastic discount factor define a game between the n agents. Settling on a market price will, therefore, be based on the behavioural assumptions that both agents adopt when they engineer their investment strategy. For example, it is possible that dominant Nash solution might not exist, that some agents have information that others do not have, that some agents have power that the others do not have (in which case, a Stackleberg solution might provide a price), and so on. For an agent, who knows how others are invested and their wealth, the propensity of other agents to invest in bonds is \( a_{rk} \) and in a risk prone index it is \( (1 - a_{rk}) \). We then have:

\[
1 = \frac{E^{M_k} \left( 1 + \frac{P_k (1 + \tilde{R}^i_k)}{(1 - P_k) (1 + \tilde{R}_{rk}^i (a_{rk}))} \right)}{(1 + R_r) \left[ 1 + \left( 1 + R_r \right) \frac{P_k}{1 - P_k} E^{M_k} \left( \frac{1}{1 + \tilde{R}_{rk}^i (a_{rk})} \right) \right]}
\]

Since \( 1 + \tilde{R}_{rk}^i (a_{rk}) = \sum_{j=k}^{n} \frac{P_j}{1 - P_k} (1 + \tilde{R}^i_j (a_j)) = \sum_{j=k}^{n} \frac{P_j}{1 - P_k} a_j (1 + R_r) + (1 - a_j) (1 + \tilde{R}^i) \) we obtain:

\[
\tilde{R}_i^i - a_{rk} (\tilde{R}_i^i - R_r) = \sum_{j=k}^{n} \frac{P_j}{1 - P_k} \left[ a_j (1 + R_r) + (1 - a_j) (1 + \tilde{R}^i) \right] \quad \text{and thereby}
\]

\[
a_{rk} = \frac{\tilde{R}_i^i - \sum_{j=k}^{n} \frac{P_j}{1 - P_k} \left( \tilde{R}_i^i - a_j (\tilde{R}_i^i - R_r) \right)}{(\tilde{R}_i^i - R_r)} = \sum_{j=k}^{n} \frac{P_j}{1 - P_k} a_j + \frac{1 - \sum_{j=k}^{n} \frac{P_j}{1 - P_k}}{(\tilde{R}_i^i - R_r)}
\]

which requires that the kth agent knows both the wealth distribution of other agents and their propensity to invest in bonds. The traditional assumption that the market can be represented by a single synthetic ‘averaged’ agent is in this case not meaningful. If all agents essentially trade in a common financial market, agents’ portfolios will also be statistically dependent, negating, therefore, the validity of the simple Euler pricing equation. The implications of a strategic multi-agent finance pricing model thus differ from the assumption of Euler’s price equation. A number of conclusions are summarised below.

- An investor with negligible wealth will be willing to pay $1 for a future rate of return while an investor which has considerable wealth will be willing to pay less.
- The price of assets in a multi-agents framework recognises the statistical dependence of rates of returns that are demanded and obtained by the agents and thus augments the investment risk. Explicitly, if financial portfolios consist of investment in a finite number of exchange-traded funds (ETFs), whether they are statistically dependent or not (which they mostly are), then buying and selling ETFs or their assets induces an inter-agents statistical dependence which is not recognised by the standard pricing models.
- Equilibrium in a multi-agent model is defined by solution of a game with n players, each player investing to maximise their returns. However, if returns are a function of aggregate investments competing for these returns, then necessarily for a single risk prone asset an
Future

Investor rate of return is a function of its wealth share invested in the risk free and the risk prone asset, as well as aggregate investments in that asset, or other assets. If we were to analyse this game then prices would consist of randomised returns which are not practical.

- The number of agent-investors (and thus the market size) matters as it affects both consumption prices as well as the economics of supply (price and so on). Of course for a fixed supply, the greater the demand, the more prices increase and thus the greater the inflation rate. Inversely, the greater the supply capacity the smaller the inflation rate. Markets size thus combined with supply capacity provide an economic model to inflation.

As a result, the future price is a function of consumers’ wealth, their investment strategy tailored to provide a random rate of return \( R_k \), as well as their decision to save, which is necessarily a function of current prices and their current need.66

Finance and future risk externalities

‘The lure for size’ is embedded in ‘economies of scale’ and the ‘power’ it provides to those who wield it. It also leads to a strategic and politicised finance and, therefore, to a departure from complete markets finance. This is the case for financial giants that have grown into networked, complex and global enterprises that have accumulated default risks too large to bear and unleash systemic risks. Banks, unlike industrial enterprises, draw their legal rights from a common trust, to manage the supply and the management of money for their own and the common good. Consequences of their failure when they are TBTF are systemic risk externalities when losses far outstrip their direct losses. For example, Rolfe Winkler (Reuters) pointed out that ‘Main Street owns much of the risk while Wall Street gets all the profits’. Given the importance of the financial system, risk externalities and their systemic risks may become a far greater focus of finance research and regulation.

The growth of companies fed by debt seems to have increased default risk significantly, threatening both creditors and borrowers alike. In fact, the growth of size through a growth of indebtedness combined with TBTF risk attitudes has contributed to moral hazard risks, with companies assuming non-sustainable growth strategies with stealth and important risk externalities.67 When size is embedded in networking, such as large global networked financial institutions, their failure may be contagious.68,69,70,71

Can economies of scale savings compensate TBTF risks? Such an issue has been implicitly recognised by Obama administration’s proposal in Congressional committees calling for banks to hold more capital with which to absorb losses. The bigger the bank, the higher the capital requirement should be.72 However, such regulation does not protect the public from the risk externalities that banks create and the ‘public’ sustains. Augmenting inexorably the capital banks must hold will necessarily reduce the ability of banks to lend and augment counter cyclical risks.

Conclusion

The future of finance is challenged by a changing and global environment, by the growing cult for regulation, by networked and stealth A-National financial enterprises, by complexity
and multiple sources of dependent and contagious risks. These create an environment where endogenous systemic risks may be far more common. The size of fewer financial entities and the power and the information they possess will alter (and already has) their comparative advantage and the nature of competitive markets. A multi-polar financial and currencies world may create an environment where financial tools can neither be transparent nor fair to financial investors outside the realm of few financial citadels. In this environment, ‘mega enterprises’ will be the dominant market makers, insurers and price setters, with investors merely the actors on a platform, manipulated to dance, sing and move at the whims, the will and the tune of gargantuan financial entities.

In this environment, risks are necessarily less predictable, at times unavoidable. The current tools of finance are challenged by future developments as they often fail to confront fundamental issues such as:

1. neglecting finance of the unlikely (explicitly and implicitly);
2. a managed (and thus far more manipulated) financial liquidity, available to some and not to others, and thereby mispriced providing systemic arbitrage to some but not to others;
3. necessity of ex-post and a robust approach to financial risk management, recovery and the ability to learn and overcome financial losses;
4. the expanded growth of financial IT risks and an urgent need to mitigate them; and
5. would providing safety nets for TBTF financial institutions that produce risks TTBTF and risk externalities adequate?

Ever since the financial crisis of 2007–2009, numerous meetings, papers and books have sought to provide some insights and reasons for the failure of financial markets and provide guidelines for a systemic risk free future. These lessons are mostly a repeat of past events that were not learned. To prevent recurrent market failures and financial inequities, regulation is called to the rescue. Regulation is a two-edged sword however. It mitigates public and systemic financial risks at a cost to the regulator and the regulated. It alters economic and financial systems with consequences that have not been fully appreciated. Risks arise due to migration to evade compliance to states’ regulations. Complex regulation (as it is increasingly the case) hampers companies’ profitability and increases the cost of compliance. It also contributes to an environment with a prior belief that all agents are guilty. Detecting compliance consists then in blanket regulation of all agents to find who is guilty and what one may be guilty of. Regulation is faced with extraordinary challenges, beset by the basic dilemma to ‘regulate more or less’. Should regulation be complex? If so, how complex? Should it be based on simple principles? And if so, how simple? These have implications that are difficult to assess as their effects have important direct and indirect consequences. Some relevant concerns include the following.

- Would the cult of regulation incite companies to further consolidate and lead to centralised TBTF companies, constrained and unable to create liquidity?
- Would banks unable to trade in financial derivatives be unwilling to provide liquidity to small and medium-sized businesses, but provide cheap credit to large corporations (increasingly global and evading many forms of national regulation)?
• What are the effects of an excessive regulated and financial economy on the creation of real jobs?
• Would regulation create mega-corporations that transform economic markets to be non-competitive?

These risks, though they may be unintended consequences, may have important implications for the future of finance, as they affect the economy and society as a whole while transforming financial agents and restructuring financial markets. Attempts to increase international regulation have currently set norms to reduce risks, but at the same time these norms can induce counter-cyclicality risks and markets dislocations. Increasingly, it is recognised that such regulation is effective in ‘normal times’ and ineffective when they are needed in abnormal times.

Financial engineers have revealed the inadequacy of many of their own methods, calling into question the efficiency of financial instruments, financial markets, or both. A theoretical backlash against fundamental and quant theories is however misplaced. It expresses investors’ and traders’ frustrations that seek refuge in the certainties of finance. When, in fact, there is no certainty but uncertainty and risk models. The ability to innovate, adapt and provide the means to fulfill its essential purposes is and will be the future of finance. It may do so by reconciling coherent financial theories with financial practice. Finance as an end has proved to be unsustainable. Finance as a means is needed for economic development, growth and the efficient allocation of society’s resources.

Bibliography
Markowitz, H, Portfolio Selection – efficient diversification of investments, 1959, Wiley.
The Alfred P Sloan Foundation is gratefully acknowledged for the financial support and grant on Systemic Risks and Financial Regulation.


7 Peter, EE, Chaos and Order in Capital Markets, 1995, Wiley.


17 Simon, HA, Models of Man, 1979, Yale University Press.


43 Simon, HA, Models of Man, 1979, Yale University Press.


Future

Chapter 12

The future of financial regulation

Patricia Jackson and Stefan Walter

Overview

This chapter reviews the future regulation of international banks and over-the-counter (OTC) derivatives markets as a result of the financial crisis and considers the implications for banking institutions and financial markets. The Basel III agreement is radical in its scope and breadth, introducing significantly higher capital requirements and buffers, a leverage ratio, and for the first time liquidity requirements to be applied internationally. In addition, the Financial Stability Board (FSB) has required that banks become more resolvable to reduce risks to taxpayers and the broader financial system. The OTC reforms are equally broad and far reaching encompassing the use of central counterparties, bilateral margining for non-centrally cleared OTC contracts, and new reporting and dealing platform requirements. But many other changes are now being contemplated, ranging from a potential fundamental rethink of the Basel risk weighting regime to a growing number of locally driven structural reforms. The article argues that the extent of change under the G20 reform package will require major adjustment to banks’ business models and financial markets and that the industry is still at the beginning of this process. The addition of possible further wide ranging reforms creates a degree of uncertainty for the industry which will increase the adjustment costs. A period of stability is now needed to enable the banks and markets to adapt at a time of continued economic fragility.

Introduction

This chapter focuses on the future of financial sector regulation and looks at how global regulation of international, and in many countries also domestic, banks is changing post crisis, as well as the other area of fundamental reform covering the OTC derivative markets. Both will change the shape of financial market activity in a profound way going forward and each is covered by a separate section which looks at the nature and implications of the reforms.

The severity of the financial crisis challenged the status quo of international bank regulation, which had been heavily focused on capital as the prime prudential buffer, but also had allowed various forms of non-equity capital to count as part of that buffer. Regulation of liquidity was underdeveloped. Moreover, supervision tended to focus at the level of individual institutions and did not consider the build-up of risk across the banking sector or across the whole financial system. Finally, there had been limited focus on putting in place
mechanisms to resolve large, interconnected financial institutions in a manner that does not put taxpayers at risk or causes contagion across the financial system.

Basel III, brought in after the crisis, has rebalanced prudential regulation to focus also on liquidity and leverage and has increased both the magnitude of the capital buffers and the importance of common equity capital in the regime. The FSB has also focused on the need for more intensive supervision (with a particular focus on risk governance and stress testing) for global systemically important companies. Another area of focus is the resolution of banks to limit systemic risk and the likelihood of public funds being used in a bail-out.

The failure of Lehman Brothers and AIG drew policy-makers’ attention to the size of counterparty exposures among the major financial institutions and to the buildup of large opaque positions in OTC derivatives. This in turn led to calls to tighten capital requirements on counterparty risk and to change the OTC markets to make them more transparent. Regulators are putting in place incentives to promote the use of central counterparties (CCPs) through higher capital requirements on OTC derivatives and through the introduction of bilateral margining between counterparties of uncleared OTC derivative transactions.

Looking to the future, it will be critical that the globally agreed reforms are implemented in a consistent manner across G20 countries to avoid distortions, arbitrage opportunities, and a race to the bottom. The Basel Committee on Banking Supervision (BCBS) and FSB peer review process is an important step to promote such comparability, but there are already signs that countries and regions are beginning to deviate from the global agreements.

An important question is the effect that the combined package of reforms, probably the most far reaching put in place, will have on the banking industry and the structure of financial markets. This is a combination of the effects of the reforms agreed to date and the uncertainty regarding future regulation, as major policy proposals continue to be debated in areas as fundamental as the treatment of risk weighted assets, the degree of permissible banking sector maturity transformation, and the amount of additional restructuring to be required of systemically important financial institutions. A key concern is that new reforms are being added to the original G20 package, increasingly driven at the local level, before the effect of the original reform package can be assessed. This process of fragmentation threatens the coherence of the original G20 agreements and its consistent application around the globe. It also increases the ‘noise to signal’ ratio around the end state to be achieved while financial institutions seek to implement the reforms and as they transform their business models at a time of continued financial and economic fragility.

Key drivers of the financial reform agenda

The financial reform agenda is a response to what are considered to be the main lessons of the financial crisis. The reforms can be characterised according to a number of dimensions. From a timing perspective the reform agenda can be viewed as evolving with the depth of the crisis, which can be characterised as playing out in three broad phases: (i) the market value and liquidity phase; (ii) the solvency phase; and (iii) sovereign risk and bank resolvability
The future of financial regulation

Another way to think about the financial reform agenda is according to efforts focused at reducing the probability of failure versus those focused at limiting the impact of failure of global financial institutions. A third dimension is the addition of macro-prudential (or system-wide) supervision to the traditional micro-prudential (or institution specific) approach. All three dimensions are addressed in the sections that follow.

Phase 1: market value and liquidity

The first phase of the crisis, beginning in the summer of 2007 and ending right before the collapse of Lehman Brothers, can be characterised as the market value and liquidity stage. It began with the peaking of the US mortgage cycle and the subsequent drop in value of assets linked to mortgage related products. The uncertainty and lack of transparency about which financial institutions were most exposed to mortgage related assets, especially asset-backed securities, resulted in counterparties becoming reluctant to lend to each other in the interbank market except in terms of overnight exposures. Central banks had to step in and provide unprecedented amounts of liquidity to the banking sector and financial markets more broadly. A number of banks with too much exposure to structured mortgage products began to experience sharp declines in the mark-to-market values of these products, with associated losses and declines in capital levels. Other factors came into play in some other cases, for example, heavy reliance on securitisation issuance by Northern Rock.

One underlying regulatory issue, exacerbating the fragility of the banking sector during this phase of the crisis, was that many banks held mortgage-backed securities and other structured products in their trading books, as well as loans being warehoused to be placed into securitisation vehicles and other illiquid positions. For the major banks the capital for these trading book positions was calculated using Value-at-Risk (VaR) models which substantially understated the credit risk and price risk associated with these illiquid mortgage-related securities, especially when the methodologies used were based on external ratings. One of the early changes post crisis was to alter the treatment of position risk under the so-called Basel 2.5 trading book reforms, which substantially increased the capital required for such securitisation exposures. These changes are likely to go further with the fundamental trading book review under consideration by the Basel Committee, which is expected to make it harder or more capital expensive to include illiquid positions in the trading book capital calculations.

Another important factor behind the scale of the crash was the size of the correlated positions that banks were holding and treating as separate exposures. Different vehicles were underpinned by similar securities but were not treated as a single exposure. This is leading to a global tightening in large exposures rules, which go beyond traditional lending exposures to incorporate a range of additional on and off-balance sheet positions, as well as groupings of economically interconnected exposures.

Once the solvency questions arose at banks with too much exposure to structured mortgage markets, they succumbed quickly because of the thinness of their liquidity buffers which had been optimised over time. Indeed many of the large holdings of retail mortgage-backed securities (RMBs) were in the treasury operations of a number of banks as a way to boost treasury returns. This shone a spotlight on the absence of an international agreement on bank liquidity which has subsequently been remedied through the Basel III agreements.
Phase 2: solvency and counterparty exposures

The second phase of the crisis began with the collapse of Lehman Brothers and can be characterised by the shift from a liquidity and market value crisis to a systemic solvency crisis affecting large parts of the banking and financial sector. Stabilising the financial system required unprecedented amounts of official sector support, including capital injections, guarantees and liquidity both at the individual institution level and for markets more broadly. Weaknesses in the underlying economy, especially housing and the financial sector, reinforced each other, requiring unprecedented amounts of fiscal and monetary stimulus, as well as special lending facilities to support the financial sector.4

Although capital was at the heart of the international approach to prudential supervision of banks, the basic framework had been established in the original 1988 Accord, which was built on the basis of a distinction between core Tier-1 capital and additional Tier-2 capital, the latter including subordinated debt. Tier-1 capital was to comprise common shares and retained earnings but was weakened over the subsequent years through the permissibility of an increasing amount of debt-like (hybrid) instruments in the core capital base. The crisis revealed that these forms of non-equity capital did not absorb losses on a going concern basis. In many cases they were not even required to absorb losses in a resolution scenario. Of the 8% combined Tier-1 and Tier-2 minimum capital to risk-weighted asset requirement, only 2 percentage points needed to be comprised of common equity. A key response of the Basel III reforms was, therefore, to concentrate regulation on Core Tier-1 capital in the form of common equity, gradually phasing out debt-like capital instruments from the capital framework and substantially raising the level of high quality common equity capital. In addition, deductions, for example for goodwill, now largely have to be taken out of common equity.

In addition to raising the minimum common equity requirement, Basel III addressed another fundamental weakness of the prevailing capital regulation, which was that the minimum buffer could never be utilised to prevent failure, because the act of using it (that is, dipping through the minimum) would mean that a bank would lose its licence. An important innovation was the introduction of new buffers designed to build up capital in good times that could be drawn down when a bank or the banking system came under pressure, albeit subject to increasing restrictions on capital distribution such as dividends and share buy-backs.5

The ripple effects through the markets after the failure of Lehman also led to a focus on the size of the counterparty risk charges in place. Basel III moved to substantially increase these and to ensure they better reflected the risks associated with the deterioration in a counterparty’s credit quality, as occurred with AIG and the monoline credit guarantors.

The failure of banks to identify key company-wide risk concentrations also resulted in questions by the authorities and by boards about the strength of risk management and governance practices. The observed shortcomings led to an enhanced focus in the Basel Committee and the FSB on recommendations to strengthen risk governance and risk management practices and the associated supervisory review process of these company-wide risk oversight capabilities. The industry also issued a range of recommendations to strengthen risk management and governance practices.6

While there are a wide range of initiatives that comprise the G20 response to the financial crisis, the effort to raise the level and quality of capital in the banking sector is arguably
the most important of these reforms, as it reduces the leverage in the core of the financial system, making banks more resilient to future periods of stress, the source of which will always be difficult to predict. More capital also increases the chance of an orderly resolution and reduces the risk of loss in confidence on the funding side. At the same time, it will be critical to ensure that these reforms do not cause credit intermediation to shift to the so-called unregulated or loosely regulated shadow banking sector. The initiatives of the FSB to develop an appropriate monitoring framework and policy response to address shadow banking risks will be particularly important to financial stability in the future.7

Phase 3: sovereign risk and bank resolvability

The third stage of the crisis can be characterised as the sovereign phase, and it is the direct consequence of the efforts in Phase 2 to support the financial system and the subsequent effects of a deteriorating economic environment in many advanced economies. This phase of the crisis, which is ongoing, took on various degrees of intensity across G20 member countries. The most severe manifestation has been in Europe, where weaknesses in many national banking sectors and high levels of sovereign debt are causing a negative feedback loop between the public and financial sectors. This phase of the crisis has raised serious questions in some countries about the capacity of the public sector to support the financial sector, and even where such capacity exists the willingness of any country to commit taxpayer resources to support systemically important institutions.

Exhibit 12.1 illustrates the changing position of the public sector following the financial crisis. From 2007 to 2012, public sector debt-to-GDP ratios grew at unprecedented rates (at least in peacetime) in many of the G20 countries most affected by the crisis, most notably in the US, the EU and Japan. By contrast, public debt levels remained quite stable in the largest emerging market economies over the course of the crisis.

In the case of Europe, these trends, and their deleterious impact on the banking and financial sectors, have resulted in the recent initiative to build a banking union for the Eurozone. The banking union plans include the creation of a single supervisory mechanism (SSM) under the European Central Bank, the establishment of a European Resolution Authority, and the creation of a European wide deposit insurance scheme. While the EU is progressing most quickly with the SSM, a European-wide resolution regime will also be necessary to stabilise the European banking sector and ultimately to scale back its vulnerability to developments on the sovereign side.

Given the constraints on the public sector to support financial institutions, and the associated moral hazard resulting from any such support, there is now a growing focus on limiting the distress that global and domestic financial institutions can pose for the economy in periods of crisis. As a result, there is a consensus among policy-makers focused on making financial institutions more resolvable in a way that limits disruption to the economy and does not expose taxpayers to losses. These concerns are producing efforts to strengthen national and regional resolution regimes, and to make banks more resolvable through so-called living wills. There is also a growing focus on using bail-in as a way to recapitalise a failed banking institution without resorting to public sector bail-outs.
Future

Exhibit 12.1

The changing position of the public sector following the financial crisis

![Chart showing the changing position of the public sector following the financial crisis]

Gross debt percentage of GDP.

Source: Trading Economics

Capital regulation

Agreed changes

A core response to the crisis is the shoring up of bank capital to reduce the probability of bank failures, with an additional cushion for globally and domestically systemic financial institutions.

The new Basel III regime introduces higher minimum capital requirements, the capital conservation buffer, the countercyclical buffer, and the surcharge for domestic and global systemic banks which are in the process of being phased in between 2013 and 2019 across the G20 countries. The design of the new requirements is geared towards encouraging banks to build up capital in good times that can be drawn down when needed. This is an innovative approach. To achieve this, the Committee has put in place both a higher minimum requirement and the capital conservation buffer which must be held above the minimum. The capital conservation buffer can, however, be utilised (subject to certain penalties) in difficult times for the bank. The penalty is that depending on where the bank is in the capital conservation range, there are restrictions on payout of earnings through bonuses, share buy backs, or dividends to force capital accumulation. Exhibit 12.2 shows the phase-in of the requirements out to 2019, with the amount of Core Tier-1 capital (equity and retained earnings) steadily
The future of financial regulation

Rising from the current 2% minimum to a minimum of 4.5%, with a further 2.5% required for the capital conservation buffer. Over and above this, national authorities can require a countercyclical buffer for overheating credit markets. Global systemically important banks (GSIBs) will have to carry a further capital surcharge ranging from 1% to 2.5%, depending on their size, complexity, interconnectedness and substitutability.8

The rise in capital requirements is substantial. Minimum common equity increases from 2% of a bank’s risk-weighted assets to 7% allowing for the new capital conservation buffer. If one also factors in the additional stringency resulting from the deductions from capital, this represents up to a seven times increase in the effective minimum capital requirement. For global systemically important financial institutions (G-SIFIs), the requirement is significantly greater due to the capital surcharge and tougher capital requirements applied to capital market activities. Moreover, should there be a credit boom, the capital requirement on particular lending could rise even further.

Although banks have been given a phased in approach to building up capital requirements in the Accord itself, they have been forced by a mixture of supervisory pressure in some regions and market pressure across the board to increase capital faster. The 2013 EY and Institute for International Finance survey of risk management shows that by the end of 2012, banks surveyed in the US and Europe had increased Tier-1 common equity to risk weighted asset ratios from an average of 6% pre-crisis to a current average of 10%.9 These

---

**Exhibit 12.2**

**Capital timeline**

<table>
<thead>
<tr>
<th>Year</th>
<th>Countercyclical buffer</th>
<th>Capital conservation buffer</th>
<th>Tier-1 capital</th>
<th>CET-1 capital</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0%–0.625%</td>
<td>0.625%</td>
<td>4%</td>
<td>2%</td>
</tr>
<tr>
<td>2013</td>
<td>0%–1.25%</td>
<td>1.25%</td>
<td>4.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>2014</td>
<td>0%–1.875%</td>
<td>1.875%</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>2015</td>
<td>0%–2.5%</td>
<td>2.5%</td>
<td>4.5%</td>
<td>4%</td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td>4.5%</td>
<td>4%</td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
<td>4.5%</td>
<td>4%</td>
</tr>
<tr>
<td>2018</td>
<td></td>
<td></td>
<td>4.5%</td>
<td>4%</td>
</tr>
<tr>
<td>2019</td>
<td></td>
<td></td>
<td>4.5%</td>
<td>4%</td>
</tr>
</tbody>
</table>

**Source:** Authors’ own
ratios do not take full account of the new deductions from capital, such as goodwill and deferred tax assets, under a fully loaded Basel III approach.

Exhibit 12.3 shows the relationship between the different capital buffers and the economic cycle. Banks will have to maintain a minimum of 4.5% common equity at all times and 7% (including the conservation buffer) in good times. The countercyclical capital buffer could also come into play in boom conditions. Moreover, global systemic banks will face an additional capital surcharge.

Exhibit 12.3

**Capital buffer**

The Committee is also working to finalise the leverage ratio, which is to serve as a backstop to the risk-based capital requirements. The leverage ratio is calibrated at 3% of risk weighted assets using the tougher Basel III definition of tier-1 capital. While 3% may seem low compared to the traditional 4% leverage ratio employed in the US, the proposed BCBS definition of what is included in the asset base of the leverage ratio is much more comprehensive, including not only the accounting balance sheet, but also off-balance sheet commitments and the potential future credit exposure of derivatives (calculated as a percentage of the national amount), and the notional exposure of written credit derivatives. The definition of assets in
the leverage ratio will be considered further over the course of 2013, including the overall calibration and the treatment of netting for derivatives and securities financing transactions, as well as the recognition of collateral. Indeed, in their recent leverage ratio proposal, the US agencies estimate that the 3% BCBS leverage ratio would translate to 4.3% under the current US definition. Moreover, the recent BCBS proposals to make the leverage ratio definition of assets even tougher would result in an even higher US equivalent ratio.

**Future direction**

While Basel III reformed the definition of capital, raised the level of capital (including through capital buffers), made certain improvements to the risk sensitivity of risk weighted assets, especially for trading and derivatives exposures, and introduced a leverage ratio as a backstop, it largely left the Basel II risk weighting regime intact (the main changes to risk weighted assets covered the trading book and counterparty credit exposures). Over the coming years, we can expect to see a significant debate about the direction of regulatory capital risk weights, which underpin the overall capital ratio of banks.

Since the introduction of a global Basel capital framework with the Basel I Accord in 1988, the debate around capital regulation has been about the appropriate balance between simplicity and risk sensitivity. Basel I was a very simple framework barely differentiating between the riskiness of different private sector exposures – for example, an exposure to a leading corporate and an exposure to the highest risk startup business all carried the same 8% capital requirement. This led to substantial regulatory arbitrage, particularly in the US, and helped to drive the growth of the securitisation industry. The extent of this was highlighted by a study by the BCBS of the effects of Basel 1 and led to the search for a mechanism to risk adjust the capital requirements. Two approaches were finally offered under Basel II: (i) use of the standardised approach based on external ratings; and (ii) the internal ratings based approach (IRB), where banks set the probability of default, loss given default, and exposure at default for different borrowers/exposures and a supervisor formula then set the capital requirement using these inputs. Basel II was just being implemented as the crisis started to unfold with standardised and foundation approaches implemented January 2007 and the advanced IRB approach implemented January 2008.

Following the crisis, with banks under pressure to increase capital ratios, the industry has been working to raise capital ratios through a combination of earnings retention, share issuance where possible, and the reduction of risk weighted assets (RWAs). RWA reduction has been achieved through a mix of actions. Banks have been exiting high capital utilising business lines, selling higher capital exposures, and taking more collateral, and in some cases buying credit protection. There has also been a general sorting out of a range of data issues which banks faced in 2008 that necessitated higher buffers in some IRB models. This has, however, led to confusion outside the banks on how the RWA figures have been reduced and concerns on the part of some elements of the official sector and the analyst community regarding gaming. This has in turn led to various studies of the comparability of RWAs across banks. The first study to be published is the BCBS review of the market risk RWAs using VaR models. This study shows considerable differences across banks, but a quarter is due to different supervisory approaches to multipliers for the raw VaR numbers in different
jurisdictions and much of the remainder reflects different decisions in banks regarding the length of historical data used to calculate the VaR. The BCBS issued a second study on the banking book IRB RWAs. The overall conclusion of this study is that there is significantly less unexplained variability in banking book RWAs than was observed in the trading book. It is likely that harmonisation of a few key parameters, such as the trading book historical look-back period, requiring through-the-cycle probability of default (PD) estimates, and providing additional guidance for the treatment of low default portfolios, as well as greater consistency of supervisory approval processes, would address most of the RWA variability issues. In addition, enhancements and greater consistency in Pillar 3 disclosure practices, building on the recommendations of the Enhanced Disclosure Task Force, would further improve transparency around any remaining RWA variability.

One response by some supervisors has been to put floors into the calculations for some portfolios and others have required banks to remove hard to model portfolios from the IRB approach. In addition, the adoption of the leverage ratio is seen as an important backstop to the risk based requirements – biting if RWAs fall too low. Some countries are also looking at publication of the standardised calculations alongside the IRB as a comparator.

This debate has also led to demands from some quarters to replace the risk-based framework with a leverage ratio on the grounds that it would be simpler and more comparable. This does beg the question: comparability of what? As was demonstrated by the Basel I calculations, the Basel ratio, under a very simple leverage type requirement, means very different things for different banks depending on the underlying riskiness of their portfolios. So while the ratios can be made to appear more comparable through ever more simple regulatory imposed ratios, these simpler regulatory measures would mask an increasing variability of the actual risk profile across the banking industry. It also encourages regulatory arbitrage enabling banks to sell higher quality assets, increasing the overall riskiness of the portfolios, while apparently retaining the same capital ratio as a lower risk competitor.

It is not yet clear how this debate will play out. Ultimately there is a need for a balanced approach that retains appropriate risk sensitivity combined with several important elements: (i) sufficient transparency and comparability in disclosures to be able to understand the differences in RWAs across banks; (ii) rigorous supervisory approvals of models; and (iii) a leverage ratio that serves as a back stop to individual institutions and the banking system, but that is not binding for banks that measure the risk in a prudent manner. Relying on just the leverage ratio as the regulatory minimum, with the RWA approach employed through a supervisory backstop, would result in a lack of transparency and comparability about the actual risk profile of the institution. The leverage ratio alone would show capital to total assets not the relationship of capital to risks. The BCBS recently issued a discussion paper to help shape decisions about the appropriate balance between risk sensitivity, simplicity and comparability when considering further reforms to the capital framework.

There is also a renewed focus on the appropriate boundary between the trading and the banking book since, during the crisis, many illiquid products were warehoused in the trading book without appropriate measurement of the risk under the prevailing VaR approach, which assumes capacity to sell or hedge an exposure within 24 hours. In particular, there is a need to define more clearly which types of exposure lend themselves to active management in the trading book, and the degree to which banks have the capability to manage the risks given
The future of financial regulation

the depth of the markets and the associated risk factors to hedge the exposures. A focus on the trading/banking book boundary, with limitations on the positions that can go into the trading book, is an important complement to higher trading book capital, which, if excessive in relation to the prudently measured risks, could create other distortions.

Another question relates to the treatment of external ratings and sovereign risk weights under the Basel III regime, two issues which are closely related. Currently, various components of the Basel capital framework rely on the use of external credit ratings as the basis for setting capital requirements. Most notably, the Basel standardised approach relies on external credit ratings as the basis for differentiating between the riskiness of corporations, banks and sovereigns in the capital requirements. External ratings also are embedded into the securitisation framework and there are a range of other areas where they play a role within the regulatory capital framework. The Basel Committee has initiated certain steps to scale back the excessive reliance on external ratings, for example by proposing a non-external ratings-based alternative to the securitisation framework.17

The Committee, as part of a broader FSB review of external ratings, is taking a more fundamental look at reliance on such ratings. It is not clear how far this review will go in scaling back the current reliance on external ratings. One way forward is to take a menu-based approach (as proposed for securitisations) to accommodate the US, which has adopted legislation fully banning the use of external ratings in its regulatory framework. However, the alternatives to external ratings can be quite crude. For example, in the case of sovereign exposures, the US approach relies on OECD country risk weightings, which were developed for a different purpose, namely currency convertibility risk. In the case of corporate exposures in the US, there has been a return to a flat risk weight, increasing concerns about capital arbitrage that existed under the original Basel I framework.18

The issue of external ratings also comes into play when considering the treatment of sovereign exposures under the capital framework. Under the standardised approach, sovereigns can receive a zero percent risk weight if they are highly rated by external credit rating agencies. However, under the Basel II framework for credit risk, banks that adopt the internal measurement approach would be required to develop internal ratings frameworks to assess sovereign exposures, scaling back such reliance on external credit ratings agencies. The Basel Committee has announced that in the future it will conduct a more fundamental review of how sovereigns should be treated under the regulatory framework, with potentially profound implications for banks’ capital requirements. Sovereigns too could be impacted by the imposition of higher bank capital requirements and large exposure limits on sovereigns although banks in effect have to hold government bonds as part of the liquid assets regimes.

One new trend is the increased focus on the use of stress-testing to set the effective capital required of banks, in particular the buffers that banks will need to hold above the regulatory minimum requirements. This was always the case in regimes that actively applied Pillar II of Basel II, which required an overarching risk assessment to be made and capital to be held against the risks if higher than the Pillar I minimum capital requirements. Part of this assessment was a stress-test conducted by the bank. However, stress-tests have now moved to centre stage in some jurisdictions. In the US (and to a significant extent for EU banks as well), regulatory stress-testing is now in effect determining the amount of capital that banks have to hold factoring in system-wide economic and financial risks. While the US
Future

has incorporated the countercyclical buffer into its final Basel III rule, stress-testing is likely
to serve as the prime macro-prudential tool, serving as the link between developments in the
macro-economy and the levels of capital that banks will need to hold over the credit cycle.19

In the case of Europe, the countercyclical buffer is seen as a central macroeconomic
tool. Indeed, the struggle over the finalisation of CRD IV (the European directive imple-
menting Basel III) was in part about the right of member countries unilaterally to invoke
the countercyclical buffer to protect their national banking regimes against credit bubbles,
with the UK and the ECB expressing strong support for this approach. In contrast, the
EU Commission had argued against such national discretion under its effort to introduce
‘maximum harmonisation’ of CRD IV among its membership, where the minimum standards
would also serve as the maximum. The final CRD IV gives scope for countries to increase
macro buffers within limits.

Going forward, questions remain about the interaction among the evolving national
and regional stress-testing requirements, and the other Basel-based buffers, in particular the
conservation and countercyclical buffers, as well as the systemic surcharge for GSIBS and
domestic systemically important banks (D-SIB).

As a result of these various developments and initiatives, banks will face new challenges
regarding how RWAs, the leverage ratio and capital buffers are to be calculated going forward,
with implications for the capital planning process. This uncertainty around the direction of
consolidated capital requirements is further complicated by regional and national develop-
ments, requiring banks to hold capital buffers at the local level. Most notably, the recent
US proposal extending Enhanced Prudential Standards to foreign banking institutions would
require foreign banks to establish a separately capitalised US intermediate holding company
covering all banking and non-banking entities in the US.20 These entities would have to meet
all US capital requirements and be subject to the US supervisory stress-testing regime. The
outcome of the stress-test applied to the US operations of foreign banks would determine the
ability to make capital distributions from the US operations to the parent. Going forward,
with more capital committed at the local level and not available to meet losses elsewhere
in the group, banks may be required by the home supervisor to hold additional buffers to
meet unexpected losses at the group level.

As a precautionary measure, banks are likely to hold additional capital until there is
more certainty around the global, regional and national regulatory environment guiding the
capital planning process.

Liquidity regulation

In contrast to the 25 year experience with capital regulation as a globally co-ordinated Pillar
I requirement, globally harmonised liquidity regulation is a new development. The Basel
Committee had discussed the possibility of liquidity requirements for much of its history but
without agreement being reached on the necessity or possible approach.21

On 10 December 2010, the Basel Committee reached agreement on two minimum global
liquidity ratios, the liquidity coverage ratio (LCR), focused on ensuring enough liquidity over
a 30-day period of stress, and the net stable funding ratio (NSFR), focused on the structural
liquidity profile over a one-year time horizon. But given that this represented the first
ever minimum global liquidity standard, the Committee noted that work would continue to finalise these ratios as part of the so-called observation period to ensure that implementation did not result in unintended consequences. The recently finalised LCR will have a significant effect on banks’ balance sheets. The 2013 EY/IIF survey shows that US and European banks believe they will have to hold an average of 10% to 20% of their balance sheet in high quality liquid assets, a substantial increase from pre-crisis levels.22

Exhibit 12.4

**Liquidity requirements**

<table>
<thead>
<tr>
<th></th>
<th><strong>LCR</strong></th>
<th><strong>NSFR</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stock of high-quality assets &gt; 100%</td>
<td>Available amount of stable funding &gt; 100%</td>
</tr>
<tr>
<td>Total net cash outflows over the next 30 days</td>
<td>Required amount of stable funding</td>
<td></td>
</tr>
</tbody>
</table>

- The LCR will prescribe the quantity of high-quality liquid assets a bank must have at any given time.
- It aims to ensure that each institution maintains an adequate level of unencumbered, high-quality assets that can be converted into cash to meet its liquidity needs for 30 days under a specified acute liquidity stress.
- The net outflow is the cumulative expected cash outflow minus cumulative expected cash inflow over a 30-day period (using specified stresses).

- The NSFR is designed to provide incentives for banks to seek more stable forms of funding.
- Assets are weighted according to their required funding based on factors such as maturity, credit quality and liquidity. Liabilities are weighted according to the likelihood they will roll off during a period of stress (that is stability). For example, 100% of illiquid assets need to be backed with stable funding, but this is 65% for qualifying residential mortgages.

*Source: Authors’ own*

In January 2013 the Basel Committee used the observation period to make revisions to how the LCR is to be calculated and calibrated.23 That final calibration resulted in a range of compromises to make the LCR less binding than the original December 2010 agreement. To a large extent, the changes reflected caution over the possible financial system and macroeconomic effects of the LCR. If too large a portion of the balance sheet has to be held in liquid assets this reduces scope for lending, but some uncertainties remain over the final effects. The LCR will not be fully phased in until January 2019. Thus it will take until the next decade before policy-makers will have all the necessary information to assess the impact of the LCR and whether further refinements are necessary, although, as with capital, a significant degree of market pressure will result in front loading of the requirement.

Another area of difference between the LCR and the capital regime relates to its status as a binding minimum requirement. In the case of capital regulation, the traditional approach was that the minimum requirement could not be breached without triggering supervisory consequences, which were likely to result in the closing or restructuring of the bank. This has now been changed with the capital conservation buffer, which can be drawn down when the bank is under pressure, and the countercyclical buffers, which can be utilised in severe
times. The LCR regime, too, has been designed to be drawn on during periods of financial stress. In good times, banks would be expected to remain above the LCR 100% minimum requirement, but in a period of market stress the intention is for the liquid assets to be available to absorb stressed liquidity outflows. A key issue in this regard is the relationship between this objective, and the recently proposed LCR disclosure requirements. Banks may well be constrained in their ability to draw down the liquid asset buffer through fear of the market reaction if disclosure requirements are too detailed and frequent.

The second key reform affecting the future of liquidity regulation is the finalisation of the NSFR. As noted, while the LCR is a liquid assets buffer to cover stressed outflows over a 30-day period, the NSFR is intended to ensure that banks have a stable structure of liabilities to support the composition and maturity profile of their assets. The Basel Committee has announced that it will work to finalise the NSFR by the end of 2014. It could be argued that this initiative is even more important than the LCR, as the NSFR cuts to the core of maturity transformation in the banking sector. Given the importance of the NSFR, and the caution with which the BCBS has proceeded in finalising the LCR, it is likely that an equally careful approach will be taken to finalise the NSFR. It will be important to understand the scope of banks to raise sufficient stable funding to back longer term lending. This assessment of the capacity to raise stable funding may be a difficult task because savers’ willingness and capacity to place the requisite deposits in banks will vary by market.

The core question in relation to the NSFR is the considerable variability in access to stable funding bank by bank and market by market, making a uniform calibration more challenging than was the case for the LCR. It will be important that the NSFR structure is appropriate for different prudently run bank business models. Moreover, the current one year threshold will create new cliff effects and distortions in the markets. It will be necessary to review this assumption and determine how a more graduated maturity profile could be introduced into the framework, with appropriate granularity and risk sensitivity of the overall framework to reflect very different bank business models.

Finally, individual liquidity regulations cannot be assessed in isolation. There are a growing number of requirements which are putting ever more pressure on a limited pool of high-quality, unencumbered, liquid assets. These regulations include the following: (i) the LCR; (ii) the requirement to move OTC derivatives to central clearing; (iii) the near final proposal to require initial two-way margining for non-cleared OTC derivatives; (iv) growing supervisory expectations for the management and control of intraday liquidity risk; and (v) growing demands by host supervisors for branches and subsidiaries of foreign-owned banks to hold liquidity buffers ‘on shore’. The cumulative impact of these regulations is not fully understood. A recent study by the Committee on the Global Financial System argues that there appears to be a greater supply of collateral that could meet the growing demand, but this could come at the expense of greater interconnectedness, market concentration and procyclicality, among other broader systemic risks.24

**Risk governance and stress-testing**

Supervision of banks is undergoing a fundamental shift away from traditional asset quality and risk control reviews towards an assessment of overall business strategy, risk appetite,
culture and board effectiveness. This changing focus is grounded in the objective to move risk governance, and supervision, towards a more forward looking assessment of the risk profile of the company under both normal and stressed environments. The core of the issue is that however high the capital and liquidity buffers are, banks which make very poor strategic decisions, poor credit decisions, and carry excess concentration risk could still fail.

Traditionally, supervision and regulation of risk governance had taken a bottom-up approach. For each risk area (market risk, credit risk, operational risk and liquidity risk) supervisors had developed separate requirements, which included expectations for various degrees of board and senior management oversight.

More recently, supervisors have begun to take a top-down approach to articulating the expectations for the risk governance process and the role of the board. The Basel Committee initiated this process through the issuance of principles for enhancing corporate governance in October 2010. The Committee followed up in January 2013 with Principles for Effective Risk Data Aggregation and Reporting, which raise the bar for the role of the board in defining the type of risk information it should receive to guide the risk profile of the organisation. The principles also require that the data feeding such board risk reporting meet higher standards for accuracy, timeliness and comprehensiveness as well as clarity.

This focus on risk governance is reinforced by the recently released FSB peer review on risk governance, and the principles for an effective risk appetite framework. Through these peer reports, supervisors have for the first time articulated criteria for assessing the effectiveness of banks’ risk governance frameworks. A key focus is on how well the risk appetite informs strategic decision-making and risk behaviour throughout the company. There is also a growing focus of the regulators on risk culture and the tone at the top of the company. Moreover, the governance process around internal capital assessments, liquidity stress-testing, and recovery and resolution planning will move to the foreground of the supervisory assessment process.

It will be essential for banks to demonstrate the links between risk capacity, risk appetite, business strategy, capital planning, budgeting and stress-testing. Banks have made significant progress to strengthen the composition and skillsets of the board and its risk committee, to strengthen the reporting lines of the chief risk officer (CRO) to the board, and to raise the stature of the CRO role (prior to the crisis many banks did not have a board risk committee but that is now almost universal). Banks are also in the process of improving their stress-testing tools, including reverse stress-testing. However, as identified by the FSB peer review, banks still need to strengthen how they embed the risk appetite statement into the organisation, based on a common language for risk appetite, risk capacity and limits. Moreover, much still needs to be done to strengthen the risk infrastructure and data aggregation capabilities, and the quality and clarity of information provided to the board as a way to inform about the risk appetite.

Supervisors will be more likely to intervene in companies’ business strategies and growth plans if there are concerns about the quality of the risk governance process. An example of this is the proposed early remediation requirements under the US Enhanced Prudential Standards, which would result in restrictions on banks if there are shortcomings in the capital and liquidity risk governance processes. Another is the result of the US 2012 annual
stress-testing exercise, where some banks were cited for qualitative weaknesses in the capital planning process. In the UK, regulators are increasingly focused on the viability of banks’ business models.

**Strengthening recovery and resolution frameworks**

While capital, liquidity and enhanced risk governance reduce the probability of bank failure, enhanced recovery and resolution planning (RRP) frameworks are intended to reduce the severity of bank failures should they occur. At the global level, RRP related efforts are governed by the FSB paper, ‘Key attributes of effective resolution regimes for financial institutions’. The FSB paper sets out the type of reforms that countries need to undertake to make resolution regimes more effective, and they establish the steps that banks need to take to become more resolvable. G-20 members are at different stages in making the necessary legislative reforms to implement the FSB principles. In July 2010, the US introduced an Orderly Liquidity Authority as part of the Dodd-Frank Act to be able to resolve both the non-bank activities of bank holding companies, as well as any other financial institution designated as systemic by the Financial Oversight Council. In June 2013, the EU reached agreement on its proposal for a directive to harmonise national approaches to resolution across its member states. It also has introduced a proposal for an EU-level resolution authority to mirror the move to an EU-wide bank supervisory authority under the ECB.

Supervisors will elaborate on the presumptive paths to a company-wide resolution strategy based on either a single-point-of-entry (SPE) or multiple-point-of-entry (MPE) approach to resolution, using bail-in as the mechanism to recapitalise the institution. The purpose of bail-in is to ensure that bank creditors are the ones to recapitalise a bank that is no longer viable (through the write down or conversion of wholesale liabilities), instead of the public sector. Key issues in this regard are the instruments that should be subject to such a bail-in requirement, balancing the need for market discipline against the objective of ensuring financial stability and runs on the bank. A second key issue is the amount of bail-in debt that would be needed to restructure and recapitalise a failed banking institution. A third issue is ensuring that the investors in bail-in debt instruments of the banking sector are able to absorb the losses without causing the systemic risk of the banking sector to spill over to other sectors.

The SPE approach can be applied to banking institutions organised as holding companies with the operating entities structured as subsidiaries of the holding company. Under this approach, only the holding company is placed into resolution, while the operating entities would continue to function as going concerns. The MPE approach to resolution would apply to banking organisations that are structured in a more decentralised manner, and where it may be necessary to apply resolution at various entity levels in a co-ordinated manner. This may be the case for global banks with subsidiaries used to conduct business in different jurisdictions, especially in retail and corporate banking. Under either approach, key areas of cross-border focus include how to manage orderly derivative close-outs, how to ensure the cross-border provision of liquidity, and how to maintain access to financial market infrastructures during the resolution of a systemic financial institution. With increased incentives to move OTC derivatives to central counterparties, it will be even more important to develop sound resolution frameworks for such entities.
Banks from home countries with a clear presumptive path to resolution are likely to have a competitive advantage over those that do not have such a framework to guide their orderly cross-border resolution because host supervisors would be less likely to impose restrictions on inbound operations within their jurisdictions. However, it is important to note that it will take time for the many outstanding cross-border impediments to resolution to be addressed by home and host country supervisors.

Significant progress towards a global approach to cross-border resolution is critical to reducing pressure for host country ring-fencing and the associated fragmentation of banks’ global business models discussed below. The Federal Deposit Insurance Corporation (FDIC) has made good progress to flesh out the SPE approach to resolution, and it is co-ordinating closely with the UK and EU authorities. But more work remains to be done to address outstanding cross-border issues, including how to avoid the host supervisor from preemptively resolving the local branch or subsidiary before a global resolution can be attained. A second issue is how to prevent the premature close-out of derivative contracts in the host jurisdictions when the parent holding company goes into resolution (the stay on early termination of derivative contracts can be invoked in the home country but does not extend to the host jurisdiction). Preventing such outcomes will either require the change of local laws to operate in co-ordination with home country resolution framework, or derivative contracts will have to be reworked to prevent such automatic closeouts.

Over the coming years, supervisors will begin to flesh out presumptive paths to resolving cross-border banking institutions. Supervisors will begin pushing banks to take steps to remove impediments to their effective recovery and resolution, although this is likely to be an iterative process versus a single big-bang event. Banks can expect to face potential, additional structural requirements to make their resolution plans workable, such as perhaps moving to a holding company and operating subsidiary structure, requiring a minimum amount of bail-in senior unsecured debt, increasing the alignment between business lines and legal entities, and requiring the establishment of stand-alone service companies that can survive the resolution of parts of the group.

**Structural reforms**

While the focus of this chapter has been on global reforms, it is important to note that there are a range of additional regulations, largely locally driven, that will have a major impact on banks’ existing global business models. The result will be a major shift in banks’ approaches to capital management, funding and governance.

Over the past decades, the emphasis of international supervision and regulation has been on promoting consolidated risk management frameworks. Moreover, the Basel Concordat placed the home supervisor at the centre of the process for co-ordination among home and host supervisors. Since the recent financial crisis, however, there has been a growing supervisory concern that as going concerns, banks are managed globally but in stress they fail locally, with substantial costs to host supervisors and local economies. This point was reinforced in particular by the experiences of host supervisors with the failure of Lehman Brothers and that of the Icelandic banks, among others.
As a result of these developments host supervisors have asserted themselves, placing greater restrictions on the operations of global banks within their jurisdictions. Moreover, a number of home supervisors have taken steps to protect their economies from the failure of complex institutions engaged in capital markets activities through the introduction of so-called structural reforms focused either on limiting the types of activities that banks may perform or the entities in which those activities may be located.

More specifically, the drivers of structural reform in the banking sector include the emerging recovery and resolution planning requirements, discussed earlier. They also include the EU's Liikanen proposal to put a ring-fence around trading activities, the UK Independent Banking Commission initiative to ring-fence retail activities, and the US push-out provision for OTC derivatives (the so-called Lincoln Amendment). Local activity restrictions, like the Volcker rule and similar French and German initiatives, will also contribute to restructuring of the banking industry, particularly capital trading and capital markets activities.

Finally, host country efforts to constrain the operations of in-bound banks are compounding the trend towards fragmentation of banks’ global business models. Most notable is the recent US Notice of Proposed Rule-making addressing the enhanced prudential standards for foreign banking organisations in the US, which would require the establishment of a US intermediate holding company for foreign banking organisations (FBOs) with more than US$10 billion in non-branch assets. The rule would also require that FBOs match liquid assets and liabilities inside the US. Finally, there is the proposed requirement to establish a US level risk governance infrastructure. There also is growing pressure in the UK to move branches towards a de facto subsidiary structure and to discourage the remote booking of derivatives by overseas banks into their UK branches or subsidiaries. Within Asia, Singapore has required that retail activities be performed through subsidiaries.

A clear goal of these structural reforms is to protect local economies from the stress arising from trading losses and from cross-border banking activities, particularly in the area of capital markets. To date, with the exception of recovery and resolution planning, these structural reforms have been the sole purview of national and regional supervisors, with limited effort at global co-ordination. This absence of global co-ordination of local initiatives poses the risk that the end result of the individual national efforts produces an aggregate outcome that is suboptimal from the perspectives of both global stability and efficiency. Moreover, there is a growing risk that other countries will respond to the actions of host country supervisors, imposing restrictions of their own on the activities of in-bound foreign banks. Countries that do not take similar steps to protect their local market from a foreign bank failure could find themselves subordinated to creditors in those countries that have ring-fenced assets to support local liabilities.

As a result of these developments, many cross-border banks will need to rethink their global business models, legal entity structures, booking models, capital structures, funding models and governance structures. Reassessing all of these elements in an integrated manner is as important as it is challenging, especially given that structural reform initiatives continue to evolve at both the level of home and host supervisors.
OTC derivatives markets – regulatory reform

Reform of the structure and transparency of OTC derivatives markets is one very important aspect of regulatory change post crisis, required by the G20 to strengthen resilience by helping reduce the interconnectedness among major financial institutions and thus the severity of failure. The changes are far reaching and will change the shape of markets for both standardised and non-standardised contracts. Standardised contracts will have to be traded on an exchange or an electronic platform and cleared through a central counterparty. Non-centrally cleared derivatives will be subject to higher capital requirements and new margin requirements. All OTC derivatives will have to be reported to trade repositories. These developments will represent a major change to the organisation and structure of the OTC derivatives markets and the key participants in these markets.

Background to the changes

The OTC derivative markets (which include interest rate and credit default swaps), which play a central role in risk management by financial institutions and corporates and also in risk taking, grew rapidly over the 2000s. In 2011, the counterparty exposure of US banks through derivative contacts (on the mark-to-market value) was over US$1trillion. The size of the exposures partly reflects the nature of the market, where market risk exposures are offset not by selling an existing contract to a third party but by entering into a new contract to create the desired risk profile. Each contract potentially leads to counterparty risk because of the replacement cost if a counterparty reneged on its obligation. Another feature of the market is the degree of concentration. Because the contracts are bilateral promises to pay, which depends on the creditworthiness of the entities involved, a sizeable proportion of risk rests with the largest players at the centre of the market who are seen as most creditworthy. Mengle shows that in mid-2010, the largest 14 derivative dealers held 82% of the total notional amount of OTC derivatives outstanding.

There have been waves of change to elements of the OTC derivatives market to reduce risks over the past two decades. The introduction of the prototype for the ISDA Master Agreement for interest rate and currency swaps in 1987 followed by the publication of the Master Agreement in 1992, which supported netting of exposures between counterparties with legally enforceable netting agreements in place, allowed the market to grow massively. Also to reduce the ever increasing counterparty exposures, many banks included clauses to reset the market value as reference rates moved and the MTM reached pre-set levels. More use of collateral was introduced, although many types of counterparty still do not provide collateral against exposures to them.

Changes were also made in the credit default swap market. Over the course of the 1990s and 2000s there had been an explosion in the use of credit derivatives, particularly single name credit default swaps, which give the buyer protection against the default of the reference entity. The notional amount of credit default swaps had reached US$34 trillion by end 2006. Processing capabilities in the banks for credit derivatives had not kept pace with the growth in the market and backlogs in formal confirmations of trades had developed,
creating the risk of losses because of errors. One reason was the use of manual processes for confirming trades. Another problem was that some counterparties had assigned their side of existing trades to new parties without notifying the dealer. Since 2005 there has been a significant improvement in introducing automated confirmations in credit derivatives and procedures were developed to ensure that there was appropriate notification to all parties in the case of trade assignment.

In addition, by December 2006 a CCP was managing the risk in a significant proportion of inter-dealer single currency interest rate swaps. CCPs had been used extensively for exchange traded futures and options. The shift was to use them for OTC contracts. By 2009, clearing and settlement infrastructure was in place for many contracts including FX, credit default swaps (CDS) and equity derivatives. But the majority of OTC contacts were still arranged bilaterally, and cleared and settled bilaterally.

The failure of Lehman Brothers, given the company’s position as a major derivatives dealer, highlighted a range of continuing issues with the market both in terms of resolution of the failed entity and in terms of contagion, with fears regarding large bilateral exposures with insufficient collateral. Concerns about the risks in the derivatives markets triggered a further wave of far reaching reforms post crisis.

The changes are being introduced in different markets in different ways. The Dodd-Frank Act is being used to introduce the changes in the US. In Europe, the changes will be part of the European Market Infrastructure Regulation (EMIR). Other jurisdictions in Asia are also moving ahead with reforms to their OTC markets, particularly in Hong Kong, Singapore and Japan. The FSB is working to co-ordinate the G20 OTC derivative reform process through a series of high level principles, time-lines, and peer review monitoring, but major concerns remain about missed deadlines and cross-border conflicts, inconsistencies and duplication of requirements.

**Central counterparties**

One of the key reforms proposed by the G20 was that all standard OTC contracts should have moved to CCPs by end 2012 but this shift is still ongoing.

The CCP is designed to standardise and mutualise counterparty credit risk by intermediating between counterparties. In a swap trade, each party will contract with the CCP rather than with each other. In effect it acts as a circuit breaker and mutualises losses. Companies which deal with the CCP are clearing members and each contributes to a default fund to protect the CCP from default. The CCP takes initial and variation margin to reduce its exposure to counterparties. The initial margin is set to cover possible losses (usually up to a particular confidence level) in liquidating a defaulting member’s portfolio. Variation margin is taken daily to cover any mark-to-market changes in the value of contracts – the net amount owed by the member to the CCP. If a clearing member failed, the loss would be absorbed first by the collateral posted as margin by that member. If losses were larger than this then they would be covered by the default fund contribution of that member. For losses above this level, each CCP would have a waterfall of loss absorption setting out the order with which different components come into play – the mutual default fund provided by all clearing members, pre-funded, would probably absorb
losses next, and at some point the CCP’s equity (provided by shareholders) would come into play. In addition, there are usually commitments by members to provide further funds to support the CCP. The sequential buffers all reduce the likelihood of market players having to absorb losses, and, if finally they do have to absorb part of the loss, the burden is shared across all the clearing members, reducing the likelihood that another company would be pushed into default. Following a default, however, clearing members are required to recapitalise the default fund. If the loss absorption procedures cannot meet all the losses then the CCP would have to go into resolution, but the procedures for this have yet to be worked out.

The CCP reduces credit and market risks in other ways as well. In bilateral contracts between two counterparties, although each counterparty can net its various positive and negative mark to market values on different contracts with the other counterparty (if the right netting agreements are in place), multilateral netting across a range of counterparties is not possible. In contrast, when dealing with a CCP as the counterparty there is scope to net across all of the contracts a company has with the CCP across many counterparties. This keeps individual exposures lower. To make this multilateral netting most effective, a small number of CCPs is ideal.

Another issue is that with bilateral trades between market players, if one fails there will be a fire sale of the collateral posted by the failed company with its counterparties. This in turn drives prices lower in the affected markets, reducing the value of collateral held and also of outright positions, creating more market disruption and spreading problems further across the market. The CCP, in contrast, can allocate the collateral it holds to other players rather than releasing it.

Another aspect of the wave of disruption in bilateral OTC markets, caused by the failure of a large dealer, is that other companies have to replace the contracts lost (to keep positions hedged) and this also creates market disruption. The CCP potentially helps in this area too because it will auction the failed company’s contracts, enabling companies to replace contracts more easily. The CCP can also facilitate the transfer of customer positions from the failed company to a sound company. In addition, more certainty is provided to market players regarding the margin they have provided against their exposures, which could be caught up in the bankruptcy process if they were a direct counterparty to the failed company rather than counterparty to the CCP.

However, the reduction in systemic risk is not a given. The CCP too can create ripples across the market by increasing the margin requirements sharply when the players are short of liquidity, effectively draining more liquidity out of the system. Also the CCPs are themselves very large nodes in the system. Although the smallest number of CCPs is best from a netting perspective, this makes the importance of each CCP larger. If one failed the disruption would be enormous. Although failure is rare, three have failed since 1974 – Paris, Kuala Lumpur and Hong Kong.

The systemic risk to the system from failure of a CCP is so high that there is a tension between ever higher buffers to protect CCPs and market efficiency. To mitigate the risk of CCP failure, the FSB has recently issued draft proposals for the recovery and resolution of CCPs, which is critical to financial stability given the incentives to move a growing share of OTC derivatives to CCPs.
**OTC contracts covered by mandatory clearing**

It was decided that the range of OTC contracts that would be subject to mandatory clearing would be determined by the authorities in each jurisdiction. This approach allows for the evolution of the list of products covered as markets change. The IOSCO technical Committee makes the point that ‘there may be products that cannot be effectively managed by a CCP today and therefore are not subject to mandatory clearing. Over time however, the product characteristics could change such that products can be risk managed and so become suitable for mandatory clearing’.

Two approaches are used in assessing what products should be captured by mandatory clearing. Under the first, so-called bottom-up approach, the products that a CCP clears or proposes to clear become mandatory for clearing. Many authorities are proposing to use this approach. Under the second, called the top-down approach, the authority identifies the OTC derivative contracts where clearing is desirable, irrespective of whether they are currently covered by a CCP.

In considering if a product should be subject to mandatory clearing the FSB stated that authorities should take into account a number of factors:

- the degree of standardisation of a product’s contractual terms and operational processes;
- the nature, depth and liquidity of the market for the product; and
- availability of fair, reliable and generally accepted pricing sources.

Additional criteria are set in the legislation in different jurisdictions. In the US, the Dodd-Frank Act requires the authorities to consider the effect on competition and mitigation of systemic risk when considering mandatory clearing for a product.

With regard to existing contracts and whether they should be moved into central clearing, the FSB suggested that both legal and practical considerations (including the ability for the CCP to manage the increase in volumes) should be taken into account.

**Requirements for non-centrally cleared derivatives**

Despite the increasing use of CCPs there is an expectation that contracts with notional amounts equivalent to trillions of dollars will remain outside them because the contracts are not standardised. The authorities are concerned that this could lead to an unlevel playing field between cleared and un-cleared contracts and lead to the buildup of systemic risk. Under current market treatment, whereas the CCPs impose margin requirements, many bilateral agreements do not have margining. To deal with this it is now proposed that parties to non-cleared OTC derivatives contracts should also impose bilateral margining (unless they benefit from exemptions built in for smaller players and end-users). In addition, because of continuing concerns that the risks will be higher in bilateral trading, market participants will have higher capital requirements on counterparty exposures on un-cleared contracts.
**Margin requirements**

The Basel Committee with IOSCO has produced near final proposals on margining for all derivatives transactions not cleared by a CCP.\(^47\) The proposals would require bilateral initial and variation margin but allow for a fairly wide pool of collateral, including equities in main indices as well as covered bonds and gold in addition to high quality government, central bank and corporate bonds. Each supervisor will establish a list of allowable collateral for their jurisdiction which meets the Basel principles, for example, it must be possible to liquidate securities held as collateral in a reasonable time frame and (after allowing for margin) they should hold their value in times of stress.

The move to bilateral margin requirements will have a major effect on the collateral needs of companies. A quantitative impact study (QIS) carried out by the Basel Committee indicated that collateral required for these contracts would increase from US$100 billion to US$700 billion.\(^48\) This estimate assumes about a 50% reduction in non-cleared OTC derivatives and the full use of the internal modelling option (see below). If the standardised margins were used, the impact would be 6 to 11 times greater than this estimate. The QIS also shows that the margin requirements set by CCPs are a fraction of the requirements proposed for non-cleared counterparties, driven by the fact that CCPs are able to take advantage of multilateral netting.

The requirements will be phased in, with those applying to variation margin effective from 15 January 2015 and two way initial margin on new contacts being phased in between 2014 to 2019, with quicker phase in for the largest OTC derivative players.

Under the proposals, all financial companies (subject to various size thresholds) and systemic non-financial companies must exchange initial and variation margin reflecting the counterparty risk on these transactions. Initial margin must be exchanged by both parties gross (that is, without netting). There are also rules about treatment of margin in a bankruptcy to try to mirror protection afforded by use of a CCP. The margin must be immediately available to the collecting party in the event of the counterparty default. In addition, the posting party must be protected if the collecting party enters bankruptcy (to the extent possible under applicable law).

Banks can use models to calculate the base line initial margin as well as haircuts. The potential future exposure should reflect an extreme but plausible estimate of an increase in the value of the contract consistent with 99% confidence (one tailed) over a 10-day horizon. The models may account for diversification, hedging and risk offsets within well-defined asset classes, such as currency/rates, equity, credit, or commodities but not across asset classes. In addition, the proposal includes standardised approaches for both initial margin and collateral haircuts. These must be used if banks do not obtain approval to use their internal models.

**Implications**

The implications of these proposals for major OTC derivative players and their counterparties will be profound. It will create a substantial incentive to use standardised contracts that can be cleared through a CCP to take advantage of more favourable capital and margining
requirements. This will drive changes in contracts going forward. But some contracts will not be amenable to standard clearing because they need to be customised to reflect the particular risk profile of an end-user. It is not yet clear if currency contracts will be captured by the bilateral margining requirements, but if they are this would have the largest effect on the corporate sector.

For contracts which cannot fit into the CCP framework, banks will have to maximise the degree to which they can net the variation margin. This will give an incentive for players to conduct these transactions with a small number of high quality counterparties. This focus on a smaller number of counterparties would tend to concentrate exposures even further, although the new proposed large exposures rules would place a limit on exposures to any one counterparty. Even so, the cost of trades will increase and charges will have to increase for all non-standard contracts. The bilateral margining of uncleared trades will also generate substantial systems change and modelling costs in the industry.

Derivatives trading and reporting

Interest rate swaps and other OTC derivatives are generally traded bilaterally between two counterparties. The G20, to improve transparency, decided that standard contacts should be executed on an exchange or electronic platform. In line with the requirement that standard contracts should be cleared through CCPs by end 2012, the same date was chosen for the use of trading platforms or exchanges for these contracts. Non-standard trades or trades involving a corporate end-user will be unaffected. Other contracts will have to be traded through a US type swap execution facility (SEF) or designated contract market (DCM). SEFs use an order matching mechanism, for example a central limit order book or a request for quote system, allowing the posting of bids and offers. The order matching mechanism will have to be linked to systems to handle communications and produce trade data. The CFTC decided in the US that most swaps and futures should be transacted on a DCM.49

In addition, the G20 agreed in Pittsburgh,50 that OTC derivatives contracts should be reported to trade repositories; electronic platforms which collect information on derivatives trades.

Progress on reforms

The FSB set out the state of play on OTC derivative market reforms in its fourth report in October 2012.51 Probably the greatest progress has been made in central clearing but even here it is patchy. Although the FSB make clear that it is not possible to determine precisely how much of the total market (standardised and non-standardised products) is centrally cleared, estimates indicated that only approximately 10% of outstanding CDSs were centrally cleared, whereas approximately 40% of outstanding interest rates derivatives had been centrally cleared as of end-August 2012. The greater coverage for interest rate contracts probably reflects the greater uses of CCPs in this area pre crisis.

The FSB reports that CCPs are available to clear some OTC derivatives in each of the five asset classes (commodities, credit, equity, foreign exchange and interest rates) and there are plans by many CCPs to expand their services in the near future, especially for interest rate
derivatives. There are also plans to create links across CCPs and to add more participants. However, regulatory uncertainty (particularly across jurisdictions) and operational challenges are seen as obstacles to further expansion by some.

The FSB set out information for 19 CCPs across 9 jurisdictions. One message from the data from the 19 CCPs is that it can take time for new products to be added, particularly if they are complex – a range of timeframes were cited, 4 weeks to 21 months. Part of the time reflects regulatory approval.

Exhibit 12.5 sets out the information provided by the FSB on central clearing of contracts related to various asset classes. For a number information was not available.

Exhibit 12.5

**Summary of CCP coverage from FSB (2012)**

<table>
<thead>
<tr>
<th>Location</th>
<th>Percentage of outstanding transactions</th>
<th>Notional outstanding US$ billions equivalents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Commodities</strong></td>
<td>Brazil, Germany, Singapore, Sweden, UK, US</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Credit</strong></td>
<td>France, Germany, Japan, UK, US</td>
<td>12%</td>
</tr>
<tr>
<td><strong>Equity</strong></td>
<td>Brazil, Canada, Sweden</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Foreign exchange</strong></td>
<td>Brazil, India, Singapore, UK, US</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Interest rates</strong></td>
<td>Brazil, Singapore, Sweden, UK, US</td>
<td>40%</td>
</tr>
</tbody>
</table>

*Source: Authors’ own; Financial Stability Board, ‘Implementing OTC derivatives market reforms’, fourth progress report on implementation, October 2012*

With regard to trade reporting, FSB\textsuperscript{32} stated that trade reporting exists for transactions in all asset classes and is extensively available for interest rate and credit risk derivatives (covering 90\% of trades). However, significant gaps remain, in particular for OTC derivatives in commodities, equities and foreign exchange – for the latter only around 50\% are reported. The FSB also noted that progress in using the data by the official sector is hampered by data aggregation issues. Dealing with these issues will require more work on compatible data formats for reconciliation.

The area where there has been the least progress is the trading infrastructure. The FSB concluded that this partly reflects ‘uncertainties about the scope and form of future regulatory frameworks for organised platform trading’.

Certain derivative products can be traded on organised platforms, with most in Europe or the US – some for dealer-to-dealer transactions while others cater for dealer-to-customer. The basic structure, according to the FSB, is global access on line with local offices and trading screens in other markets. The FSB finds that in some countries organised trading is most common for credit and interest rate swaps. This reflects the fact that these are the
most standardised contracts. In addition, the FSB finds that some organised platform facilities exist for a limited range of foreign exchange and equity derivatives.

The FSB expected that new platforms would start to spring up once the regulatory framework for mandatory organised platform trading was in place. But it will also take more standardisation of products for this goal to be achieved.

**Overall implications of the OTC reforms**

The changes discussed will fundamentally affect markets and players and it is important that there is a harmonised approach across the main markets. The absence of harmonisation could drive regulatory arbitrage between markets. Currently there are differences between the various regulations implementing the changes, for example between Dodd-Frank in the US and EMIR in Europe. The regulations are also being introduced at different speeds in different markets. A number of smaller markets are not expected to implement trade execution rules.53

The CCP rules and the rules for non-cleared derivatives will change the cost of contracts for end-users, as well as traders, although clearing through CCPs should improve transparency, including with regard to pricing. Institutional investors which do not currently have to put up margin on OTC derivatives are faced with systems and data challenges of margining trades with CCPs (they may escape from the non-cleared margining). Nonetheless, the CCP margin will be substantial for these players. BNY Mellon54 cites a typical 5-year vanilla interest rate swap as having initial margin equivalent to 1% to 3% of the contract with longer dated contracts requiring more – perhaps 10% of notional for a 15-year swap. BNY Mellon argue that institutional investors are more likely to be adversely affected by initial margin because they have directional plays, whereas dealers are more likely to be matched enabling margin with the CCP to be netted down.

The reforms in terms of use of CCPs and bilateral margining for non-cleared trades all put pressure on available collateral. This is particularly the case because of other calls on collateral from the post crisis reforms – in particular the new higher liquidity pools for banks under Basel III. Eligible collateral can be found through securities borrowing from institutions, but these assets are already heavily used in the secured funding markets or repo. Euroclear has launched a collateral superhighway that will bring collateral in from the Far East, which will then be lent to market players. But this will not be enough to alleviate the pressures.

BNY Mellon55 points to fears of a ‘collateral crunch’, where yields on these higher quality assets are driven lower and lower. One solution they see is the use of collateral upgrading or transformation through clearing brokers of custodians. Customers are able to use the clearing broker to swap lower quality securities for those acceptable to the CCP, but they see this as only helping some end-users. It also will not alleviate market wide shortages.56 Nonetheless, despite the greater costs of trading, the increased standardisation of contracts and market pressure to move them into CCPs is likely to encourage more market liquidity.

**Conclusion: the wider effects of regulatory change**

The lessons of the crisis clearly pointed to the need for more and higher quality capital, minimum global standards for liquidity, stronger risk governance, and the move towards a
The future of financial regulation

A macro-prudential approach to supervision to address procyclicality and interconnectedness among financial institutions, including through reforms to the OTC derivative markets. Moreover, there is a broad consensus on the need to make global financial institutions more resolvable in a manner that limits systemic risks and the potential for taxpayer bail-outs.

However, the continued pace of regulatory change raises several fundamental questions. One is the long run effect on the shape of the industry and markets. Another is how long it will take to reach a steady state and the transition costs.

The adjustment is already underway. Banks are deleveraging and de-risking post crisis, but are also reassessing their fundamental business in the light of the overall costs. This includes the capital and liquidity buffers and the OTC derivative reforms. With much higher capital buffers, returns on equity (ROEs) have been adjusted downwards sharply, but the banks are now under pressure from institutional investors to try to increase returns.

In terms of adjustment costs there are several areas of strain. The message from some institutional investors is that the regulatory uncertainty surrounding the industry has made it almost impossible to assess future earnings. This reflects the fact that Basel III was not a clean step change to a new tighter capital and liquidity environment. The changes are continuing in a variety of areas. There is a further change expected to trading book treatment, further work on liquidity (necessary to ensure the ratios are workable), work on operational risk measurement, standardised counterparty and market risk treatments and so on. All these areas create considerable uncertainty but that is now outweighed by the additional uncertainty created by a fundamental revisiting of all aspects of the risk weighting regime not just for the trading book, but also for the banking book and for operational risk. While there is a need to ensure transparency and a reasonable amount of comparability of the risk weighing regime, full standardisation as the end in itself, with a stand-alone leverage ratio as the extreme incarnation of this approach, creates a different set of risks, namely the unleashing of a new cycle of capital arbitrage, as already occurred under the Basel I regime.

Even more fundamental are the range of locally driven initiatives focused on protecting national economies and creditors to the detriment of a global approach to supervisory collaboration and risk management. In particular, there is a risk that these local initiatives undercut the global effort to arrive at a co-ordinated approach to the resolution of cross-border financial institutions.

Institutional investors react to uncertainty by heavily discounting future earnings and price to book ratios of banks are indeed at a 20 year low. Although this is in part because of continuing economic uncertainty and concerns about possible issues regarding asset values in some European banks, the regulatory uncertainty is undoubtedly an important factor. This uncertainty could make the adjustment phase more difficult and costly. For many banks retaining earnings is the only practical way to increase common equity Tier-1 capital. Given the market pressure to move faster to completing Basel III, this is creating some strains in banking systems where profits are lower and core business models remain in question.

In the long term bank business models have to change. Proprietary trading and capital markets activities are more difficult to remunerate given the much higher capital charges, and many banks are scaling back considerably or considering focusing on managing customer flow rather than own positions. Ring-fencing of trading activities will also make it hard for banks to continue active proprietary trading and other capital markets activities because...
funding costs would be higher. Moreover, the line between proprietary trading and market making is difficult to draw. OTC derivatives activity will continue but the new large exposures rules and the higher margin required for bilateral exposures and exposures to CCPs makes it difficult for the size of interbank exposures between the core players to be maintained. This will lead to a change in the market with positions being closed out faster. Across all portfolios, banks are looking for collateral to reduce risks as well as managing information on collateral much better than in the past.

A significant share of traditional banking activities, both credit creation and trading, will likely move into the shadow banking area. In terms of credit extension, institutional investors are increasingly providing direct secured lending – for example different types of project finance – driven by a combination of search for yield in a low interest rate environment and opportunities given the retrenching of banks. Proprietary trading is moving into hedge funds. The inter-linkages between the banking and shadow banking sectors are changing. One question here is the extent this will push new forms of risk concentrations below the official radar. Oversight of shadow banking is being enhanced but may catch issues too late. The effect of the changes could be to reduce market making capacity in some markets compared with more opportunistic position taking in others.

Banks are retreating to their core geographies and their core small, medium enterprise (SME) and retail customers. But even here, to set a profitable strategy banks will have to increase margins, reduce costs, and exit products which will not be sufficiently profitable. However, as more and more banks focus in on these activities, competition could drive down margin rather than allowing an upward adjustment. Competition will also come from shadow banks as well as other banks with new types of platform such as peer to peer lending which are currently very small but could grow. Banks have large quantities of information, not all of which is used in an optimum way. It will be important to mobilise the information to sift the customer base and use differentiated pricing, to keep ahead of the competition.

The area of retail where banks with less access to customer data (because they do not have the primary transactions accounts) have felt more comfortable transacting is mortgages, because of reliance on security. The focus back onto retail could again lead to bubbles developing in housing markets. The macro-prudential tools of the authorities can be brought into play but 2.5 percentage points on capital requirements will probably not have sufficient effect if housing prices are rising by 10% per annum, unless the general level of interest rates is used as well (that is, bringing monetary policy into play), or the use of other tools such as loan to value (LTV) limits or higher risk weights for the asset class in question. Given the challenges associated with the use of macro-prudential tools, it will be all the more important to focus on resiliency through appropriate capital and liquidity buffers, limits on large exposures to systemic counterparties, sound governance practices, and the robustness of CCPs and other core market infrastructures.

The regulatory changes are both radical and sweeping. Much of the change is focused on the prudential buffers the banking industry has to hold which changes the costs of business and will force a retreat from some activities. Other changes such as the recovery and resolution planning, bail-in, OTC derivatives reform, and liquidity regulation are a departure from the old order of regulation, and the full effects are hard to foresee. It would be good
to give the current set of G20 reforms a chance to play out before adding even more to an already full plate.

Bibliography


We would like to thank the Ernst & Young Global Regulatory Network, in particular Thomas Huertas and Donald Vangel, for their comments.

This chapter contains information in summary form and is therefore intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. Neither EYGM Limited nor any other member of the global Ernst & Young organization can accept any responsibility for loss occasioned to any person acting or refraining from action as a result of any material in this publication. On any specific matter, reference should be made to the appropriate advisor.

The articles, information and reports contained within this book are generic and represent the views and opinions of their authors. They do not represent the view or opinions of Ernst & Young LLP or any member of the Ernst & Young network, nor does it express any form of assurance. Where articles, information and reports are produced by Ernst & Young LLP they do not contain advice and should not be regarded as comprehensive or sufficient for making decisions nor should it be used in place of professional advice. Accordingly Ernst & Young LLP accepts no responsibility for loss arising from any action taken or not taken by those receiving the book. (EYG no. EK0144)

1 On 1 March 2013, AIG repaid the paid the final instalment of its US$182 billion government bailout, plus a profit of US$22.7 billion, to the US Treasury.
19 Bernanke, B, ‘Stress testing banks: what have we learned?’ April 2013.
20 Federal Reserve System, ‘Enhanced prudential standards and early remediation requirements for foreign banking organisations and foreign nonbank financial companies; proposed rule’, Federal Register 77(249), 2012.
37 Government Accountability Office, ‘Credit derivatives: confirmation backlogs increased dealers’ operational risks, but were successfully addressed after joint regulatory action’, Report to Congressional Requesters, June 2007.
43International Organization of Securities Commissions, ‘Requirements for mandatory clearing’, February 2012.
The future of financial regulation

49 Commodity Futures Trading Commission, ‘Core principles and other requirements for designated contract markets; final rule’, June 2012.
54 BNY Mellon, ‘To clear or not to clear…collateral is the question’, BNY Mellon Global Collateral Services, Rule Financial, 2012.
56 But also see CGFS study.
Chapter 13

Global banking after the cataclysm

Roy C Smith
NYU Stern School of Business

Overview
The financial events of 2007–2008 began with a liquidity crisis in structured financial products that drove market prices down to levels well below their intrinsic values, causing massive write-offs among large global banks. Market price adjustment spread well beyond structured finance and government intervention at an historic level followed, precipitating the largest regulatory changes in the global financial system that had ever occurred. This chapter attempts to place all these events in perspective, and addresses the key issues that will shape the future of global banking industry.

Introduction
After many years of continuous deregulation, global market integration and major advances in telecommunications technology, world financial markets became vast, liquid and efficient. The markets were efficient in the sense that they became easily accessible, were subject to fairly common pricing metrics and were marketable across countries and regions. As a result, global financial markets lowered the cost of capital to governments, corporations and others around the world, allowed risk to be allocated to investors wanting it the most, and shifted assets from the balance sheets of banks and insurance companies to institutional investors where they became tradable.

At the end of 2007, the last year before the financial system collapsed, the market value of all tradable securities and loans in the world amounted to US$202 trillion, up from US$54 trillion in 1990 (see Exhibit 13.1). Even after the crisis, markets expanded. The value of securities and loans in 2012 was about US$225 trillion, approximately 10 times greater than the GDP of the US and the EU combined, and about 3 times global GDP, up from 2.6 times in 1990. Growth of capital markets, however, has been much slower after 2007 than before, and global cross borders flows have been reduced as markets have borne the strain of recession, European sovereign distress and reorganisation of the banking industry.
Exhibit 13.1

Market capitalisation of global financial stock

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-securitised loans</td>
<td>5.5</td>
<td>4.8</td>
</tr>
<tr>
<td>Securitised loans</td>
<td>15.9</td>
<td>–0.7</td>
</tr>
<tr>
<td>Corporate bonds</td>
<td>5.1</td>
<td>9.1</td>
</tr>
<tr>
<td>Financial bonds</td>
<td>10.7</td>
<td>1.5</td>
</tr>
<tr>
<td>Government bonds</td>
<td>8.3</td>
<td>9.2</td>
</tr>
<tr>
<td>Equity</td>
<td>8.0</td>
<td>–5.5</td>
</tr>
<tr>
<td>Total</td>
<td>8.1</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Source: McKinsey Global Institute Finance Assets Database; McKinsey Global Institute analysis

www.rasabourse.com
Nonetheless, markets have become so large that changes in investor preferences, caused by fear, greed, or other, more rational, impulses, can now create enormous market forces beyond the control of governments that are capable of creating and bursting financial bubbles that can have profound effects on the global economy.

**Investment banks as intermediaries**

Investment banks are the intermediaries in capital markets, acting as underwriters, brokers and dealers (traders), and advisers in merger and other transactions. They have evolved from a large number of relatively small and specialised companies in the 1970s into a small number of very large, diversified companies in 2012. The evolution of the industry from low-risk advisory companies into high-risk, trading-intense, multiple-platform companies has been accompanied by mishaps, scandals and strategic mistakes amidst two and a half decades of profitable and rapidly increasing capital market activity. For several large commercial banks focused on servicing corporate or government clients, the investment banking business had not only become attractive for its returns, it was also thought to be essential to their long-term business strategies as capital markets displaced traditional lending businesses. In the US, banks lobbied extensively to remove the 1933 Glass-Steagall law that separated banking and securities businesses. After years of incremental lifting of the barriers, the old law was repealed in 1999 – a year after the otherwise illegal Travelers-Citicorp merger was announced. In Europe, ‘universal banks’ were never restricted in their investment banking activities, though most were never as aggressive as their American counterparts were to become after 1999.

The repeal of Glass-Steagall allowed US banks to cross over into the riskier and, for most, unfamiliar, territory of capital markets. The banks were also permitted to expand across state lines by the repeal in 1994 of the McFadden Act of 1927.

**Increased competition**

By 2007, the dozen or so ‘money centre’ banks of the 1990s had been consolidated into just three very large Bank Holding Companies (Citigroup, JP Morgan Chase and Bank of America) that engaged in a wide variety of financial services, including investment banking. These institutions, all the result of numerous merger and acquisition transactions, were seen as ‘too-big-to-fail’. Because of their size and importance, it was assumed that the government would have to intervene to protect depositors and other liability holders from any sort of market run on the bank. This feature was soon reflected in the relatively low rates at which these banks were able to fund themselves in the markets.

**Basic business models emerge**

In the US, the three large commercial/investment banks developed ‘big-balance sheet’ business models. They offered their lending capacities to corporate clients together with the cross-selling proposition that they also use the banks’ investment banking capabilities.

*Big balance sheets:* but for this approach to be convincing to clients, it had to be capable of operating at a large scale – by developing into very large, too-big-to-fail entities,
the banks would possess very large legal lending limits, which they could make available to clients seeking to complete large mergers or leveraged buy-outs. Such clients, eager to be able to arrange all or most of the financing facilities they needed with one phone call, were willing to insist that the banks be added as co-managers to its investment banking advisers. In time as their skills in investment banking increased, they were able to compete for lead manager roles.

By 2007, Citigroup and JP Morgan Chase occupied the top two market share positions among investment banks. As a result, US investment banks and those Europeans interested in capital markets, had to confront the rapidly increasing footprints of these giant US banks, both by new product innovation (such as more sophisticated forms of ‘structured finance’ involving collateralised securities), and by a willingness to increase the risk they took in offering transactions to clients.

*Flow monsters:* beginning in the 1980s, the stand-alone investment banks evolved into trading companies in response to the rapid increase in US treasury and agency debt outstanding, and the need for market solutions to the savings and loan crisis of that time. As they did so, they became more comfortable with the practice of ‘bought deals’, in which they would offer to purchase an entire block of debt or equity securities, either through an underwriting with a corporate client, or a block-trade with an institutional investor. Taking this risk on their books meant that the companies had to develop exceptionally strong resale capabilities (placing power) and the ability to hedge positions they retained. Thus they became large-scale market-makers to both issuers and investors in all types of securities, and invested heavily in their market information and sales networks and in technology to support them. With these resources in place, the companies could feel confident that they were ‘seeing all the order flows’ in the markets, and could use the information so derived to manage their positions and inventories. These included ‘proprietary trades’, in which the company was investing some of its capital, not just to support a client, but for its own reward.

**Ten banks dominate markets**

The result was that market capacity increased significantly. In 2006, the volume of global capital market new issues reached a record of US$14.7 trillion (US$10.2 trillion of securities new issues and US$4.5 trillion of syndicated bank loans).

The top 10 global capital market banks then accounted for 94% of these and other investment banking transactions (the top 5 had 57%). The top 10 banks in 2006 comprised the 3 large US banks, 4 large stand-alone US investment banks (Goldman Sachs, Morgan Stanley, Merrill Lynch and Lehman Brothers) and 3 European universal banks (UBS, Deutsche Bank and Credit Suisse).

Since then, the financial crisis triggered the acquisitions of Merrill Lynch by Bank of America, of the US business of Lehman Brothers by Barclays Bank, and of Bear Stearns by JP Morgan. The top 10 companies managing capital market transactions of various types in 2012 are shown in Exhibit 13.2. In that year of reduced capital market activity, the top 5 companies accounted for 49% of the total market and the top 10 for 81%, well below the market shares of the leaders in 2006.
## Exhibit 13.2

### Global capital market share

<table>
<thead>
<tr>
<th>Company</th>
<th>Rank 2012</th>
<th>Rank 2011</th>
<th>Global debt</th>
<th>Global equity</th>
<th>M&amp;A advisory</th>
<th>Syndicated bank loans</th>
<th>Total</th>
<th>Market share</th>
<th>Total cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>JP Morgan Chase</td>
<td>1</td>
<td>1</td>
<td>425,237.4</td>
<td>47,266.1</td>
<td>440,351.1</td>
<td>355,807.0</td>
<td>1,268,661.6</td>
<td>11.9%</td>
<td></td>
</tr>
<tr>
<td>Goldman Sachs</td>
<td>2</td>
<td>3</td>
<td>222,710.8</td>
<td>55,946.6</td>
<td>700,391.8</td>
<td>42,554.5</td>
<td>1,021,603.7</td>
<td>9.6%</td>
<td></td>
</tr>
<tr>
<td>Citigroup</td>
<td>3</td>
<td>4</td>
<td>304,256.5</td>
<td>53,115.5</td>
<td>441,518.4</td>
<td>214,268.8</td>
<td>1,013,159.2</td>
<td>9.5%</td>
<td></td>
</tr>
<tr>
<td>Bank of America/ Merrill Lynch</td>
<td>4</td>
<td>2</td>
<td>291,905.4</td>
<td>46,798.4</td>
<td>349,585.6</td>
<td>298,284.2</td>
<td>986,573.6</td>
<td>9.2%</td>
<td></td>
</tr>
<tr>
<td>Barclays</td>
<td>5</td>
<td>6</td>
<td>341,641.5</td>
<td>34,811.1</td>
<td>412,989.0</td>
<td>108,423.0</td>
<td>897,864.6</td>
<td>8.4%</td>
<td>48.6%</td>
</tr>
<tr>
<td>Deutsche Bank</td>
<td>6</td>
<td>7</td>
<td>354,800.1</td>
<td>37,196.1</td>
<td>387,402.8</td>
<td>92,092.5</td>
<td>871,491.5</td>
<td>8.2%</td>
<td></td>
</tr>
<tr>
<td>Morgan Stanley</td>
<td>7</td>
<td>5</td>
<td>206,839.2</td>
<td>52,188.1</td>
<td>504,355.2</td>
<td>61,160.7</td>
<td>824,543.2</td>
<td>7.7%</td>
<td></td>
</tr>
<tr>
<td>Credit Suisse</td>
<td>8</td>
<td>8</td>
<td>192,293.5</td>
<td>36,171.0</td>
<td>380,131.3</td>
<td>73,187.7</td>
<td>681,783.5</td>
<td>6.4%</td>
<td></td>
</tr>
<tr>
<td>UBS</td>
<td>9</td>
<td>9</td>
<td>139,392.7</td>
<td>31,544.2</td>
<td>219,282.9</td>
<td>26,106.6</td>
<td>416,326.4</td>
<td>3.9%</td>
<td></td>
</tr>
<tr>
<td>BNP Paribas</td>
<td>10</td>
<td>10</td>
<td>175,631.1</td>
<td>3,118.8</td>
<td>109,126.8</td>
<td>71,523.2</td>
<td>359,399.9</td>
<td>3.4%</td>
<td>81.3%</td>
</tr>
<tr>
<td>Industry total</td>
<td></td>
<td></td>
<td>2,654,708.2</td>
<td>398,155.9</td>
<td>3,945,134.9</td>
<td>1,343,408.2</td>
<td>8,341,407.2</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

*Source: Thomson Financial 1.973.622.5200 date: 2 January 2013*
Cataclysm

Many factors contributed to the massive market collapse of 2008 that resulted in extensive government intervention to contain a systemic failure that was unprecedented in modern financial history. These factors included a period of relaxed bank regulation that relied on market forces to contain excesses, a prolonged period of low interest rates after the bursting of the technology stock bubble in 2000, a major shift in demand for relatively high-yielding mortgage-backed securities, and a variety of turbo-charging mechanisms devised by global banks to meet this demand.

The market for mortgage-backed securities began to soften in early 2007 as house prices, unexpectedly, began to decline. The market deteriorated slowly for more than a year before the decline was sharply accelerated.

In May 2008, the Bank of England, in a Financial Stability Report, suggested that write-offs, under pressure from auditors, might have been excessive. Using fairly pessimistic assumptions about defaults and loss rates, the bank determined that the US$900 billion of mortgage-backed securities it studied had a fundamental worth of 81% of par value, yet market prices (determined by credit default swap prices) were then trading at only 58%. The bank predicted that arbitrage trading would buoy the falling prices, but this did not happen. The market panic related to mortgage-backed securities, indeed, proved to be far greater than the reality. At no point in the financial crisis, according to Standard & Poor’s, did the ‘default or near-default rate’ for any of its investment grade structured finance products exceed 8.5%, or 2% for any of its AAA rated products.

The sliding market precipitated the seizure of US mortgage financing giants Federal National Mortgage Association (FNMA) and Federal Mortgage Acceptance Corporation (FMAC) by the US Treasury in 2008 and caused further write-offs by banks and other financial institutions. These seizures were closely followed by the bankruptcy of Lehman Brothers and the bailout of AIG. These latter events led to a massive collapse of all financial markets in the US and elsewhere.

As much as the market failure of 2008 has been attributed to misconduct and misjudgments by banks and their regulators, equal attention must be given to the avalanche-like effects of the post-Lehman market collapses.

The focus of this chapter is on the institutional effects on the global capital markets industry in the aftermath of the worst global financial crisis since the 1930s.

Regulatory responses

After most financial crises, market corrections return the system to equilibrium before serious regulatory changes begin to be contemplated or put into place. After the stock market crash of 1929, and the subsequent Great Depression, a major overhaul of banking and securities regulations occurred in 1933 to 1934. The bursting of the technology bubble in 2000 inspired the Sarbanes-Oxley Act of 2002, but otherwise, previous crises did not generate important regulatory reforms.

This time, however, major reforms were enacted in the US and elsewhere in global financial centres.
Future

Basel III

Soon after the crisis began, the Basel Committee on Banking Supervision, an organisation of central bankers from 27 subscribing countries, met to re-examine their standards of risk-adjusted capital adequacy for banks. The committee was formed in the 1980s to agree a common standard of minimum capital adequacy for banks competing in global markets to minimise risk to the global financial system. These standards had been revised prior to the crisis, but had not been adequate to prevent the systemic collapse that indeed occurred. Hurriedly the committee met to revise the standards to toughen them up (Basel III). Risk adjustments would be more severe, and the percentage of risk-weighted assets to be covered by capital would be increased significantly. Further, additional new standards would be introduced to limit leverage and liquidity risks. These measures would be implemented gradually over a seven-year period ending on 1 January 2019 when total capital requirements (in two ‘tiers’) would increase from 8.0% of risk-weighted assets to 10.5%. At least 6% of the 10.5% would have to be in the form of Tier-1 common equity.

As at 31 December 2012, the top 10 capital market banks had average Tier-1 capital ratios – scored according to Basel I – of 15.7% (all but one had ratios above 12%). These ratios were roughly equivalent to 8% to 9% under the stricter standards of Basel III. All of these banks wanted to demonstrate that they were fully capitalised under current and expected regulations, and to satisfy debt markets that they remained good credit risks, despite the difficult circumstances they were in.

G20 bank standards

In 2010, the G20 group of nations (of which the US, Japan and the largest European and emerging market countries are members) created a Financial Stability Board. In November 2011, the Board designated 29 global banks as being so important to the global financial system that they must hold more capital than others. Of the banks listed, 17 are from Europe, 8 are US banks and 4 are from Asia.

The G20 also endorsed a core capital requirement surcharge (over the Basel III requirement) – starting at 1% of risk-weighted assets and rising to 2.5% for the biggest banks – which must be phased in over three years from 2016. This surcharge is called ‘Basel 2.5’.

Dodd-Frank

Dodd-Frank was signed into law in August 2010. It is a massive undertaking (848 pages; Glass-Steagall was 37 pages) that has been described as reflecting the opinions of outraged members of Congress more so than those of the Treasury or Federal Reserve. It was intended to address and contain systemic risk in the financial system, but the bill went well beyond that. The law essentially is a set of instructions to regulatory agencies to write about 400 new rules for financial markets. By the end of 2012, two and a half years after the bill was passed, only about a third of the new rules, involving approximately 9,000 pages of new regulations, had been completed and the process of implementing the law was way behind schedule. The new rules will replace or extend existing ones, add additional regulations
and overlap extensively with the other financial regulatory bodies (see Exhibit 13.3). By all accounts Dodd-Frank will take several additional years to implement in full, and the cost to the banks of both implementing and complying with it will be very high, though how high no one presently knows.

**Exhibit 13.3**

**The reach of Dodd-Frank**

![Diagram of financial agencies and lines of reporting]

Financial agencies:
- Old
- New
- Old with new powers
- Affected parties

Lines of reporting:
- Can request information
- Has authority to examine

Source: The Economist, ‘Too big not to fail’, 18 February 2012

Dodd-Frank addresses systemic risk in several ways. It designates all US banks with assets of more than US$50 billion (about 30) as ‘systemically important’ and requires the new Financial Stability Oversight Council (FSOC), to which all other financial regulators are meant to report) to designate those ‘non-banks’ that are to be regarded as systemically important. As at December 2012, these non-bank designations had not been made.

All systemically important banks and non-banks are to be subject to a requirement to hold more capital against losses than other banks – premium amounts presumably set by Basel III and the G20 requirement. All banks, however, are to be subject to the Volcker Rule (a part of the Dodd-Frank law which prohibits ‘proprietary trading’ and related activities); the actual rules governing the Volcker Rule had also not yet been released in December.
2012, though they were scheduled by Dodd-Frank to become effective in July 2012. This provision has involved extensive efforts by the banks to persuade the Federal Reserve, and other regulators, to moderate its terms so as not to prevent the banks from performing their traditional market-making functions for clients.

Similarly rules requiring the trading of derivative contracts on exchanges or clearinghouse platforms were still under development in December 2012. These rules will apply to about 70% of all derivatives trades and most likely will subject banks to margin requirements they did not have to meet while dealing entirely over-the-counter.

Dodd-Frank did not seek to break up the largest banks (though that would have been difficult with these banks on average holding assets of more than US$1 trillion; they would have to be broken up into 20 or more separate banks that would be small enough not be systemically important, that is, with more than US$50 billion in assets). Neither did the law attempt to re-impose the Glass-Steagall limitations on dealing in securities, which also would be difficult to enforce when banks can operate freely outside the US.

But Dodd-Frank does impose a system for intervention and ‘risk mitigation’ to be undertaken by the FSOC, though bailouts as practiced in the past are no longer to be permitted. The risk mitigation process is cumbersome – it requires an official designation of a bank as being a ‘grave threat’ to financial stability (requiring a majority vote of the FSOC), which can be challenged in court and appealed, followed by one or more mitigation efforts imposed by the FSOC to be implemented when and as possible. The whole process could take many weeks, but once a bank has been designated as a ‘grave threat’, investors in the designated bank (and others like it) will immediately take flight, fearing they will not be protected. In other words, the mitigation process may actually create the run on the bank that the FSOC is trying to prevent. But once the run comes, what can it do but take over the bank through the FDIC?

The Swiss ‘finish’

After bailing out UBS, which it did reluctantly, the Swiss government formed a ‘Committee of Experts’ to advise it on the ‘too big to fail’ problem. The government, as in the UK, was concerned that the country’s largest banks had assets well in excess of the Swiss GDP and Swiss taxpayers were unwilling to guarantee such a large and disproportionate amount under too-big-to-fail conditions. Made up of top regulators, bank executives and other industry experts, the Committee announced in October 2010 that UBS and Credit Suisse, its two largest banks, must hold almost twice as much capital as set out in the Basel III standards. The two banks must hold at least 10% of risk-weighted assets in form of common equity (Tier-1) and another 9% (Tier-2), which could be ‘contingent’ bonds, taking the current total capital requirements to 19%.

Ring fencing

The UK, like the Swiss, also appointed an independent banking commission to advise the government on its exposures to too-big-to-fail situations. The combined assets of the four largest UK banks were several times the size of the GDP of the UK, according to the
Commission’s chairman, Sir John Vickers, and thus exposed British taxpayers to considerable expense if any of the top four should become distressed. The Vicker’s Committee issued its report (which the government said it would adopt) in September 2011. Banks would be forced to ‘ring fence’ their domestic retail businesses to separate and protect them from their ‘casino’ investment banking arms. Business inside the fence would be eligible for financial assistance from the government if needed, but those outside would not be. The commission suggested that between one-sixth and a third of the £6 trillion of UK bank assets would be inside the fence, the rest not.

The Committee described the ring fence as ‘high’ and said that both the fenced and unfenced parts of the bank should have its own board of directors and be legally and operationally separate from one another. Similar to the Swiss rules, ring-fenced banks are to have a capital cushion of up to 20% – comprising Tier 1 equity of 10% together with an extra amount of other capital such as contingent capital bonds. The largest ring-fenced banks should have at least 17% of equity and bonds and a further loss-absorbing buffer of up to 3% if ‘their supervisor has concerns about their ability to be resolved without cost to the taxpayer’. Capital could be transferred from the ring-fenced bank to the investment bank only if the Tier-1 capital ratio of the ring-fenced bank did not fall below the 10% minimum after doing so.

Operations outside the ring fence will not be eligible for government assistance, and will have to secure financing for their lending and trading activities solely on their own ‘stand-alone’ credit. It is likely that such financing would depend on the non ring-fenced business itself being very well capitalised.

In October 2012, the ‘Liikanen Committee’,10 established by the European Commission (the executive body of the EU) to address systemic risk, issued a report that also recommended ring fencing as a way of isolating risk and requiring it to be better capitalised. In this case, however, the Commission recommended ring-fencing the riskier, that is, the investment banking, businesses to be sure these were contained in high capital requirement entities not eligible for bailouts. Similar proposals were introduced in France and Germany by the end of 2012.

**European Union regulations**

In November 2010, the EU established its own financial regulatory system, creating different units for banking, securities and insurance. The European Banking Authority sprang to life quickly, first organising ‘stress tests’ and then, much to the surprise of many, imposing its own minimum capital rules for the largest banks in the EU. These rules required banks to meet and maintain a 9% ‘core capital’ (similar to Tier-1) requirement by July of 2012. These are higher standards than Basel III and went into effect much sooner. The EBA said the raised capital buffers were intended to provide reassurance to markets about the ability of European banks to withstand a range of shocks and still maintain adequate capital.

In March 2013, 26 out of 27 finance ministers of the EU (the UK dissenting) voted to approve a comprehensive ‘rulebook’ for all 8,300 European banks that adopts somewhat revised Basel III capital standards to be phased in by 2018. The rule book requires the banks to hold sufficient liquid assets to survive a 30 day financial crisis, and restricts bonuses to
be paid to bank executives after January 2014 to no more than twice their salaries. The capital standards require all banks to maintain a 7% core tier-1 capital ratio, and globally systemic banks to maintain up to 9.5%. Eleven EU countries also approved a 0.10% tax on financial transactions in January 2013. Further, the EU has restricted activities of hedge and private equity funds, and has taken other measures to prevent financial markets from becoming excessively speculative in the future.

**Changing compensation practices**

Very soon after the crisis began, public sentiment turned against bankers who were thought to be responsible for the many post-crisis economic difficulties experienced in the US and in Europe. This sentiment was encouraged by opinions often expressed by respected officials and others without much supporting evidence that the ‘excessive’ levels of incentives for bankers to increase profits had been a principal cause of the crisis.

On the contrary, the incentive programs objected to, have required that a large portion of senior executives’ compensation be in options or shares of the common stock of their banks. All of the executives at senior management levels in all of the major banks had the vast majority of their personal wealth tied up in shares of stock of their own banks, which fell drastically in value after the crisis began. Some of them lost fortunes of several hundred million dollars. Many of these executives also lost their jobs as a consequence of the crisis and faced embarrassment and considerable personal litigation as well. All of these individuals were aware that excessive risk taking as a business strategy could result in devastating personal losses and, accordingly, may not have been as motivated to take the excessive risks their critics contend they were.

Nevertheless, the public anger over bank compensation practices was not satisfied by these examples, and it spread quickly to be taken up by regulators and legislators in various countries and forums.

In the UK, a one-time 50% tax on bank bonuses was applied in 2009, and in 2012 the EU proposed a stringent cap on bank bonuses. Similar efforts were suggested in the US, but not authorised by Congress. Dodd-Frank, however, does require that compensation programs receive regulatory approval, and be subjected to a non-binding shareholder vote (‘say on pay’). In 2012 the shareholders of Citigroup did not approve a compensation plan for its CEO, and the plan was withdrawn.

US regulators have established standards for bank executives that require that a substantial portion of bonuses paid to executives be paid in company stock, not cash, that the payments be deferred for several years, and be subject to ‘clawbacks’, or cancellation, in the event of misconduct.

The EU limitation of bonuses to no more than twice salaries (with shareholder approval) was greatly resisted by the UK government and the City of London, which feared that the rule would greatly impede London-based banks from competing for top people with US and Asian banks. The banks lost the battle, however, and would commence a no doubt lengthy period of lobbying, negotiating and manoeuvring so as to achieve a higher effective bonus to salary ratio, while recognising that the rule did help to lower their overall compensation costs and improve returns on capital.
Market corrections and other responses
The market effects of the financial crisis and the Great Recession that followed were felt in several different ways by global banks.

Bank stock prices collapse
The stock prices of the major capital market banks were shattered by the financial crisis and its very slow recovery. For the six-year period from December 2006 until December 2012, three US banks saw their stock prices drop by an average of 82%. The stocks of the five largest European capital market banks were closely packed around an average decline of 62% (Exhibit 13.4).

Exhibit 13.4
Declining stock prices of US and European capital market banks
(a) US banks
Markets recovered in 2009 from the worst of the crisis, also lifting several of the capital market banks’ stocks. Their shares, however, are still trading at near historically low levels relative to their book values. Usually, a bank’s book value (its net assets) is thought to be a fair representation of its liquidation value – it holds only financial assets for most of which ready markets exist – so a discount from book value suggests either that the bank’s valuation of its stated assets is incorrect, or that its ‘going concern’ business is in some sort of distress.

The average market price-to-book value ratio of the top 10 capital market banks declined during this period from 1.08 in December 2009 to 0.58 in December 2011, before recovering to 0.74 in December 2012.

The banks had, however, raised Tier-1 capital to an average of 15.3% by December 2012 to reassure markets of their solvency. But this increase in reserve capital, on which low returns were earned, together with protracted weak conditions in trading markets, and considerable restructuring and legal costs contributed to a sharp decline in returns on equity (ROE) by December 2012 to an average of only 3.3%, or 9.5% below the average cost of equity capital of these banks. This differential (ROE less cost of equity capital) is known as ‘economic value added’ (EVA) (Exhibit 13.5).
Exhibit 13.5

**EVA of global banks (31 December 2012)**

<table>
<thead>
<tr>
<th>Ranked by market capitalisation</th>
<th>Market capitalisation (US$ billion)</th>
<th>Market to book ratio</th>
<th>Return on equity</th>
<th>Beta</th>
<th>Cost of equity capital*</th>
<th>EVA RoE-CoEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>JP Morgan</td>
<td>167</td>
<td>0.86</td>
<td>11.00</td>
<td>1.32</td>
<td>9.49</td>
<td>1.51</td>
</tr>
<tr>
<td>Bank of America</td>
<td>120</td>
<td>0.55</td>
<td>1.27</td>
<td>2.34</td>
<td>15.49</td>
<td>-14.22</td>
</tr>
<tr>
<td>Citigroup</td>
<td>114</td>
<td>0.64</td>
<td>4.10</td>
<td>2.58</td>
<td>16.91</td>
<td>-12.81</td>
</tr>
<tr>
<td>BNP</td>
<td>77</td>
<td>0.67</td>
<td>8.90</td>
<td>1.53</td>
<td>10.70</td>
<td>-1.80</td>
</tr>
<tr>
<td>Goldman Sachs</td>
<td>60</td>
<td>0.88</td>
<td>10.70</td>
<td>1.39</td>
<td>9.87</td>
<td>0.83</td>
</tr>
<tr>
<td>UBS</td>
<td>60</td>
<td>1.16</td>
<td>-5.20</td>
<td>1.72</td>
<td>11.82</td>
<td>-17.02</td>
</tr>
<tr>
<td>Barclays</td>
<td>53</td>
<td>0.60</td>
<td>-1.90</td>
<td>2.66</td>
<td>17.39</td>
<td>-19.29</td>
</tr>
<tr>
<td>Deutsche Bank</td>
<td>42</td>
<td>0.57</td>
<td>1.10</td>
<td>2.22</td>
<td>14.78</td>
<td>-13.68</td>
</tr>
<tr>
<td>Morgan Stanley</td>
<td>38</td>
<td>0.62</td>
<td>-1.27</td>
<td>1.56</td>
<td>10.88</td>
<td>-12.15</td>
</tr>
<tr>
<td>Credit Suisse</td>
<td>32</td>
<td>0.82</td>
<td>4.30</td>
<td>1.40</td>
<td>9.93</td>
<td>-5.63</td>
</tr>
<tr>
<td>Average</td>
<td>76</td>
<td>0.74</td>
<td>3.30</td>
<td>1.87</td>
<td>12.72</td>
<td>-9.43</td>
</tr>
<tr>
<td>Top 5 average</td>
<td>68</td>
<td>0.76</td>
<td>7.20</td>
<td>1.57</td>
<td>10.95</td>
<td>-3.75</td>
</tr>
<tr>
<td>Bottom 5 average</td>
<td>83</td>
<td>0.74</td>
<td>-0.28</td>
<td>2.05</td>
<td>13.78</td>
<td>-14.05</td>
</tr>
</tbody>
</table>

*Cost of capital = risk free rate + (equity risk premium x company beta).
31 December 2012, ERP = 5.78%, 10-yr. UST = 1.86%.

*Source: Author’s own*

The average EVA for the top 10 banks has been negative for all of the 12 quarters since December 2009, when it was -4.0. EVA has varied widely among the banks, however. The 5 banks with the lowest EVAs averaged -14.1% at December 2012, and the 5 highest averaged -3.8%.

*Management changes*

The first response to a crisis that seriously affects stock prices is often a change in top management. Between 2007 and 2012, 7 of the top 10 global banks made CEO changes, four of which were made ‘effective immediately’. Two of the banks (plus Merrill Lynch, now part of Bank of America) changed their CEOs twice. Four of them replaced CEOs with trading backgrounds with executives with little trading experience who announced intentions to scale back trading as a percentage of the total company’s business.
**Reduced capital market activity**

There was also a sharp reduction in demand for capital market services: global new issues of debt and equity, mergers and acquisitions, and secondary market trading declined by about 30% after 2007, through the end of 2012 (Exhibit 13.6).

**Exhibit 13.6**

**Value of capital market transactions (US$ trillions)**

<table>
<thead>
<tr>
<th></th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>Percentage decline from peak year</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Debt new issues</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td>4.20</td>
<td>4.00</td>
<td>2.60</td>
<td>3.00</td>
<td>3.00</td>
<td>2.50</td>
<td>2.70</td>
<td>35.7%</td>
</tr>
<tr>
<td>International</td>
<td>5.60</td>
<td>4.60</td>
<td>4.30</td>
<td>4.20</td>
<td>3.30</td>
<td>3.20</td>
<td>3.70</td>
<td>33.9%</td>
</tr>
<tr>
<td>Total</td>
<td>9.80</td>
<td>8.60</td>
<td>6.90</td>
<td>7.20</td>
<td>6.30</td>
<td>5.70</td>
<td>6.40</td>
<td>34.7%</td>
</tr>
<tr>
<td><strong>Equity new issues</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td>0.23</td>
<td>0.29</td>
<td>0.26</td>
<td>0.3</td>
<td>0.29</td>
<td>0.21</td>
<td>0.27</td>
<td>6.9%</td>
</tr>
<tr>
<td>International</td>
<td>0.21</td>
<td>0.42</td>
<td>0.36</td>
<td>0.49</td>
<td>0.46</td>
<td>0.3</td>
<td>0.37</td>
<td>11.9%</td>
</tr>
<tr>
<td>Total</td>
<td>0.44</td>
<td>0.71</td>
<td>0.62</td>
<td>0.79</td>
<td>0.75</td>
<td>0.51</td>
<td>0.64</td>
<td>9.8%</td>
</tr>
<tr>
<td><strong>Total new issues</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.24</td>
<td>9.24</td>
<td>7.5</td>
<td>8.07</td>
<td>7.04</td>
<td>6.84</td>
<td>6.98</td>
<td>31.8%</td>
</tr>
<tr>
<td><strong>Global M&amp;A</strong></td>
<td>2.08</td>
<td>2.73</td>
<td>1.76</td>
<td>1.1</td>
<td>1.48</td>
<td>1.54</td>
<td>1.38</td>
<td>49.5%</td>
</tr>
<tr>
<td><strong>Value of shares traded</strong></td>
<td>67.46</td>
<td>98.82</td>
<td>80.52</td>
<td>80.42</td>
<td>63.96</td>
<td>66.42</td>
<td>32.8%</td>
<td></td>
</tr>
</tbody>
</table>

**Source:** Author's own

Part of the reduction in capital market activity is explained by the decline in new issues and mergers by banks and other financial services organisations as principals, and by the approximately 80% decline in new issues of mortgage-backed securities after 2008 as compared to its peak year in 2006, in which US$3.3 trillion of such debt was issued.

**Increase in bank funding spreads**

The funding costs for global banks increased considerably after 2008. This is indicated by the widening of five year credit default swap (CDS) spreads from as little as 10 to 20 basis points in 2007 (reflecting too-big-to-fail assumptions) to a peak of over 1,200 basis points.
in late 2008, after which they settled into the area of 300 basis points at the end of 2012 for the weaker banks and 175 basis points for the stronger, reflecting much greater disbelief in the likelihood of federal support that would bailout bank creditors in the future (Exhibit 13.7).

Exhibit 13.7

**Increased bank funding spreads**

Source: Goldman Sachs

**Downgraded credit ratings**

This concern was also reflected in the deterioration of credit ratings issued for banks by the major credit rating agencies. These agencies progressively downgraded the ratings of the largest banks after the crisis. On 21 June 2012, Moody’s announced that it had lowered its ratings of the largest capital market banks at their holding company levels, effectively dividing them into three tiers, because of ‘more fragile funding conditions, wider credit spreads, increased regulatory burdens and more difficult operating conditions... that together with inherent vulnerabilities... and opacity of risk, have diminished the longer term profitability and growth prospects of these companies’.11
The first tier (senior debt ratings between Aa3 and A2) – HSBC, Royal Bank of Canada and JP Morgan Chase – was seen to have strong capital buffers from their non-capital market activities and a record of good risk management.

The second tier (senior debt ratings between A2-A3) – Barclays, BNP-Paribas, Credit Suisse, Deutsche Bank, Goldman Sachs and UBS – had high contributions from trading.

The third, and weakest, tier (senior debt Baa1-Baa2) – Morgan Stanley, Citigroup, Bank of America – Moody’s said had experienced problems of risk management and volatility.

Moody’s also awarded ‘standalone’ ratings of two to three notches lower for all of these banks, reflecting their credit positions without regard for possible government support or assistance.

**Liquidity squeeze in Europe**

As a result of growing concerns about exposures to European sovereign credits and to other banks similarly exposed, interbank credit markets began to resist European bank paper. Also, money market funds in the US that typically rolled over substantial quantities of European bank certificates of deposit (which paid a higher rate than comparable US banks) began to liquidate their positions. European banks keenly felt these market pressures. Accordingly, Mario Draghi, who replaced Jean-Claude Trichet as President of the European Central Bank on 1 November 2011, declared an ‘unlimited’ access to ECB funds by European banks for up to three-years at low rates. This action, considered bold and controversial because of the open-ended nature of the commitment, clearly established the ECB as Europe’s lender of last resort for banks. On 21 December 2011, an auction was held in which 523 banks borrowed €489 billion; a second auction was held on 29 February 2012 and 800 banks borrowed €530 billion. These actions substantially reduced the borrowing pressures for banks in Europe, but increased the assets on the balance sheet of the ECB from €1.3 trillion in January 2008 to €3 trillion in February 2012.

**Legal settlements and reputation loss**

More than four years after the crisis, President Obama announced (in his 2012 State of the Union address) the formation of a special task force to work with enforcement officials of the States to ‘bring to justice’ those banks and others involved with financial fraud during the crisis. Regulators were active on many fronts, and brought charges against all the major banks, which were settled out of court for large amounts. There were several suits for securities fraud totalling US$2-US$3 billion, and a major, US$25 billion, settlement with the five largest US mortgage servicers for fraud related to mortgage processing and foreclosures.

In addition, the Federal Housing Finance Authority, the regulator of FNMA and FMAC, and others (regulators, prosecutors, investors and insurers) also sued 17 large banks in 2011 for US$200 billion, claiming that they defrauded the mortgage finance giants into buying US$1 trillion of faulty mortgage-backed securities. The plaintiffs claimed that the banks underwrote securities that were backed by mortgages that did not meet the standards described in the offering materials, and therefore they should be allowed to ‘put back’ the deficient securities.
at their original cost to the banks. Put back claims were estimated by some analysts to be as large as US$300 billion, though most believe the settlement amounts would be less. 13

European banks were also exposed to some of these and other charges related to activities in the US. UBS settled a criminal charge of aiding its clients in evading US taxes for US$780 million in 2009, and the Department of Justice then turned its attention to Credit Suisse and other Swiss banks, which it claimed did the same thing. UBS, Royal Bank of Scotland, and Barclays paid US$2.5 billion in settlement of charges by the US Commodities Futures Trading Commission and its European counterparts of rate rigging in the Libor market, for which several other banks are expected also to be charged. In addition, Barclays also paid US$2.5 billion for violating US anti-money laundering laws, and US$450 million to settle charges brought by the US Federal Electric Regulatory Commission in 2012.

Further, Barclays and other UK banks were expected to have to pay about US$30 billion to settle claims for fraudulent sales of payment protection insurance, a retail product sold in the UK.

Altogether, these various actions are expected to result in legal costs to the banking industry of US$100 billion to US$150 billion dollars, an unprecedented sum that would involve substantial write-downs of capital that would have to be replaced.

Diagnostics

By the end of 2012, three important elements had begun to bring the future of the global banking industry into focus.

A protracted downcycle

The crash in September 2008 opened the door to the Great Recession, which since then has limited US GDP growth to a five-year average of less than 1%. This prolonged and not yet ended period of slow growth, sustained by consumer anxieties, regulatory uncertainties, fears of European economic difficulties spreading to America, and concerns about the US fiscal position, has been poisonous to investment returns, with the S&P 500 remaining virtually flat over the five-year period during a time of decreasing fixed income yields.

All of this produced occasions when levels of stock market volatility were higher than at any time in the past 20 years. In October 2008, the volatility index (VIX) spiked at 80%, in March 2009 it was 50%, and in May 2011 and September 2012 it spiked again at 40%. The 20-year average VIX is about 18%. Considering that the value of securities outstanding in global markets now exceeds US$225 trillion, the magnitude of financial assets subject to fear and panic has never been higher. A sudden change in investor attitudes of just 5% (well less than average volatility levels) could release financial flows of US$11 trillion onto secondary markets, causing major, powerful shifts in liquidity that can affect prices correspondingly. Liquidity affects are felt not just in stocks but also in all financial assets.

Bankers have referred to this period as one of cyclical downturn, from which a recovery to ‘normal’ (that is, what it was before 2006) is expected to accompany a return to economic growth rates of 3% to 4%. Capital markets have always been cyclical, but rarely – not since the 1930s – have they been stuck in a slump of such an extended period.
The loss of lucrative fee income from merger advisory services, equity underwriting, and derivatives trading has been considerable during this period, but this has been accompanied by a substantial reduction of trading income, particularly in fixed income areas where, on average, the top 10 banks had allocated about 50% of their capital.\(^{14}\)

According to one report, the estimated global investment banking revenue pool from all transactions declined by 33% from US$358 billion in 2009 to US$240 billion in 2012.\(^ {15}\)

In response to these difficult conditions, US banks have hunkered down, cut costs, reduced leverage and risk-weighted assets and increased capital reserves in order to ride out the storms. European banks have done the same, though more so because their capital positions were weaker and they were more exposed to European sovereign and bank debt.

**New structural constraints**

The second element is the need to comply with structural reforms imposed by regulators to take effect over the next seven years, during which Basel III and other rules will be implemented fully. Major banks are burdened by the prospects of complying with the great variety of new regulatory requirements imposed by Dodd-Frank, Basel III, EU and EBA requirements and special rules adopted by Switzerland and the UK and other countries. The regulations will require considerably higher capitalisation ratios than before the crisis, provide restrictions on leverage, liquidity, and certain previously important trading activities (proprietary trading, derivatives), and the need to comply with restrictions on compensation and other regulatory requirements.

These regulations will be accompanied by the need for extensive additional compliance and reporting systems that will involve considerable initial investment and ongoing expense to the banks.

These new requirements will certainly cause ROEs to be reduced to levels well below the 15% to 20% they were before the crisis. A 2011 study by Morgan Stanley and Oliver Wyman noted that as much as a third of peak year ROE was the result of high leverage levels that will be curtailed by Basel III, and that other regulatory factors will decrease ROE for major global banks to 5% to 10%, before strenuous management efforts to mitigate these to a more tolerable 8% to 12%.\(^ {16}\)

Such mitigation efforts (including cost reductions and improved uses of technology) together with cyclical recoveries, may improve returns to the 10% to 15% range, the study concludes, but this will depend on making major adjustments to fixed income activities and on rethinking basic business models.

Two 2012 studies by Alliance Bernstein also concluded that deleveraging, increased capital requirements, and decreased margins from trading would lower returns to 5% to 6%, and for these to be mitigated to achieve ‘reasonable’ returns (that is, 8% to 12%), ‘compensation expense of trading units must be reduced to approximately 40% of revenues and the amount of capital employed by the units must decline by approximately 30%’.\(^ {17}\)

These studies suggest that even after the return to normal levels of capital market activity, which may still be some distance ahead, ROEs of the capital market businesses of major banks may be limited to around 10%, a level that at present does not exceed the units’ cost of equity capital. If the downturn is prolonged for more than a year or two,
competitive pressures within the industry are likely to erode fees, commissions and trading spreads further, which could mean that ‘new normal’ levels of market activity might provide returns of less than 10%.

**The regulatory dilemma**

The third element is the uncertain and contradictory role of governments in seeking to prevent a future crisis.

This is immediately visible in the continuing efforts in the US and Europe to seek large legal settlements from banks for actions taken years ago by individuals no longer employed by them. These settlements, mainly intended to satisfy public demand for punishment, are paid from capital that banks must replace, even while they are adjusting to a regulatory regime that increases the amount of capital they must have. The broader dilemma, however, is that in order to protect the banks from threatening the financial system, regulators may be rendering them into weakened, economically non-viable entities that are more susceptible to failure in the future than they were in 2007.

The regulatory idea adopted by governments in the US and Europe is to prevent ‘taxpayer bailouts’ in the future by relying almost entirely on a higher level of capital reserves than was required before. The theory assumes that most of the losses incurred in the 2008 crisis were the result of insufficient capital buffers, poor risk management controls, and overaggressive lending and trading activities, motivated by excessive compensation incentives.

There is little evidence, however, to support these assumptions.

In 2008, the major banks were fully compliant with Basel I capital standards and some had implemented Basel II. These standards were based on what was known at the time as credit and market risks, but did not include provisions for liquidity risk.

The most distinguishing feature of the 2008 crisis is the massive withdrawal of liquidity in the mortgage financing markets that began slowly in 2007 and peaked after the Lehman bankruptcy. During this period when mortgage-backed securities rated AAA experienced a higher than usual default rate of about 2%, prices of the securities dropped by 50% or more, a clear market over-reaction that was unexpected and caused banks, then newly subject to ‘fair value’ accounting, to take massive write-offs.

In a global securities market with over US$200 trillion of capitalisation and a history of high volatility spikes, the problem was not inadequate management of normal risks (though there was some of this), but the unpreparedness for an sudden avalanche powerful enough to carry away liquidity in all markets for an extended time period.

The avalanche might have been less powerful if governmental actions had not contributed to the fear and uncertainty. Regulators did not object to continuous deregulation or large serial mergers by banks, did not object to large concentrations of mortgage securities on the balance sheets of banks, were inconsistent in their bailout policies (why Bear Stearns but not Lehman?), and in the end committed many trillions of dollars through central banks to support and stabilise markets after September 2008 that they were not prepared to inject before then.

Basel III is a revision of previous models; it requires capital for collateralised securities, but even the arbitrary doubling of total capital to be held against tightened risk-weighted
assets may still not be enough to address a future liquidity panic with substantial marketo-market losses, particularly if governments acerbate the crisis by their own, well-intended but mistaken actions.

Most of the rest of the regulatory requirements, including the Volcker Rule and other elements of Dodd-Frank, are expensive add-ons that do not directly affect systemic risk.

The most important regulatory change since 2008 is the raised capital requirements of Basel III that most of the time will be unnecessary, but all of the time will reduce ROE and encourage banks to game the system so as to mitigate the reduction by increasing de facto leverage.

Restrictions on incentive compensation plans have led to substantial deferrals of compensation payments (turning them into future contingent liabilities that can weaken future balance sheets), and have encouraged many talented executives to look for opportunities in hedge funds or private equity where they would not be subject to such heavy regulatory pressures.

The new rules substantially constrain profitability and the ability of banks to continue to offer wholesale financing services needed for economic recovery and sustainable growth. Half of the top 10 market leaders still have very highly negative EVAs, even before the full effects of Basel III have been felt. Some major banks have already indicated their intention to cut back or withdraw from this important market sector.

**Adapting to the new order**

Large institutions can be slow to adapt to weaknesses perceived by outsiders. Inertia can propel them to continue old strategies from better days that they are no longer able to execute. Such institutions either change themselves, are changed by challenges from activist investors seeking to break them up, or, in due course, face bankruptcy.

The global banking industry is at such a point, following a decade of unsatisfactory results for investors, where it can expect challenges from activist investors. Even the largest companies, with large market capitalisations that might be assumed to make them invulnerable to unsolicited takeovers, are exposed. Chase Manhattan Bank, Bank of America, Union Bank of Switzerland and National Westminster Bank are examples of large banks, stalled by low valuations and apparent lack of progress, that were taken over by much smaller rivals during the past 20 years.18

Most of today’s top 10 banks have been committed to business strategies that would achieve rates of growth and returns on shareholder equity of 15% or more, provided by mergers, proactive trading activity, and aggressive expansion into new products and markets. Even without the challenges of a protracted downturn and significant regulatory tightening, some of these banks, having grown into enormous institutions of US$2 trillion or more of assets, have found achieving 15% growth very difficult to do in a global economy with a nominal growth rate of less than half that amount.

The Morgan Stanley and Oliver Wyman study concluded that investors are sceptical about the ability of current management of the global banks to work out their problems and re-engineer business models. Because of this, and their enormous size, the banks will continue to be considered potentially dangerous by their regulators, requiring strict controls and surveillance to protect against systemic failure.
So, the basic questions that boards of directors of the major capital market banks must ask themselves is whether the business strategies they have been pursuing for many years are still appropriate – that is, can they execute them effectively – and are they sustainable. If not, what should replace them?

**Break-up the bank**

There is a huge difference between the economic and cultural underpinnings of consumer/business banking and global investment banking, and little evidence that the two benefit much from being joined together beyond a minimal degree necessary to accommodate execution of transactions for clients.

Further, managerial skills and practices necessary to sustain a large capital market bank are very different from those of a large retail organisation. When banks have acquired investment banking units they place them under the supervision of executives with broad managerial experience but unskilled in the trenches of capital market combat where things are very different. Given the fast pace of activity, the extreme competitive pressures, and the challenges of risk retention and management, it is not surprising that significant losses and embarrassments occur. Indeed, there is considerable history of large banks stumbling after too much exposure to wholesale and investment banking.¹⁹

Some banks may benefit considerably by divesting themselves of all or most of their capital market activity. They would be guided by the contrasting experience of five of the largest global commercial and consumer banks, which though also suffering from a cyclical downturn, averaged in 2012 a positive economic value added of 3.48% and a market to book ratio of 1.45 (see Exhibit 13.8). (Two banks that might have been included in this survey, Santander and BBVA, have been excluded because of the special problems they are experiencing in the Spanish economy.)

Many regulators and academic observers would applaud a complete severance by major banks of their capital market activities, which they consider to be the riskiest part of their operations. Some banks, including Citigroup, have considered the idea carefully.²⁰ Doing so would substantially reduce risk-weighted assets, and would free the bank to be revalued by equity markets, no doubt at higher valuations.

There are problems with breaking up, however. First is the fact that the investment banking and securities units produce from one-third to two-thirds of major banks’ net income. Second is the fact that there are probably very few, if any, likely buyers of the investment banks in the current market. And third, separating the capital markets businesses would leave the new standalone investment banks (which are likely to be shorn of their banking licenses) vulnerable to further downgrading by credit rating agencies and difficulty in funding their activities in the market.

Shares in the investment banks could be distributed to bank shareholders but the market may not rise to welcome them until their ability to survive on their own could be established. Lehman Brothers was spun off to the shareholders of American Express in 1991, but it may be more difficult to repeat the action today. In any event, so far, no major bank has attempted to sell, spin off, or liquidate its capital markets subsidiary.
Future

Retreat from wholesale banking

Under considerable pressure from the Swiss government concerned about systemic risk, UBS, after several management changes, decided to shrink its investment bank considerably. In October 2012, UBS announced a major acceleration of its plan to reduce the size of its funded balance sheet by 300 billion Swiss francs by the end of 2015 and to exit most of its trading activities. UBS’s share price rose 60% from July, when the plan was first discussed, until the end of 2012. Some observers believe that a further de-emphasis of investment banking would allow UBS stock to be valued more by its valuable private banking and asset management businesses, which without its investment banking business, might be worth several times book value.

Barclays Bank has been under similar pressure from its regulators, investors and the British Parliament to improve its ‘culture’ of aggressive risk taking and profit seeking and to become a more ‘normal’ bank. In the summer of 2012, a new CEO, Anthony Jenkins, a retail banker, was appointed along with a new board chairman. Mr Jenkins promised a strategic review of the bank after six months, and in February 2013, he delivered his report. The old culture, he said, would be ‘shredded’, and a more compliant and less costly business model would replace it. Though he announced another round of layoffs in the investment bank, he was compelled to accept the fact that, because the Barclay’s capital markets unit contributed more than two-thirds of the bank’s profits, he could not cut it back quickly or even much further.

Mr Jenkins did not address a serious threat to his ‘clean up and preserve’ strategy: the almost certain imposition by the UK government of the ring fencing proposals that could

Exhibit 13.8

EVA of global retail banks (31 December 2012)

<table>
<thead>
<tr>
<th>Ranked by market capitalisation</th>
<th>MV US$ billion</th>
<th>Market to book ratio</th>
<th>P/E TTM</th>
<th>YTD ROE</th>
<th>Tier-1</th>
<th>Total assets</th>
<th>Beta</th>
<th>Cost of equity capital</th>
<th>EVA RoE-COC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSBC</td>
<td>194</td>
<td>1.06</td>
<td>13.80</td>
<td>8.40</td>
<td>13.40</td>
<td>2,693</td>
<td>1.31</td>
<td>9.43</td>
<td>–1.03</td>
</tr>
<tr>
<td>Wells Fargo</td>
<td>183</td>
<td>1.24</td>
<td>10.00</td>
<td>12.95</td>
<td>11.75</td>
<td>1,423</td>
<td>1.32</td>
<td>9.49</td>
<td>3.46</td>
</tr>
<tr>
<td>Royal Bank Canada</td>
<td>89</td>
<td>2.17</td>
<td>12.69</td>
<td>19.50</td>
<td>13.10</td>
<td>825</td>
<td>1.28</td>
<td>9.26</td>
<td>10.24</td>
</tr>
<tr>
<td>Toronto Dominion</td>
<td>76</td>
<td>1.31</td>
<td>12.28</td>
<td>14.39</td>
<td>12.60</td>
<td>807</td>
<td>1.68</td>
<td>11.57</td>
<td>2.82</td>
</tr>
<tr>
<td>Standard Chartered</td>
<td>63</td>
<td>1.47</td>
<td>12.14</td>
<td>12.80</td>
<td>13.4</td>
<td>637</td>
<td>1.56</td>
<td>10.88</td>
<td>1.92</td>
</tr>
<tr>
<td>Average</td>
<td>121</td>
<td>1.45</td>
<td>12.18</td>
<td>13.61</td>
<td>12.85</td>
<td>2,058</td>
<td>1.43</td>
<td>10.13</td>
<td>3.48</td>
</tr>
</tbody>
</table>

Source: Author’s own
put Barclays’ capital markets activities into an uncompetitive position, having to fund itself entirely in the markets without any expectation of assistance from either its parent or its government. However, ring-fencing is not expected until 2019 and as a result, Barclays has more time to see how things develop and make further changes if necessary.

In April 2012, Michael O’Neill, a tough-minded banking turnaround expert, became chairman of the board of Citigroup, replacing one of the last denizens of the Sandy Weill era, Richard Parsons. In October, Vikram Pandit, who replaced Charles Prince as CEO in 2007, was himself replaced by Michael Corbat, a 29-year Citigroup veteran who had been leading the effort to shed assets of Citi-Holdings. O’Neill and Corbat immediately began an accelerated effort to cut costs and further reduce Citi-Holdings, without offering any announcements about future changes to its business model, in which securities and wholesale banking accounted for 39% of profits for the fourth quarter of 2012.

Before 2008, when it acquired Countrywide Financial and Merrill Lynch, Bank of America was a very successful national retail banking organisation whose share price from 2000 significantly outperformed its large US rivals. The Countrywide acquisition, however, in terms of its legal and related exposures to the mortgage industry, was disastrous. In July 2012, Bank of America’s stock price traded at 0.41 times its book value, the lowest then of all US global banks.

The bank’s Merrill Lynch acquisition, on the other hand, though troubled at the beginning, proved to be successful in the sense that it contributed 36% of Bank of America’s 2012 revenues, and most of its profits in a year of continuing write offs in the mortgage sector. How Bank of America intends to change its strategy with respect to investment banking is unclear, though it has yet to demonstrate that the global capital markets business of Merrill Lynch (as opposed to its national retail brokerage business) is a manageable fit with its commercial and consumer businesses.

Re-engineer the trading model

For those banks determined to remain as market leaders in the global wholesale banking business, it will be necessary to rethink the role of trading in their business models, that is, to determine how much of their capital and other resources are to be committed to trading beyond what is necessary to conduct top of the line underwriting and advisory businesses.

Dodd-Frank is expected to disallow some of the proprietary trading that ‘flow traders’ routinely perform. Banks are still likely to use their balance sheets for selective ‘mandate seeking’ by offering to put up capital at competitive prices in exchange for the assignment to manage a merger transaction. Such actions can involve banks making bridge or other loans to be distributed or refinanced when market conditions permit. However, the capital costs of supporting large inventories of loans and securities purchased for such purposes will be economically challenging at best for systemically important companies under the new rules. These banks will have to rely more on their distribution capabilities rather than on their balance sheets in making markets for clients. They will have to take cost out of their trading businesses with technology (including greater use of electronic exchanges) and by eliminating low margin activities. They will also have to reduce compensation paid to traders – market-makers need not be paid as generously as proprietary traders, most of
which will be eliminated. New approaches to compensation that lowers its annual expense as a percentage of net revenues will be necessary to improve ROE.

Goldman Sachs is the most committed of all the top 10 banks to trading and the management of ‘alternative assets’. In 2012 it still had a substantial proprietary trading and investing business and owned hedge funds, private equity and real estate funds that would be disallowed by the Volcker Rule, though the company will have several years to disengage from them. This business has been a profitable source of captive trading and investment banking business for Goldman Sachs, which it is reluctant to give it up. Nevertheless, if its extensive alternative asset management business (which is comparable to Blackstone’s, an industry leader) could be operated as a non-systemic non-bank without the regulatory burdens that Goldman Sachs itself cannot escape, the company may decide that its shareholders would be better off if it distributed that business directly to them.

**Strategic mergers**

In 2010 James Gorman, a retail brokerage executive who joined the company from Merrill Lynch in 2006, succeeded John Mack, a fixed-income trading executive, as CEO of Morgan Stanley. The previous year, Morgan Stanley formed a joint venture with Salomon Smith Barney (of which initially it owned 51% and Citi-Holdings 49%), committing itself to acquiring, in stages, the portion still held by Citigroup over the next few years. Doing so would make Morgan Stanley one of the largest retail brokerage companies in the world, and reduce the concentration of the company’s investment banking and trading business to less than half. In September 2012, Morgan Stanley announced it would acquire a further 14% stake in the joint venture, and the rest of it over the following three years.21

The retail brokerage business is not as capital intensive as investment banking, so absorbing the joint venture could be beneficial to Morgan Stanley, but continuing to be ‘systemically important’ under Dodd-Frank could offset the benefits. (Possibly Morgan Stanley could give up its status as a Bank Holding Company, though this alone might not enable it to escape the capital requirements if it were to be designated a systemically important non-bank.) If the burden of being systemic is too great, Morgan Stanley may decide to separate the brokerage and investment banking businesses.

In 2010, Deutsche Bank announced the acquisition of a majority interest in Deutsche Postbank, the former government-owned postal savings bank, becoming Germany’s largest retail banking organisation. In 2012, it acquired most of the remaining shares in Postbank, bringing its ownership to 94%. Retail banking is a difficult industry in Germany because there are so many subsidised savings organisations owned and promoted by local and regional governments. Deutsche’s effort to enter this business, however, will increase its source of retail deposits, which can be used to fund its considerable European commercial and wholesale banking activity.

In 2012, the bank also installed two co-CEOs, Jürgen Fitschen, a German, to run the domestic businesses of Deutsche Bank, and Anshu Jain, an Indian-born UK citizen who supervises the investment bank. In 2012, the bank announced a comprehensive ‘2015+ Plan’, an initiative to lower costs, raise capital, and prepare for a more tightly regulated future. It formed a non-core businesses unit (for disposals), took substantial write-offs to reflect legal
charges and goodwill losses, and tightened its efforts to de-risk the bank, increase capital, and substantially lower operating costs, including executive compensation. Deutsche says it is committed to operating its two main units as a universal banking whole, but their structures and management are such that, if necessary, they could be separated fairly easily.

**Adapting to survive**

The investment banking business goes back to the 19th century when some of the present leading companies were established. Observers have noted that over the time the industry persists – capital has to be raised and invested – but individual companies come and go. In the 1930s, US banks were required by law to divest their securities units and the industry was changed radically as adaptations were made. In the 1960s and 1970s, technology developments and important regulatory changes occurred that forced companies to adapt again. The pattern of continuous adaptation has lasted until the present – today’s companies will have to adapt to regulatory changes as profound as those of the 1930s amidst global markets of enormous size and volatility.

Adaptations are uncertain events. Some first movers set the stage for others to follow, even though the success of the moves may be doubtful. Some companies will hesitate to change, either out of inertia or indecision, and they may suffer from their caution, or not, as only the future can reveal. But all have to think about how they might best adapt their particular businesses to the new conditions.

It may be possible to sustain negative EVA for a temporary period – even an extended period – of transition, but negative numbers point to non-viability in the long run and thus they must be addressed. A more optimistic outlook for improving economic conditions and the banks’ own restructuring efforts lifted all bank stock prices by more than 20% during 2012, but the future structural impact on bank stocks has to be respected, too.

All of the major banks are considering how they might adjust – some are waiting for improved markets to sell or spin off parts of their businesses, others are waiting for a more definitive understanding of the new rules before acting. There are indeed quite a few important rules that are still being waited for.

For the top 10 global capital market banks, the last four years have been a period of recovery and adjustment. Even so, total assets of the group have increased (not decreased) from an average of US$1.5 trillion to US$1.8 trillion while their Basel I Tier-1 ratios have increased from an average of 12.8% to 15.7%, suggesting a significant decline in risk-weighted assets. Banks’ total assets divided by book value of equity (gross leverage) declined from a ratio of 21 in 2009 to 17.5 in 2012, but total assets divided by market capitalisation has risen from 19 in 2009 to 24 in 2012. Banks are clearly searching in the realm of Basel III for low risk weighted assets that can be leveraged to the point of profitability, much as they did a decade ago (Exhibit 13.9).

Nevertheless, EVA has declined steadily since 2009. This is a condition banks cannot endure for very much longer without attracting activist shareholder attacks. All of the banks have struggled to overcome this condition through extensive cost cutting efforts, though in many cases it is clear that there is still a long way to travel to regain positive territory, and cost cutting alone may not be enough.
Adaptations are not made in a vacuum, however. Even as some banks withdraw some of their competitive energy from the markets, others will insert more in order to capture some of the market share that will become available. Those seeking to gain market share from this period of adjustment will include some large banks with lesser amounts of capital market participation, some boutiques and ‘shadow banking’ players (for example, private equity and hedge fund managers), but the large trading companies already in the markets will be formidable competitors as well.

It is difficult, however, in the light of the increased regulatory requirements and intense scrutiny under which the large, systemically important companies must operate, for them to revert to the aggressive, high-risk high-reward business strategies that characterised the pre-crisis period. Instead these companies will adjust expectations, as several have already begun to do to lower levels of growth and ROE, and with higher expectations of predictable cash flows with higher dividend payouts.

In this way, some of the systemically important companies can be expected to revert to business models of the 1960s, in which large banks accepted roles as important financial public utilities, while increasing amounts of risk is distributed outside the banking system.

---

2 The top 10 global investment banks have evolved into what they are today as a result of a long period of mergers and acquisitions, often of failing or weakened companies, in which approximately 40 US investment banks, 12
US money centre banks, 10 UK merchant banks and 10 European securities companies were absorbed into 10 main players.

An earlier banking crisis occurred in the US in 1984 after the collapse of Continental Illinois Bank, which was deemed to be too-big-to-fail by the Federal Reserve and the FDIC, which took it over after guaranteeing all depositors, lenders and bond holders. This action precipitated the recognition that many other large banks were undercapitalised and in difficulty, several of which were rescued by acquisition by banks from other states, which received waivers from applicable laws preventing inter-state banking. As banks recovered from the banking crisis of 1984–1994, they asked for relief from Glass Steagall to be able to compete more effectively. Paul Volcker chaired the Federal Reserve until 1987 when Alan Greenspan, who favoured a more open, competition-enhancing regulatory structure for banks, replaced him.


See standardandpoors.com: default rates for global structured finance products.


In January 2012, seeking to test the law, The Public Citizen, a public interest group, petitioned the FSOC to declare Bank of America, which it described as being in very tenuous condition for such a large bank, a ‘grave threat’ and begin mitigation steps. The FSOC has not responded to the petition.


Fixed income obligations that convert into common stock (or become worthless) if the issuer’s Tier-1 ratio falls below a pre-set limit.

Named for the governor of the Central Bank of Finland.

Moody’s Investor Service, ‘Announcement: Moody’s reviews ratings for banks and securities companies with global capital market operations’, 15 February 2012; and ‘Key drivers of rating actions on companies with global capital market operations’, 21 June 2012.

The government has not been successful in convicting any leading banking, investment banking, or mortgage-banking figures of criminal activity, though several efforts have resulted in acquittals. A few individuals have settled civil charges for cash amounts covered by insurance.


Chase Manhattan was acquired by Chemical Bank in 1991, Bank of America was acquired by Nations Bank in 1997, Union Bank of Switzerland was acquired by Swiss Bank Corp in 1998 and National Westminster Bank was acquired by Royal Bank of Scotland in 2000.

John Reed discovered this when he was CEO of Citicorp and shed its investment banking business in the 1990s before its merger into Travelers; American Express discovered it too after years of seeing its stock price dragged down by its ownership of Lehman Brothers. Barclays and National Westminster Bank also learned how difficult investment banking could be in London (Barclays withdrew from most of it until it decided to re-enter with its acquisition of the Lehman Brothers franchise in 2008, NatWest was so weakened by its experience that it was taken over by a smaller Royal Bank of Scotland, which subsequently failed and had to be taken over by the UK government).


After long discussions Morgan Stanley and Citigroup agreed on a valuation of the joint venture of US$13.5 billion and Citigroup announced it would take US$2.9 billion after-tax write-down from the transaction. After receiving approval from the Federal Reserve, Morgan Stanley accelerated its plans to acquire the 35% minority interest in the joint venture in June 2013.