Navigating Uncharted Waters

A roadmap to responsible innovation with AI in financial services

Part of the Future of Financial Services series

Prepared in collaboration with Deloitte

Foreword

Consistent with the World Economic Forum's mission of applying a multistakeholder approach to address issues of global impact, the creation of this report involved extensive outreach and dialogue with numerous organizations and individuals. These included the Forum's financial services, innovation and technology communities, as well as leaders from academia and the public sector. The outreach comprised more than 200 interviews and ten international workshop sessions, conducted over the past 20 months, with the aim of capturing insights into and opportunities relating to the impact of artificial intelligence on the financial services industry.

The holistic and global content of this report would not be as rich without the support of, and contributions from, the subject matter experts who assisted in shaping our thoughts about the future impact of AI on the financial services industry. In particular, we thank this project's Steering Committee and Working Group. Their expertise and patient mentorship have been invaluable. Also critical has been the ongoing institutional support for this initiative from the World Economic Forum and the leadership of our chairman, whose vision of the Fourth Industrial Revolution has been inspirational to this work.

Finally, we are grateful to Deloitte* for its generous commitment and support, in its capacity as the official professional services adviser to the World Economic Forum, on this project.

Contact

For feedback or questions,
please contact:

R. Jesse McWaters, Lead Author

Matthew Blake, Head of Financial and Monetary System Initiative
matthew.blake@weforum.org
+1 (212) 703-6621

* Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms. This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional adviser. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Cover image credit: "Monochrome Photo of Waves" by Matt Hardy on Pexels

Editor’s note

Artificial intelligence is a critical aspect of the Fourth Industrial Revolution; in recent years, during discussions at the World Economic Forum’s Annual Meetings, we have observed an exponential increase in both the interest in, and awareness of, the opportunity that AI presents for financial services. However, this excitement has been coupled with significant concern and uncertainty about the potential negative impacts of AI.

The World Economic Forum has a successful track record of providing detailed multistakeholder analysis of the changing landscape of the financial ecosystem, particularly through our *Future of Financial Services* series. We felt our approach could be effective in cutting through the sensationalism surrounding AI to provide valuable insights for the public and private sectors alike. In 2018 we published *The New Physics of Financial Services*, an analysis of the operational and competitive transformation AI is driving in the financial sector; that work is furthered in this document, with a detailed exploration of the risks inherent in deploying AI in the financial sector, as well as strategies for how those risks might be mitigated.

Contrary to popular opinion, we found that AI does not require the development of a “new ethics” for financial services. Instead, it demands an openness to new governance and regulatory approaches that are better suited to the needs of complex AI-enabled systems that seek to meet – and exceed – the same standards of fairness, stability, transparency and accessibility to which the financial sector has been held for years.

We hope that this document will be helpful to you and your institution as you seek to navigate the turbulent changes on the horizon, enabling you to manage the risks and opportunities presented by AI to the betterment of your customers and society at large.

R. Jesse McWaters

Project Lead, Future of AI in Financial Services
World Economic Forum

Rob Galaski

Partner, Deloitte Canada;
Global Leader, Banking & Capital Markets, Deloitte Consulting

Past reports from the *Future of Financial Services* series

The Future of Financial Services
How disruptive innovations are reshaping the way financial services are structured, provisioned and consumed

2015

The future of financial infrastructure
An ambitious look at how blockchain can reshape financial services

2016

A Blueprint for Digital Identity

2016

Beyond Fintech: A Pragmatic Assessment Of Disruptive Potential In Financial Services

2017

The New Physics of Financial Services

2018

The Next Generation of Data Sharing in Financial Services: How Privacy-Enhancing Techniques Can Unlock New Value in the Financial Services Industry

2019

Members of the Steering Committee

Josh Bottomley
Global Head of Digital,
HSBC

Nick Cafferillo
Chief Data & Technology Officer,
S&P Global

Vanessa Colella
Head, *Citi Ventures* and Chief Innovation Officer, *Citi*

Juan Colombas
Executive Director and Chief Operating Officer, *Lloyds Banking Group*

Robert Contri
Global Financial Services Leader,
Deloitte

David Craig
Founding Chief Executive Officer and Board Member, *Refinitiv*

Tony Cyriac
Enterprise Chief Data & Analytics Officer,
BMO Financial Group

Rob Goldstein
Chief Operating Officer & Global Head of BlackRock Solutions, *BlackRock*

Greg Jensen
Co-Chief Investment Officer,
Bridgewater Associates

Prof. Dr Axel P. Lehmann
President Personal & Corporate Banking Switzerland, *UBS*

Lena Mass-Cresnik, PhD
Chief Data Officer,
Moelis & Company

Max Neukirchen
Head of Strategy,
JP Morgan Chase

Kush Saxena
Chief Technology Officer,
Mastercard

Nicolas de Skowronski
Member of the Executive Board, Head of Advisory Solutions, *Julius Baer*

Michael Zerbs
Group Head & Chief Technology Officer,
Scotiabank

Members of the Working Group

Sami Ahmed
Vice-President & Head of Data, Analytics and AI Transformation, *BMO Financial Group*

Secil Arslan
Head of R&D and Special Projects, *Yapi Kredi*

Tim Baker
Global Head of Applied Innovation, *Refinitiv*

Beth Devin
Managing Director & Head of Innovation Network & Emerging Technology, *Citi*

Roland Fejfar
Head Technology Business Development & Innovation EMEA/ APAC, *Morgan Stanley*

Gero Gunkel
Group Head of Artificial Intelligence, *Zurich Insurance*

James Harborne
Head of Digital Public Policy, *HSBC*

Milos Krstajic
Head of AI Solutions, Claims, *Allianz SE*

Wei-Lin Lee
Senior Director, Strategy & Growth, *PayPal*

Juan Martinez
Global Head of APIs, Identity & Connectivity, *SWIFT*

Michael O'Rourke
Head of Machine Intelligence and Global Information Services, *Nasdaq*

Jennifer Peve
Managing Director, Business Development and Fintech Strategy, *DTCC*

Jim Psota
Co-Founder and Chief Technology Officer, *Panjiva (S&P Global)*

Nicole Sandler
Innovation Policy Global Lead, *Barclays*

Annika Schröder
Executive Director, Artificial Intelligence Center of Excellence, *UBS*

Chadwick Westlake
Executive Vice-President, Enterprise Productivity & Canadian Banking Finance, *Scotiabank*

Members of the Project Team

Project leadership

The Future of AI in Financial Services project leadership team includes the following individuals:

World Economic Forum

R. Jesse McWaters, Lead Author, Project Lead

Matthew Blake, Head of Financial and Monetary System Initiative

Professional Services Leadership from Deloitte Canada

Rob Galaski, Co-Author, Project Adviser

Project authors

The World Economic Forum expresses its gratitude to the following individuals on the project team:

Deloitte Canada

Ishani Majumdar, Senior Consultant

Hemanth Soni, Senior Consultant

Special thanks for contributions from:

Felix Mueller, Allianz

Verena Treber, Allianz

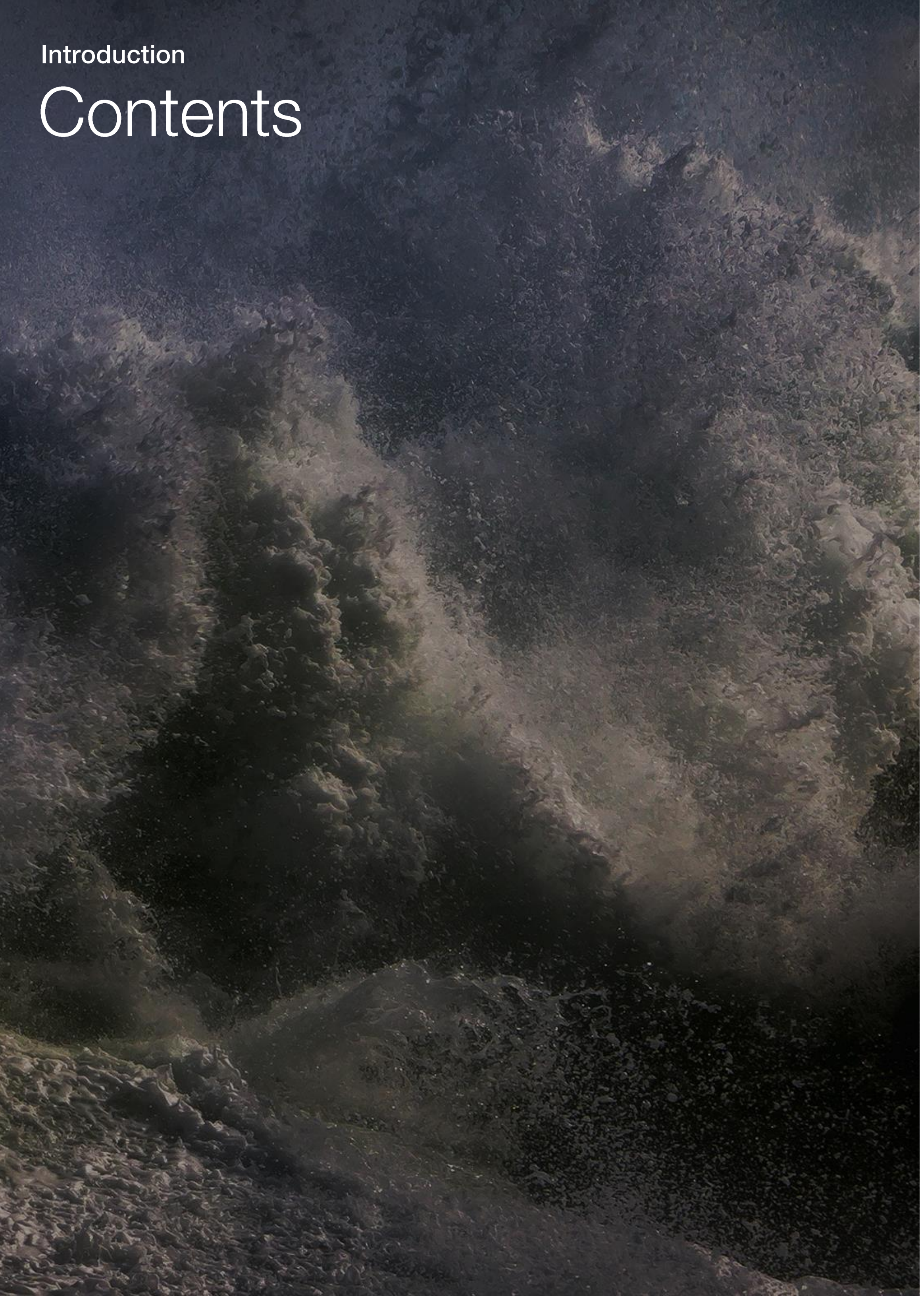
Ferdina Yarzada, Deutsche Börse

Additional thanks

The project team expresses gratitude to the following individuals for their contributions and support:

Derek Baraldi
Mary Emma Barton
Andre Belelieu
Kerry Butts
Alexandra Durbak
Natalya Guseva
Rouzbeh Hadjibaba
Mahmood Hassan
Kai Keller
Courtney Kidd Chubb
Abel Lee
Nicole Peerless
Alexandra Romic
Ryan Singel
Denizhan Uykur
Han Yik

Contents



Preface

Executive summary and key findings

Chapters

Vignettes

Conclusion

Context and approach

Last year's report in this series, *The New Physics of Financial Services*, explored the implications of AI for the financial ecosystem, raising questions for further examination

The New Physics of Financial Services explored three main facets of AI's impact:

New modes of operating

AI in financial services will make front- and back-office operations look radically different.

New market structures

AI in financial services will create major shifts in the structure and regulation of financial markets.

New societal challenges

AI in financial services will raise critical challenges for society to solve.

The report raised two questions that warranted further exploration:

1 Sharing data to unlock new value

How might financial institutions capitalize on the opportunity to use shared data to create new value for themselves, their customers, regulators and societies at large?

This question formed the basis of the following white paper:

The Next Generation of Data Sharing in Financial Services: How Privacy-Enhancing Techniques Can Unlock New Value in the Financial Services Industry

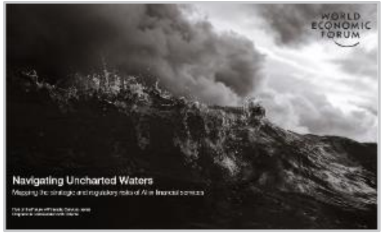
The Next Generation of Data Sharing in Financial Services

How privacy enhancing techniques can unlock new value in the financial services industry

2 The dilemmas of deploying AI responsibly

How can financial institutions ensure the responsible use of AI in the financial sector, reaping the benefits of new capabilities while effectively navigating the new risks it introduces?

This question formed the basis of this report:



Navigating Uncharted Waters

Mapping the strategic and regulatory risks of AI in financial services

Navigating Uncharted Waters

A roadmap to responsible innovation with AI in financial services

To explore the challenge of responsible AI, the World Economic Forum and Deloitte led one of the world's largest studies into the use of AI in financial services

250+ Contributions from subject matter experts and leaders across incumbents, innovators, academics and regulators

7+ Global workshops that brought together stakeholders from different backgrounds

10+ Months of extensive research

Working with leading incumbents...

SWIFT
RBS The Royal Bank of Scotland
Morgan Stanley
AlixPartners
Fidelity Investments
S&P Global
Rabobank
ING
Societe Generale
Bank of America
DTCC
Standard Chartered
Nordea
Microsoft
BBVA
UBS
Prudential
Nomura
J.P.Morgan
Barings
BlackRock
Swiss Re
BMO
Zurich
NYSE
BNP Paribas
Viking Global Investors LP
The Rock Creek Group
PZU
OakNorth
Safra
Credit Suisse

...with leading innovators and regulators...

omniscience
ethics.net Registered Member
mymoneybank
riskgenius
Cognitive Finance
Borealis AI
ComplyAdvantage
Temenos
WilmotML
aperio
Decipher
FINOS
ocean
Element AI
10x
Savvy
Plato
FSB Financial Stability Board
Luminoso
onfido
Socure
FCA Financial Conduct Authority
DFSA
ForwardLane
Immuta
sedicii
MAS

...and hosting interactive discussions in financial capitals around the world.

New York

London

Zurich

Washington DC

Davos

This report focuses on understanding the risks and governance requirements of AI in financial services through the lens of five frequently cited areas of concern

Throughout our research several thematic questions were frequently repeated:

Risks:
Where are the real risks? Which fears are overstated?

Best practices:
What are the best practices in governing AI systems?

Opportunities:
Where might the risks of AI be turned into opportunities?

Unknowns:
Which topics require further information and discussion?

These questions are particularly relevant for five concerns surrounding the use of AI in the financial sector:

AI explainability
How does business context shape what we need to know about our AI?

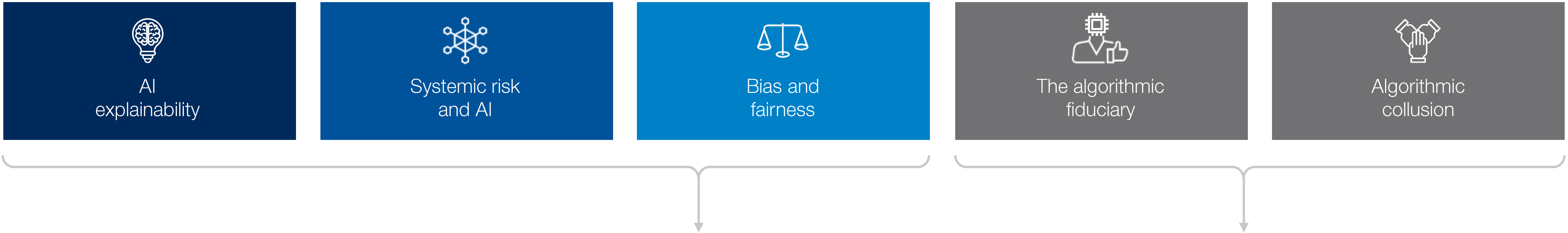
Systemic risk and AI
Could algorithms destabilize the financial system?

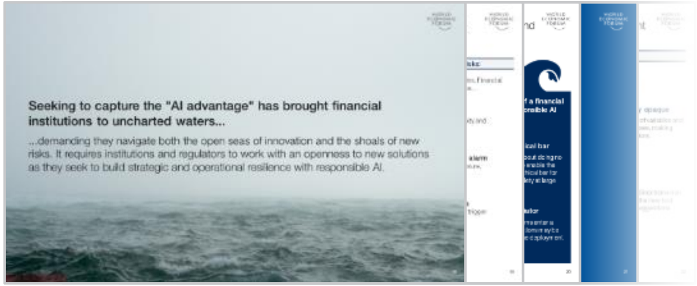
Bias and fairness
How can institutions ensure their systems do not discriminate against a specific group?

The algorithmic fiduciary
Could an AI be trusted as a fiduciary?

Algorithmic collusion
How can we manage AI systems that learn to engage in anti-competitive behaviour?

This report includes a cross-cutting executive summary and key findings, and an exploration of the key concerns about the use of AI in financial services





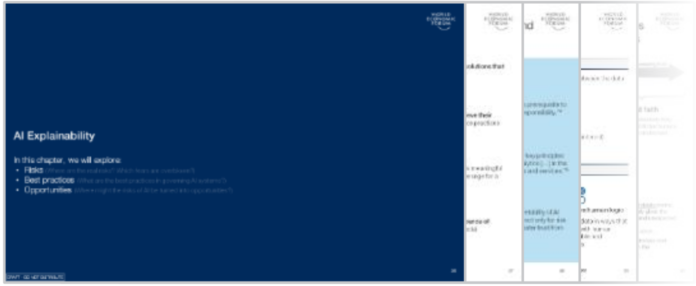
Seeking to capture the "AI advantage" has brought financial institutions to uncharted waters...

...demanding they navigate both the open seas of innovation and the shoals of new risks. It requires institutions and regulators to work with an openness to new solutions as they seek to build strategic and operational resilience with responsible AI.

Executive summary and key findings

(page 14)

Our synopsis of the evolution of governance as a result of institutions pursuing AI-enabled strategies



AI Explainability

In this chapter, we will explore:
- Risks (Where are the real risks? Which fears are overblown?)
- Best practices (What are the best practices in governing AI systems?)
- Opportunities (Where might the risks of AI be turned into opportunities?)

Detailed exploration of the priority concerns

(page 32)

Our exploration of the three highest-priority concerns for institutions/regulators, with insights on how these concerns might be addressed

The Algorithmic Fiduciary

In this vignette, we will briefly explore:
- Risks (Where are the real risks? Which fears are overblown?)
- Opportunities (Where might the risks of AI be turned into opportunities?)

Brief vignettes on other uncertainties

(page 99)

Our synthesis of the conversation surrounding two less commonly discussed uncertainties

This report aims to provide industry executives, regulators and policy-makers with context-specific frameworks for the responsible use of AI in financial services

✓

This report will…

- **Propose decision-making frameworks to address key concerns** surrounding the use of AI in financial services
- **Explore the strategic upside** of investments in "responsible" and "trust-first" AI business models
- **Highlight areas of regulatory uncertainty** where public/private engagement is needed to forge additional clarity

✗

This report will not…

- **Delve into the technical details** of how AI technologies or the associated tools (e.g. explainability algorithms) work
- **Outline implementation strategies** for how financial institutions can deploy new AI governance systems
- **Provide detailed recommendations** or suggest policy positions for specific financial institutions

This report seeks to help…

- **Strategic decision-makers at financial institutions** build a map of the risks of deploying AI in financial services, and mitigate those risks while capturing the opportunity of AI-driven business models and capabilities
- **Regulators and policy-makers** understand the new challenges of an AI-enabled financial ecosystem, and what responses are necessary to protect consumers, institutions and broader society alike

Executive summary and key findings

Seeking to capture the ‘AI advantage’ has brought financial institutions, regulators and policy-makers to uncharted waters...

...exposing the financial system to new hazards and posing uncertainties to AI adoption. Navigating safely will require stakeholders to work together, opening themselves up to new paradigms of governance and supervision to build a better financial ecosystem for all.

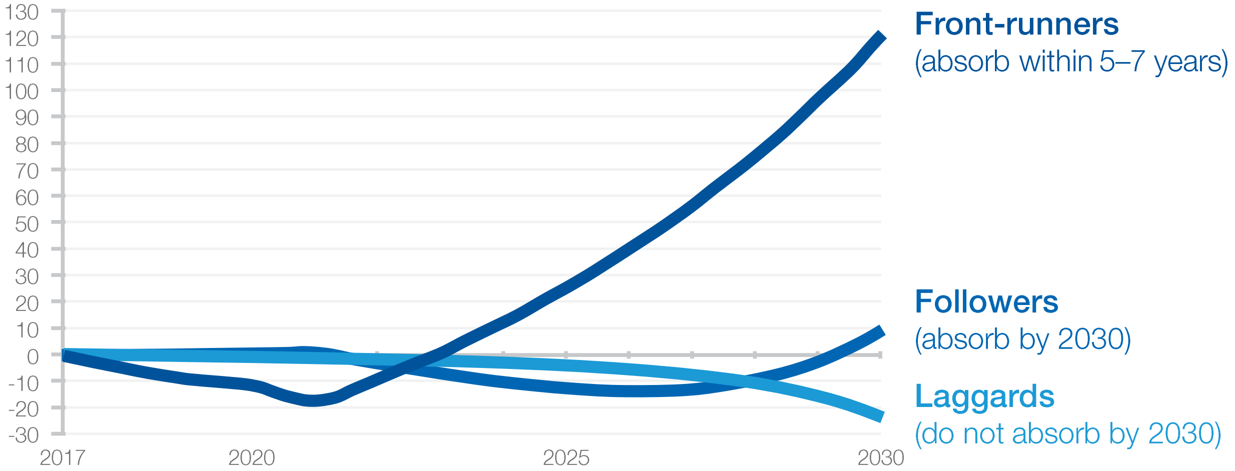
Institutions that succeed in the application of AI tend to do so by being early movers and establish a defensible competitive position – but being first comes with risks

Early AI adopters will reap an outsized share of the rewards…

Over the next decade, a remarkable gap is expected to emerge between the institutions that adopt and absorb AI quickly vs. those that follow or lag behind.

Relative changes in cash flow by AI adoption cohort[1]

% change per cohort, cumulative



As a result, financial institutions that innovate rapidly and move to implementation early have the potential to realize the greatest competitive gains.

…but being an early AI adopter comes with new risks:

Being first to market comes with a host of risks and uncertainties. Financial institutions pursuing revolutionary AI-enabled strategies will face…

The risk of customer backlash
from AI failures that damage brand equity and trust

The risk of triggering regulatory alarm
in the form of additional scrutiny or censure, and depleting goodwill

The risk of alienating employees
by depriving them of human agency or triggering panic about layoffs

This report explores how this uncertain landscape can be navigated.

This report seeks to understand the risks of AI to the financial system – both today and in the future – and to propose strategies for the mitigation of these risks

To unlock the potential of AI, financial institutions, regulators, and policy-makers should…

A

Responsibly deploy AI systems in the financial ecosystem of today

Key findings:

1. **AI raises challenging new issues because it is “foreign”**

   AI systems “think” in a way that is deeply foreign to humans and fundamentally different from the systems of the past. This creates new risks in the financial sector.

▼

2. **Managing this new foreignness requires the use of new solutions**

   Faced with AI’s foreignness, much of the old governance toolkit becomes ineffective; responsibly harnessing AI’s potential requires an openness to new modes of governance.

B

Responsibly scale the AI-ubiquitous financial ecosystem of tomorrow

Key findings:

3. **AI will drive policy shifts outside of the control of any one institution**

   The use of AI is spurring a cross-industry re-examination of competition policy, data rights and operational resilience, with profound implications for the financial system.

▼

4. **This new landscape will reshape the structure of the financial market**

   These fundamental re-examinations will shape the ways in which financial institutions can deploy AI and the broader set of strategic choices at their disposal.

C

Harness the potential of a financial ecosystem built on responsible AI

Key findings:

5. **AI presents an opportunity to raise the ethical bar**

   “Responsible AI” is not just about doing no harm; AI capabilities can also enable the financial sector to raise the ethical bar on how it serves clients and society at large.

▼

6. **At this higher bar, “trusted AI” can be a competitive differentiator**

   As the consumer’s digital norms enter a period of flux, financial institutions may be uniquely positioned to lead the deployment of “trust-first” AI models.

Key findings:
Responsibly deploy AI systems in the financial ecosystem of today

Responsibly deploy AI systems in the financial ecosystem of today

Responsibly scale the AI-ubiquitous financial ecosystem of tomorrow

Harness the potential of a financial ecosystem built on responsible AI

The ‘foreignness’ of AI systems is the source of the most serious risks and uncertainties surrounding the deployment of AI in the financial sector

The three facets of AI’s “foreignness”:

AI systems reason in “unhuman” ways

AI systems do not follow human constructs of logic; they can behave very differently from human actors given the same task.

AlphaGo

For example, AlphaGo defeated global experts at the game of Go, using strategies and techniques completely foreign to its human counterparts.

AI can evolve autonomously over time

The self-learning nature of AI systems allows them to change without direct input from human actors, potentially leading to unexpected outcomes.

Google

For example, Google’s Search improves by itself by identifying the pages where users end their search (i.e. they do not click through for additional links).

AI systems can be highly opaque

AI systems can involve a multitude of variables and many layers of intermediary processes, making them inscrutable even to their creators.

Mount Sinai

For example, researches at Mount Sinai trained an AI system to predict diseases, but the new tool didn’t provide explanations for its suggestions.

This “foreignness” is the source of the largest risks and uncertainties surrounding the use of AI in financial services:

⚠ New risks of bias:

As AI systems invent their own logic, disconnected from human notions of fairness, could they unintentionally limit a protected group’s access to financial products?

⚠ New sources of systemic risk:

As AI systems reason in new ways and provide limited visibility into their inner workings, could they create and propagate new forms of systemic risk?

⚠ The risk of unintentional collusion:

As AI systems interact with each other with greater frequency and velocity, could they learn to collude with each other and generate unfair outcomes for customers?

⚠ New risks to fiduciary duty:

As AI systems take on a broader set of customer-facing responsibilities, could they meet fiduciary responsibilities even with their foreign and opaque logic?

Many established approaches to governance and regulation of financial services may not be suited to ensuring the responsible deployment and ongoing operation of AI systems

Challenges of addressing AI's foreignness with current modes of operating:

AI systems reason in "unhuman" ways

can be a challenge for today's…

A

Human-centric accountability

Today's processes regulate the behaviour of humans and treat the systems they use as extensions of human conduct.

They do not fully account for systems that independently develop logic and conduct without explicit instruction from their human creators.

For example, if two AI systems are found to autonomously engage in price collusion, who should be held liable given that there are no human actors communicating with each other?

AI can evolve autonomously over time

can be a challenge for today's…

B

Slow-moving safeguards

Today's processes enforce safeguards on systems that seldom change during use, and infrequently re-examine a system's vulnerabilities.

They do not fully account for systems that can radically change their behaviour as they learn through real-world use.

For example, if an AI system is used to make lending decisions and it learns from its past mistakes, how can regulators and institutions be sure it does not develop a bias over time?

AI systems can be highly opaque

can be a challenge for today's…

C

Rigid auditability requirements

Today's processes have strict requirements around transparency and auditability, born from experience with static models.

They do not fully account for systems that behave in "foreign" ways, which may be inherently unexplainable but still valuable.

For example, if an AI system is found to be highly accurate (e.g. in assessing damage to a car through computer vision), should it be used even if it cannot explain its decisions?

Ensuring the responsible deployment of AI in the financial sector will require an openness to new modes of governance and new regulatory practices

Current modes of governance and regulation may need to shift to *new, fit-for-purpose approaches*:

Current modes of governance and regulation:

A **Human-centric accountability**

B **Slow-moving safeguards**

C **Rigid auditability requirements**

Potential new modes of governance and regulation:

Blame-free remediation

Designing mechanisms to remediate system accidents seamlessly will be essential in a world where the growing autonomy of AI systems makes them prone to unexpected behaviours:

- **Mechanisms for swift recourse** (e.g. speedy customer appeals) preserve trust in the face of AI failures and build institutional resilience
- **Protocols for blame-free investigations** of AI "failures" enable the organization to learn about the system as it improves through experience with failure

Real-time governance

Scanning for risk in forward-looking ways and designing dynamic safeguards will be essential in a world where the risks of AI systems keep evolving beyond their initial development:

- **War-gaming adverse scenarios** with interdisciplinary experts and fellow financial institutions prepares organizations to respond to unexpected points of failure
- **Dynamic safeguards** pegged to the movements of traditional, non-AI systems can help detect and prevent AI risks in real time

Fit-for-purpose "explainability"

Re-examining AI transparency requirements on a use-case-specific basis for the underlying algorithms and the resulting decisions will be crucial as the diversity of AI's users widens:

- **Frameworks to decide "if"** explainability is a requirement help teams prioritize the objectives of their AI (e.g. accuracy first)
- **Frameworks to decide "how"** explainability should be satisfied across diverse use cases facilitate more effective and meaningful explanations for all stakeholders involved

Key findings:
Responsibly scale the AI-ubiquitous financial ecosystem of tomorrow

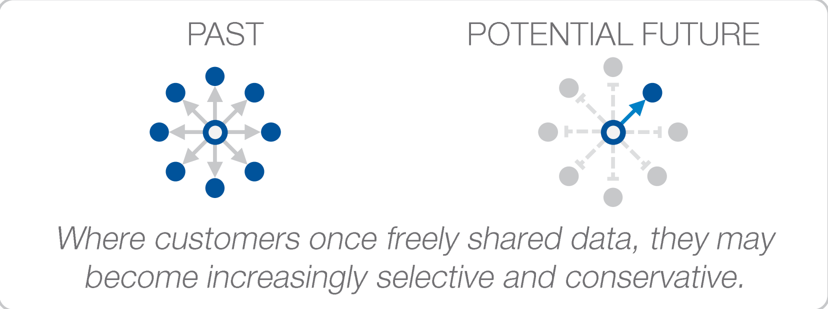
Responsibly deploy AI systems in the financial ecosystem of today

Responsibly scale the AI-ubiquitous financial ecosystem of tomorrow

Harness the potential of a financial ecosystem built on responsible AI

The global policy landscape is entering a period of flux as governments react to the disruptive impact of AI and digitization occurring across every sector

Customers are changing their data-sharing practices



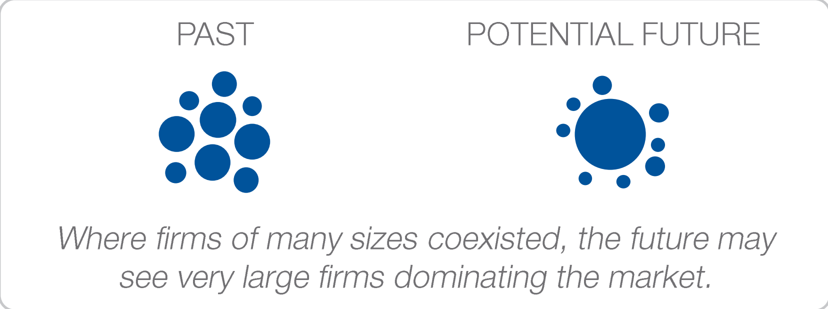
Where customers once freely shared data, they may become increasingly selective and conservative.

- Recent data privacy and misuse scandals have drawn attention to how institutions store, manage and use customers' personal data
- Individuals, institutions and regulators are increasingly aware of the value of data, and new privacy regulation is being interpreted/enforced
- This raises concern about how institutions transfer/monetize data, and whether existing regulation will be enough to protect customers

This is driving…

A fundamental re-examination of **the appropriate use of data**

AI systems are driving a winner-takes-all competitive landscape



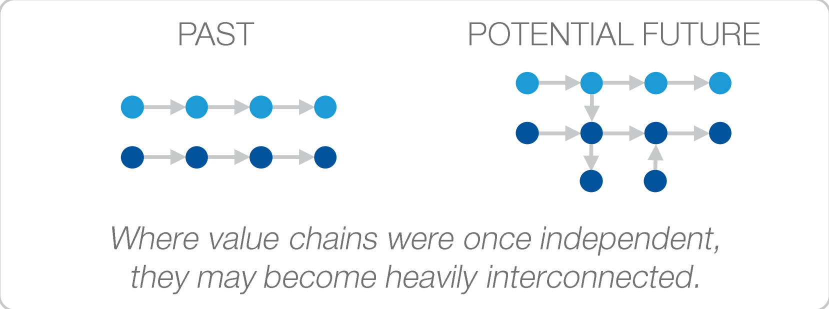
Where firms of many sizes coexisted, the future may see very large firms dominating the market.

- AI systems allow the establishment of a virtuous flywheel whereby data leads to better offerings, attracting customers, and thus more data
- This is enabling the creation of very large institutions that have extremely high barriers to entry posed by the scale of their data
- This is raising concerns that current competition policy may not be sufficient or appropriate in regulating these large market players

This is driving…

A fundamental re-examination of **competition and antitrust policy**

Value chains are becoming increasingly interconnected



Where value chains were once independent, they may become heavily interconnected.

- Institutions are increasingly using data and capabilities from third parties, and are becoming capability providers to other market participants
- These dynamics are expanding the breadth of participants in a given market and the degree of interconnectedness between them
- This raises concerns on how this more complex financial system can be effectively governed, and how new risks might be detected and prevented

This is driving…

A fundamental re-examination of **risk operations and accountability**

The resolution of these economy-wide policy uncertainties will have far-reaching implications for the operational and strategic avenues open to financial institutions

A fundamental re-examination of **the appropriate use of data**

Will raise serious questions regarding…

…the use of data:

A How will my ability to use customer data change? What new types of data will be allowed or disallowed in the provision of financial offerings?

E.g. for lenders, this affects which factors they can use as inputs in their credit models

…the sharing of data:

B How will my ability to access third-party data change? What internal data will I need to make available to other parties, and how will those parties be regulated?

E.g. for advisers, this affects the barriers to entry for other players to provide advice

A fundamental re-examination of **competition and antitrust policy**

Will raise serious questions regarding…

…the sector's competitive dynamics:

C Which market participants will I compete with (e.g. fintech, bigtech)? What policy advantage might different types of players have?

E.g. for retail banks, this affects the ability of new challenger banks to enter the market

…allowable business models:

D What types of business models will I be able to pursue? Will I be able to pursue several different business models simultaneously?

E.g. for aggregators, this affects their ability to offer their own products on their platform

A fundamental re-examination of **risk operations and accountability**

Will raise serious questions regarding…

…the regulation of new players:

E How will new systemically important players (that traditionally sat outside the financial sector) be governed? How will this affect my ability to partner/compete?

E.g. for insurers, this affects their ability to benefit from new technologies such as cloud

…firms' risk operations:

F How will new systemic risks resulting from a more complex financial system be managed? How will this affect my compliance processes?

E.g. for asset managers, this affects the total compliance burden faced by the firm

Ensuring that policy decisions take account of these operational, strategic and human capital implications **requires senior leaders to engage proactively** in policy discussions and industry consultations.

Key findings:
Harness the potential of a financial ecosystem built on responsible AI

Responsibly deploy AI systems in the financial ecosystem of today

Responsibly scale the AI-ubiquitous financial ecosystem of tomorrow

Harness the potential of a financial ecosystem built on responsible AI

‘Responsible AI’ presents the opportunity to do more than avoiding harm – it has the potential to raise the ethical bar for the financial system as a whole

The social licence of the financial sector has, and continues to be, contingent on meeting an ethical bar for:

Accessibility and fairness
In allocating financial products across the population

Transparency
In how it uses quantitative models to guide decisions

Consumer protection
In delivering products and advice aligned to customers’ best interest

Market stability
In safeguarding the long-term resilience of markets

The use of AI does not alter the importance of this ethical bar, but does create an opportunity to exceed it:

***Widening* accessibility**

NOVA

Fintechs such as Nova Credit are widening financial access to unbanked and underbanked populations (e.g. new immigrants) by inferring creditworthiness from digital footprints and psychometric data.[2,3]

***Deepening* transparency**

TEMENOS | logical glue

Logical Glue’s Explainable AI platform provides everyday users (e.g. customers, front-line business users) of automated financial decision systems with meaningful transparency that was previously available only to the developers.[4]

***Improving* client outcomes**

personetics

Personetics is able to understand a customer’s circumstances in granular ways: e.g. by continually analysing spending behaviour to locate unused funds with which to pay off student debt.[5]

***Bolstering* market efficiency**

The Bank of Italy is able to identify and track depositors’ trust in banks at any point in time through sentiments reflected in Twitter posts. Doing so allows a real-time view of threats to market stability.[6]

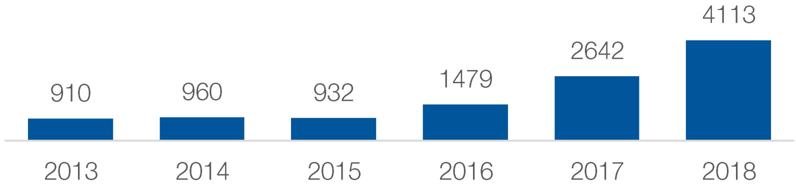
Most financial institutions that choose to “exceed the ethical bar” begin by identifying a gap in customers’ trust in incumbents and **move to capitalize on the “trust premium” by closing that gap**.

Growing customer concerns about established data practices could enable financial institutions to build a 'trust advantage', conferred in part by their highly regulated nature

A "tech-lash" is materially shifting customer expectations…

Consumer attention to their data rights and digital sovereignty is at an all-time high, especially as large technology players become mired in scandals over data misuse, and as regulatory interventions in this space intensify.[7,8]

Number of consumer complaints about the misuse of their data across Europe[9]



…putting bigtech on the defence…

Many bigtech businesses employ freemium revenue models that have scaled in laissez-faire regulatory environments. These models could be challenged over the long term if consumers and regulators demand increased digital sovereignty.

Growing public and regulatory scrutiny into bigtech business practices globally, by the numbers[10,11,12]

80%
US Consumers changed spending habits due to a brand's security practices

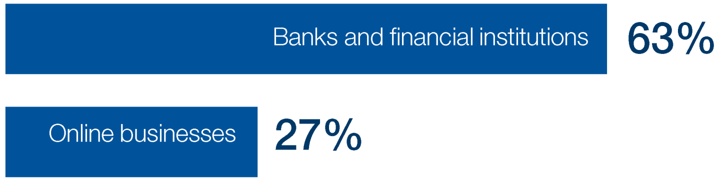
15
Open GDPR investigations into bigtechs

7
Open antitrust investigations into bigtechs

…and giving financial institutions a chance to seize the "trust advantage".

Financial institutions enjoy higher rates of customer trust than online businesses and are more experienced at navigating shifting regulatory environments. This could position them well to build on trust as the foundation of a competitive advantage.

Percentage of consumers that trust these providers to protect their personal information across Europe[13]



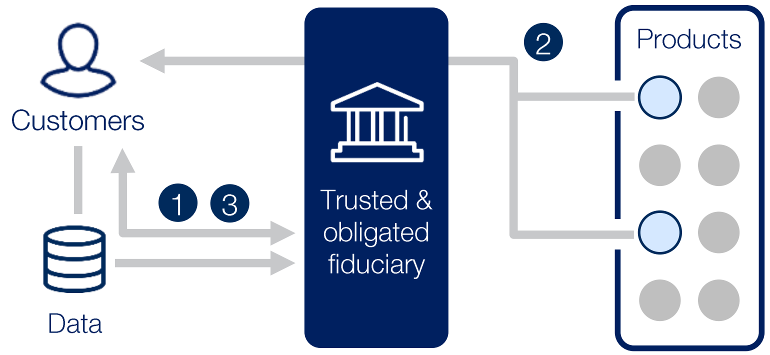
By asserting a new model for putting customers' data to work in a "trust first" framework, financial institutions could position themselves to assert a new customer value proposition.

Financial institutions have an opportunity to lead with ‘trust-first’ digital models that set the gold standard for fiduciary care, data stewardship and consumer sovereignty

Institutions can turn their regulatory burden into an advantage by leading with trust, for example, as…

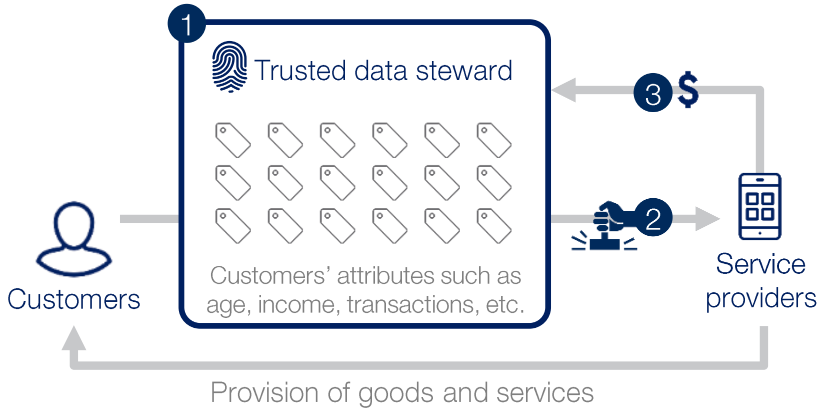
A trusted financial adviser

1. Offering **automated financial expertise** to mass consumers in more economical and accessible ways than is possible with traditional labour-intensive services
2. Aligning offerings to a **customer’s best interest,** based on a holistic perspective of their financial situation
3. Providing customers with **compelling rationale and next-best actions** with explainable AI



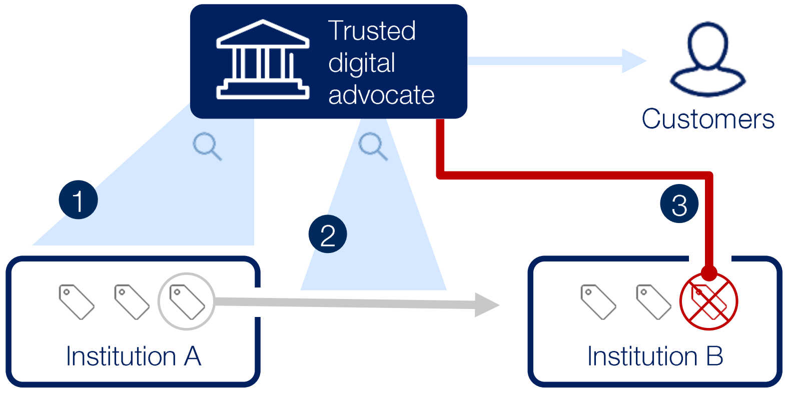
A trusted data steward

1. **Providing secure rails** for the movement of trusted data between consumers and digital service providers
2. Making **privacy-preserving attestations** of consumer identity and data for third-party service providers
3. **Creating new revenue sources** from service providers willing to pay for a more trusted and streamlined means of confirming client identity


1
Trusted data steward
Customers’ attributes such as age, income, transactions, etc.
3 $
2
Customers
Service providers
Provision of goods and services

An advocate for customers’ digital rights

1. **Scanning customers’ digital interactions** to identify when providers are using data without their permission
2. **Actively notifying** customers each time their data is collected, harvested for a new use or transferred to a new provider
3. Enabling customers with the means to **re-curate misrepresented identities** or enforce a right to be forgotten



Key findings: Implications

Implications

Strategic decision-makers at financial institutions:

AI will drive foundational shifts in firms' strategies, requiring executive attention

- The rapid adoption of AI both within, and beyond, the financial sector is triggering policy dialogues that will meaningfully shape how these technologies can be deployed
- Engaging proactively with all stakeholders will be critical to institutions seeking to play an active role in shaping their future strategic options, rather than being a "policy-taker"
- Executive involvement is critical to ensuring that regulation is defined with consideration for the operational, strategic and human capital implications of policy changes

Effective AI governance demands both offense and defence

- On one hand, institutions must invest in defensive measures to prevent the potential governance risks of AI systems from occurring
- On the other hand, excellence in AI governance and associated activities, such as contextual explanations of AI decisions, represents an opportunity to differentiate based on close alignment with customers' interests
- Balancing strategic opportunities and defensive plays will define winning institutions in an AI-enabled financial system of the future

…and regulators and policy-makers:

The responsible use of AI necessitates openness to new forms of governance

- New regulatory and governance frameworks will need to be developed and adopted in order to help ensure effective AI governance
- Regulators and financial institutions each hold a piece of this puzzle: regulators need data on the impact of AI from institutions to make informed choices about governance requirements, while institutions need additional policy clarity from regulators in order to successfully deploy AI
- Collaborations such as sandboxes and the co-building of regtech solutions are vital in facilitating the exchange of information required to support data-driven policy-making

References

1. https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20frontier%20Modeling%20the%20impact%20of%20AI%20on%20the%20world%20economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-economy-September-2018.ashx Note: The analysis referenced stems from a cross-industry study of AI's potential impact on future cash flows, which includes financial institutions
2. https://www.pymnts.com/news/banking/2018/alternative-credit-scoring-underbanked-unbanked-nova/
3. https://www.centerforfinancialinclusion.org/to-bank-the-unbanked-start-using-alternative-data
4. https://aithority.com/natural-language/information-extraction/temenos-acquires-a-saas-based-patented-explainable-ai-xai-platform/
5. https://www.bankingtech.com/2017/10/personetics-leverages-ai-to-help-chip-away-at-student-loan-crisis/
6. https://www.fsb.org/wp-content/uploads/P011117.pdf
7. https://www.politico.eu/article/tech-predictions-2019-facebook-techclash-europe-united-states-data-misinformation-fake-news/
8. https://www.pewinternet.org/2018/06/28/public-attitudes-toward-technology-companies/
9. https://techcrunch.com/2019/02/28/privacy-complaints-received-by-tech-giants-favorite-eu-watchdog-up-more-than-2x-since-gdpr/
10. https://www.mytotalretail.com/article/when-it-comes-to-security-us-consumers-put-their-money-where-their-trust-is/
11. https://blog.globalwebindex.com/chart-of-the-week/trust-data-privacy/
12. https://www.cnn.com/2019/07/17/tech/amazon-antitrust-european-commission/index.html
13. https://www.bafin.de/SharedDocs/Downloads/EN/dl_bdai_studie_en.html;jsessionid=DF0654DBF78C8FA8AE5E4AA58148CB44.1_cid381?nn=9866146. Survey question: "Different authorities (government departments, local authorities, agencies) and private companies collect and store personal information about you. To what extent do you trust the following authorities and private companies to protect your personal information?"

AI explainability

In this chapter, we will explore:

- Risks (Where are the real risks? Which fears are overstated?)
- Best practices (What are the best practices in governing AI systems?)
- Opportunities (Where might the risks of AI be turned into opportunities?)

Chapter summary

Some forms of AI are not interpretable, creating concerns for financial executives and supervisors who struggle to trust solutions that employees cannot understand or explain. In some cases, these fears have stalled efforts to deploy new technology.

While these concerns are understandable, pursuing an instinctive desire for "fully transparent" AI or no AI at all is unlikely to achieve the desired ends. Instead, the reason a stakeholder requires an explanation for a given use case should sit at the centre of governance practices.

Context, control and auditability are all reasonable things to want but may not require full interpretability. An explanation that is meaningful for a technically savvy user may be unintelligible to the everyday consumer. Businesses and supervisors will need to reconsider the urge for a one-size-fits-all solution to the explainability problem, given its context-specific nature.

The technical ability to deploy facets of explainability – particularly around decision context – may become an important source of competitive differentiation. Interactive forms of explainability that allow users to clarify and probe the rationale behind AI-based financial management advice, for example, will differentiate such products from less transparent competitors.

The ‘explainability’ of AI models is a growing concern for both financial institutions and regulatory authorities

Financial institutions

BLACKROCK

> “The industry needs to address the problem of interpretability […] before putting investor money into play.”
>
> – Head of Liquidity Research[1]

Bank of America

> “There’s no time frame for when an AI system could be deployed […] because solving the explainability problem is so important.”
>
> – Managing Director[2]

Capital One

> “We only roll out machine learning where we feel comfortable there are no biases or lack of transparency…”
>
> – Managing Vice-President, ML & AI[3]

Regulatory authorities

BaFin

> “Ultimately, [explainability] is a prerequisite to secure the very principle of responsibility.”[4]

MAS

> “[Transparency] is one of four key principles for the use of AI and data analytics […] in the provision of financial products and services.”[5]

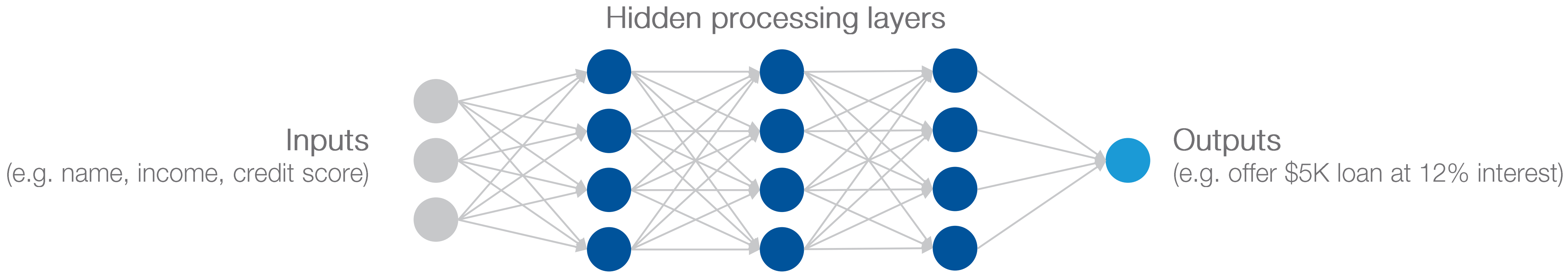
FSB FINANCIAL STABILITY BOARD

> “Efforts to improve the interpretability of AI may be important conditions not only for risk management, but also for greater trust from the general public.”[6]

The enormous complexity of some AI systems makes it difficult to obtain an interpretable 'explanation' for why the system has produced a given output

Today's most advanced AI systems are complex and multilayered*:

Cutting-edge approaches to developing artificial intelligence, such as deep learning, are highly complex, often containing dozens of "hidden" layers between the data inputs and the resulting model outputs.


Hidden processing layers
Inputs
(e.g. name, income, credit score)
Outputs
(e.g. offer $5K loan at 12% interest)

These systems can be opaque "black boxes", even to their developers, for several reasons:

Volume of factors considered

AI systems can incorporate high volumes of input factors, making it difficult to understand which inputs most heavily influenced the outputs.

Multitude of intermediary steps

AI systems can run through hundreds of intermediate steps to arrive at a decision, making it difficult to step through and follow the process.

Self-changing over time

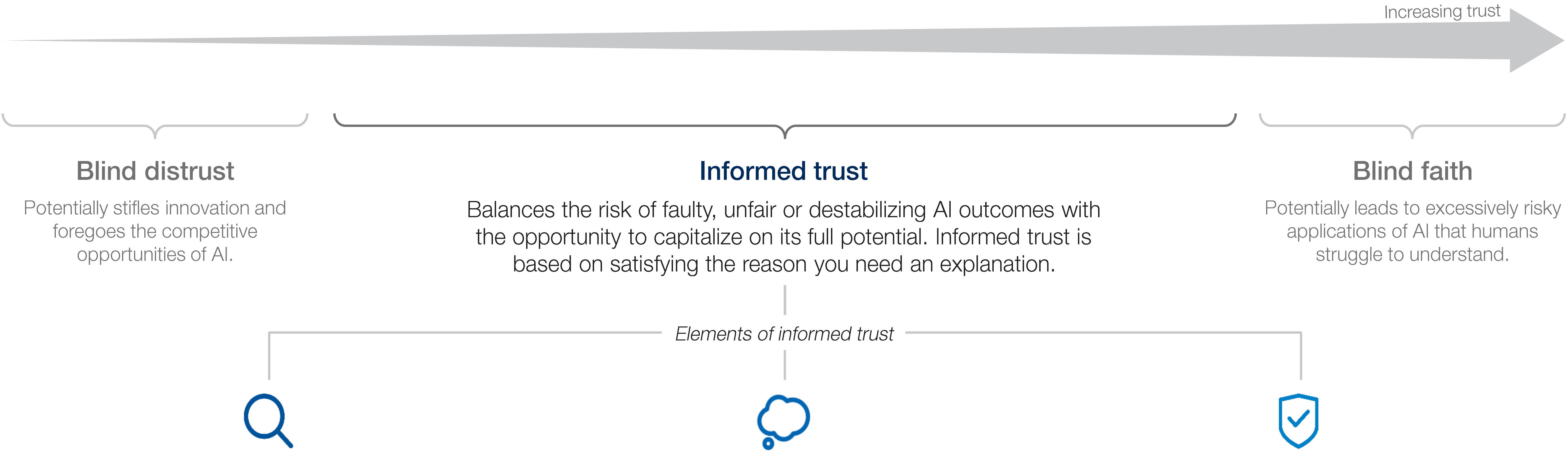
AI systems change autonomously, making it difficult to anticipate future behaviour based on the past decisions of the system.

Disconnected from human logic

AI systems process data in ways that do not always align with human 'judgement' or established constructs of fairness.

* Note: This diagram illustratively approximates "deep-learning" systems, one of many AI techniques; many other techniques exhibit characteristics that introduce various degrees of "explainability" challenges

The desire for AI explainability is grounded in a need for informed trust, which involves navigating the middle ground between blind distrust and blind faith in expert systems



Increasing trust

Blind distrust

Potentially stifles innovation and foregoes the competitive opportunities of AI.

Informed trust

Balances the risk of faulty, unfair or destabilizing AI outcomes with the opportunity to capitalize on its full potential. Informed trust is based on satisfying the reason you need an explanation.

Blind faith

Potentially leads to excessively risky applications of AI that humans struggle to understand.

Elements of informed trust

Transparency

Humans look to understand the logic driving a system and want to be able to interrogate it. They look to follow *how* a system arrived at its decision.

AI systems make this difficult to achieve, since…

Some AI systems consider a massive volume of factors through a multitude of intermediary steps, making it difficult to follow the system’s thinking.

Context

Humans need to understand *why* a system made a specific choice. They look for a few meaningful, causal relationships driving an outcome.

AI systems make this difficult to achieve, since…

AI systems can be disconnected from human logic; and abstract away their complex multistep analyses of inputs, which makes it difficult to grasp the clear, causal relationships that drive decisions.

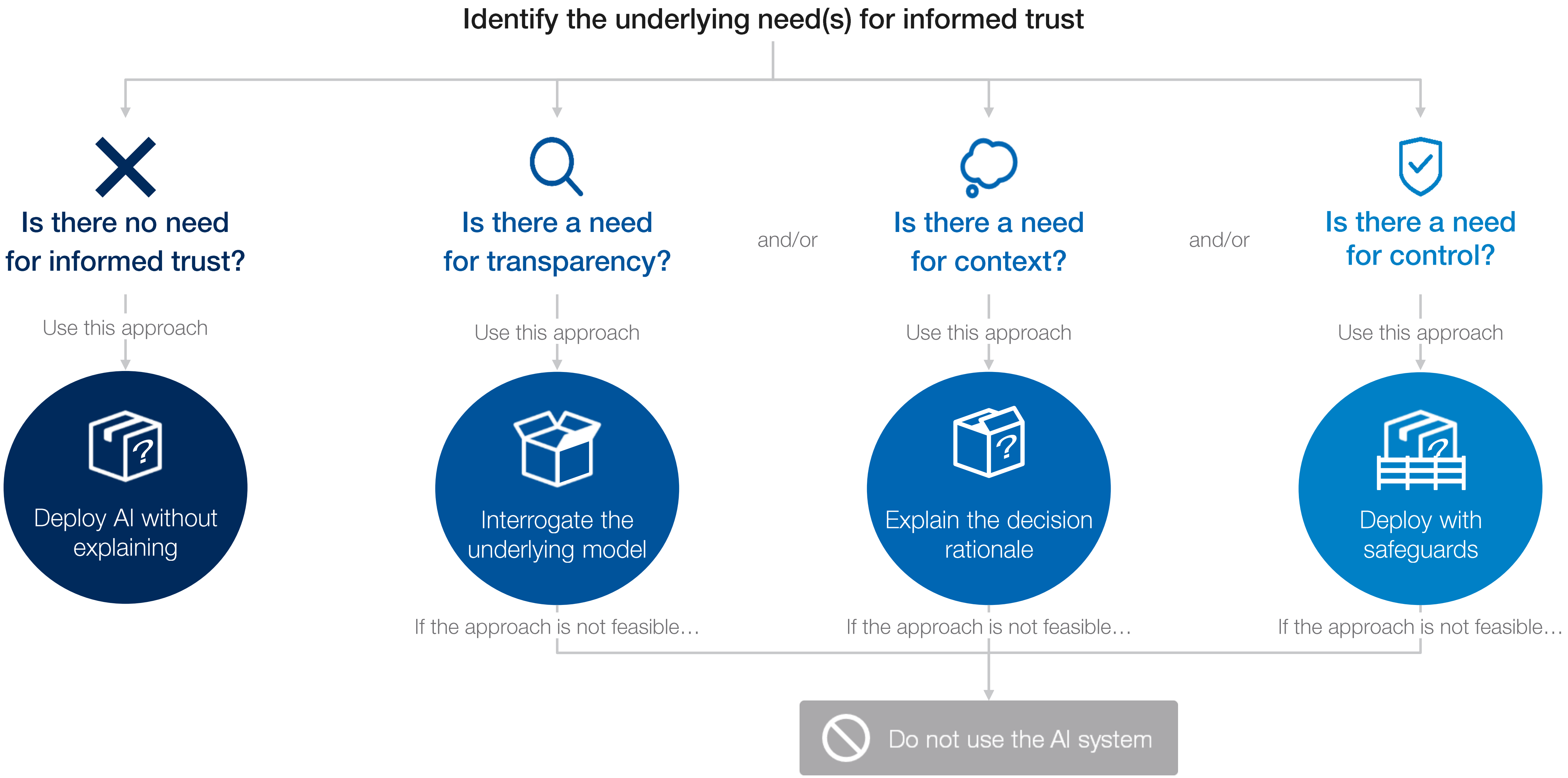
Control

Humans may need confidence that AI *will not violate* norms, laws or business requirements, especially as the outputs of AI systems can be unintuitive and unexpected.

AI systems make this difficult to achieve, since…

AI systems can evolve their decision-making process over time, producing results that do not match the expectations set by their past performance.

Not all situations call for 'full interpretability' to achieve informed trust; a use case-based framework is required to consider the context in which an explanation is needed



Identify the underlying need(s) for informed trust

Is there no need for informed trust?

Use this approach

Deploy AI without explaining

Is there a need for transparency?

Use this approach

Interrogate the underlying model

If the approach is not feasible…

and/or

Is there a need for context?

Use this approach

Explain the decision rationale

If the approach is not feasible…

and/or

Is there a need for control?

Use this approach

Deploy with safeguards

If the approach is not feasible…

Do not use the AI system

Over the following slides, we explore each of these approaches to managing explainability in greater detail

Deploy AI
without explaining

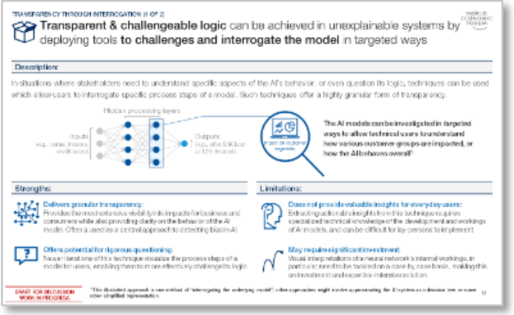
Interrogate the
underlying model

Interrogate the
underlying model

Interrogate the
underlying model

Do not use
the AI

For each of the first four approaches, we explore…



TRANSPARENCY THROUGH INTERROGATION (1 OF 2)

Transparent & challengeable logic can be achieved in unexplainable systems by deploying tools **to challenges and interrogate the model** in targeted ways

Description:

In situations where stakeholders need to understand specific aspects of the AI's behavior, or even question its logic, techniques can be used which allow users to interrogate specific process steps of a model. Such techniques offer a highly granular form of transparency.

The AI models can be investigated in targeted ways to allow technical users to understand how various customer groups are impacted, or how the AI behaves overall*

Strengths:

Delivers granular transparency: Provides the most extensive visibility into impacts for business and consumers while also providing clarity on the behavior of the AI model. Often used as a central approach to detecting bias in AI

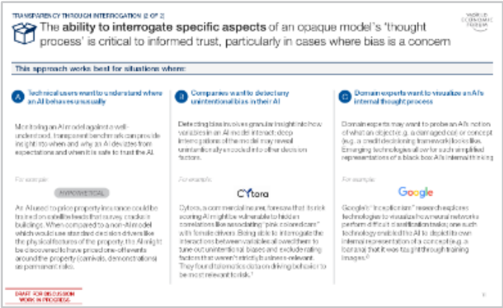
Offers potential for rigorous questioning: Newer iterations of this technique visualize the process steps of a model for users, enabling them to more effectively challenge its logic

Limitations:

Does not provide valuable insights for everyday users: Extracting actionable insights from this technique requires specialized technical knowledge of the development and workings of AI models, and can be difficult for lay-persons to implement

May require significant investment: Visual interpretations of a neural network's internal workings, in particular, need to be tackled on a case by case basis, making this an investment and expertise-intensive solution

DRAFT FOR DISCUSSION
WORK IN PROGRESS



TRANSPARENCY THROUGH INTERROGATION (2 OF 2)

The **ability to interrogate specific aspects** of an opaque model's 'thought process' is critical to informed trust, particularly in cases where bias is a concern

This approach works best for situations where:

A Technical users want to understand where an AI behaves unusually

Monitoring an AI model against a well-understood, transparent benchmark can provide insight into when and why an AI deviates from expectations and when it is safe to trust the AI.

For example: HYPOTHETICAL

An AI used to price property insurance could be trained on satellite feeds that survey cracks in buildings. When compared to a non-AI model which would use standard decision drivers like the physical features of the property, the AI might be discovered to have priced one-off events around the property (carnivals, demonstrations) as permanent risks.

B Companies want to detect any unintentional bias in their AI

Detecting bias involves granular insight into how variables in an AI model interact; deep interrogations of the model may reveal unintentionally encoded into other decision factors.

For example: Cytora

Cytora, a commercial insurer, foresaw that its risk scoring AI might be vulnerable to hidden correlations like associating "pink colored cars" with female drivers. Being able to interrogate the interactions between variables allowed them to tune out unintentional biases and exclude rating factors that weren't strictly business-relevant. They found telematics data on driving behavior to be most relevant to risk.

C Domain experts want to visualize an AI's internal thought process

Domain experts may want to probe an AI's notion of what an object (e.g. a damaged car) or concept (e.g. a credit decisioning framework) looks like. Emerging technologies allow for such simplified representations of a black box AI's internal thinking

For example: Google

Google's "Inceptionism" research explores technologies to visualize how neural networks perform difficult classification tasks; one such technology enabled the AI to depict its own internal representation of a concept (e.g. a banana) that it was taught through training images.

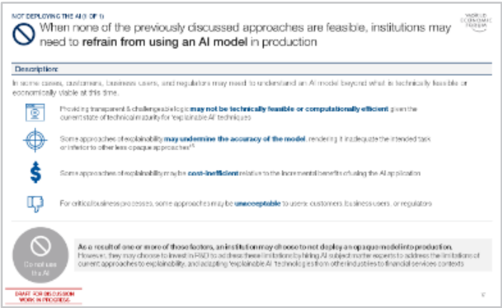
DRAFT FOR DISCUSSION
WORK IN PROGRESS

1 **Overview**
Brief description of the form of bias

2 **Assessment**
Consider its strengths and limitations

3 **Examples**
Explore real-world and hypothetical cases

For the last approach – "do not use" – we explore…



NOT DEPLOYING THE AI (1 OF 1)

When none of the previously discussed approaches are feasible, institutions may need to **refrain from using an AI model** in production

Description:

In some cases, customers, business users, and regulators may need to understand an AI model beyond what is technically feasible or economically viable at this time.

Providing transparent & challengeable logic **may not be technically feasible or computationally efficient** given the current state of technical maturity for 'explainable AI' techniques

Some approaches of explainability **may undermine the accuracy of the model**, rendering it inadequate the intended task or inferior to other less opaque approaches

Some approaches of explainability may be **cost-inefficient** relative to the incremental benefits of using the AI application

For critical business processes, some approaches may be **unacceptable** to users: customers, business users, or regulators

Do not use the AI

As a result of one or more of these factors, an institution may choose to not deploy an opaque model into production. However, they may choose to invest in R&D to address these limitations by hiring AI subject matter experts to address the limitations of current approaches to explainability, and adapting 'explainable AI' technologies from other industries to financial services contexts

DRAFT FOR DISCUSSION
WORK IN PROGRESS

1 **Overview**
Consider when the approach might be necessary

2 **Potential future**
Discuss how institutions may overcome these limitations in the future

AI explainability:
Approaches to managing explainability

- Deploy AI without explaining
- Interrogate the underlying model
- Explain the decision rationale
- Deploy with safeguards
- Do not use the AI

In a few cases, it may be sufficient for an AI system to be effective, meaning no explanation is required to ensure the system's governance and effectiveness

An explanation may not be needed when…

The potential adverse impact of an AI's decisions is negligible

For example:

KANETIX.CA

Online insurance aggregator Kanetix.ca uses AI to optimize a local step in the process from customer interest to purchased insurance, i.e. the number of likely purchasers transferred from its website to an insurer. Compared to the final quote, this AI has little potential for harm, so no effort was made to explain it.[7]

Other objectives such as accuracy take precedence over transparency

For example:

ROSS

Ross Intelligence, an AI-powered legal research tool, uses AI to find documents relevant to a case. Rather than sacrificing model quality to explain why each document was retrieved, it focused on maximizing the number of *relevant* instances retrieved for each search query.

There may be no explicit regulatory requirement for an explanation

For example:

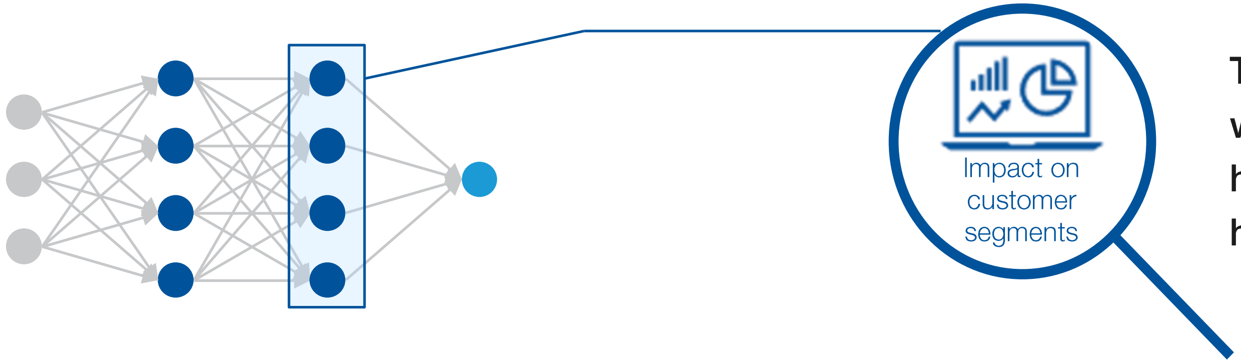
Under GDPR, transparency around targeted marketing does not fall under the same responsibility requirements as applications that "significantly impact data subjects". Thus, European banks using AI to personalize offers may not have to explain the rationale to customers[7]

Such cases are more often the exception than the norm, because without explanations users cannot establish informed trust in their AI.

Where transparency into the AI's inner workings is a priority, technical interrogation techniques offer ways to challenge and probe the model in targeted ways

Description:

In situations where stakeholders need to understand specific aspects of the AI's behaviour, or even question its logic, techniques can be used that allow users to interrogate specific process steps of a model. Such techniques offer a highly granular form of transparency.



The AI models can be investigated in *targeted* ways to allow technical users to understand how various customer groups are affected, or how the AI behaves overall.*

Strengths:

Delivers granular transparency: Provides the most extensive visibility into the impacts for business and consumers while also providing clarity on the behaviour of the AI model; often used as a key method in detecting bias in AI.

Offers potential for rigorous questioning: Newer iterations of this technique visualize the process steps of a model for users, enabling them to more effectively challenge its logic.

Limitations:

Does not provide valuable insights for everyday users: Extracting useful insights from this technique requires specialized technical knowledge of the development and workings of AI models, and can be difficult for laypersons to implement.

May require significant investment: Visual interpretations of a neural network's internal workings, in particular, need to be tackled on a case-by-case basis, making this an investment and expertise-intensive solution.

* This illustrated approach is one method of 'interrogating the underlying model'; other approaches might involve approximating the AI system as a decision tree or some other simplified representation

The ability to interrogate specific aspects of an opaque model's 'thought process' is critical to informed trust, particularly if bias is a concern

This approach works best for situations where:

A Technical users want to understand where an AI behaves unusually

Monitoring an AI model against a well-understood, transparent benchmark can provide insight into when and why an AI deviates from expectations and when it is safe to trust the AI.

For example:

HYPOTHETICAL

An AI used to price property insurance could be trained on satellite feeds that survey cracks in buildings. When compared to a non-AI model, which would use standard decision drivers such as the physical features of the property, the AI might be discovered to have priced one-off events around the property (carnivals, demonstrations) as permanent risks.

B Companies want to detect any unintentional bias in their AI

Detecting bias involves granular insight into how variables in an AI model interact; deep interrogations of the model may reveal bias unintentionally encoded into other decision factors.

For example:

Cytora

Cytora, a commercial insurer, foresaw that its risk-scoring AI might be vulnerable to hidden correlations such as associating "pink-coloured cars" with female drivers. Being able to interrogate the interactions between variables allowed them to tune out unintentional biases and exclude rating factors that weren't strictly business-relevant. They found telematics data on driving behaviour to be most relevant to risk.[8]

C Domain experts want to visualize an AI's internal thought process

Domain experts may want to probe an AI's notion of what an object (e.g. a damaged car) or concept (e.g. a credit decision-making framework) looks like. Emerging technologies allow for such simplified representations of a black-box AI's internal thinking.

For example:

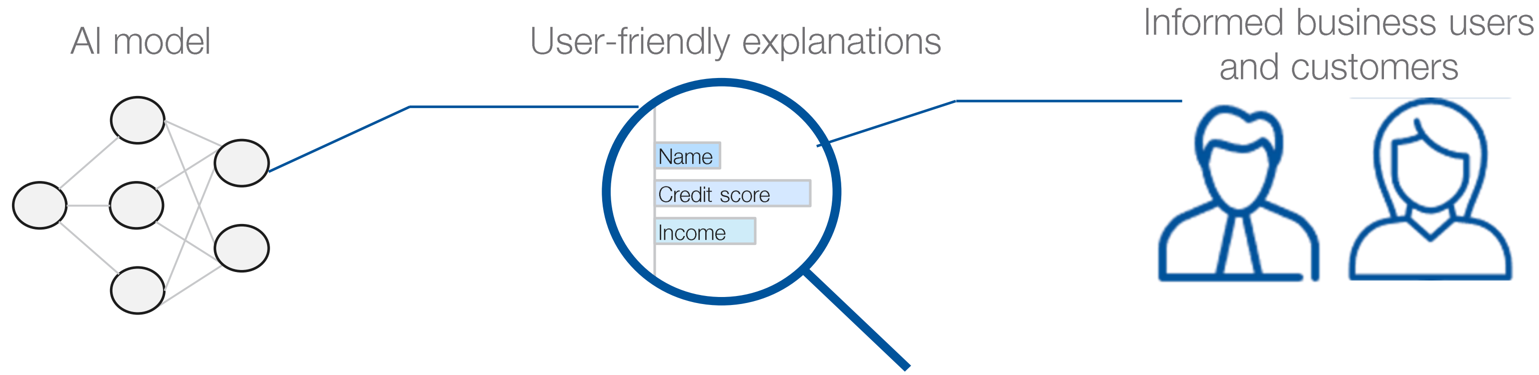
Google

Google's "Inceptionism" research explores technologies to visualize how neural networks perform difficult classification tasks; one such technology enabled the AI to depict its own internal representation of a concept (e.g. a banana) that it was taught through training images.[9]

Where users require context to trust and act on the decisions of a model, providing an intuitive rationale may be more important than code transparency

Description:

This approach provides end users, such as clients, business users or even regulators, with non-technical and actionable insight into the "reasoning" behind a model's decision. It does not provide a granular, end-to-end understanding of the AI's inner workings.


AI model
User-friendly explanations
Name
Credit score
Income
Informed business users and customers

Emerging techniques for providing context allow users to grasp the most important factors behind an AI decision

Strengths:

Delivers a rationale tailored to a user's background and goals: This approach provides users with the knowledge they need to "sanity check" a model output, and/or the knowledge to shape their future choices in response to the output.

Supports customer interactions: In addition to providing one-time explanations for financial decisions, interactively explaining the automated advice and recommendations can encourage deeper engagement with the customer or employee on the receiving end.

Limitations:

Offers limited visibility into *how* AI explanations are derived: For some use cases, simple explanations may not provide the granular details technical users need to improve a model or spot its limitations.

Requires significant expertise and investment: This approach is the focus of nascent R&D efforts to build user-friendly explainability techniques or interfaces, requiring added investment to integrate into products or services.

Explaining the basis of a decision is particularly useful in consumer-facing applications, which call for simple, actionable and intuitive rationales

This approach works best for situations where:

A Consumers need to judge if an outcome is fair

For decisions that affect customers' access to financial products, explaining the driving factors is critical to convincing them they are fair. A person in a minority group could visualize how a customer similar to her in all other regards would have fared in a credit decision, for example.

For example:

BBVA

BBVA has been experimenting with "counterfactual" explanations of AI decisions across its portfolio of companies. These work by designing a "digital quasi-twin" of a customer that is as close as possible to her profile, but would get a different decision. This helps business users identify potentially biasing variables, and may be a tool for customers in the future.[10]

B Consumers need the knowledge to shape future actions

For processes aimed at creating empathetic customer experiences, enriching an AI's decision rationale with actionable next steps for the users can improve outcomes and promote a deeper sense of human agency and engagement.

For example:

HYPOTHETICAL

An online mortgage decision-making AI could explain to a customer, "You didn't qualify because you did not pay your rent for the past two months. If you make the next four payments in a row, your score may improve sufficiently to obtain an approval." Providing an actionable rationale requires drawing on more granular insights within the AI.[11]

C Consumers need to understand clear, causal relationships

Simplifying an AI's design to produce only meaningful causal relationships can help businesses avoid the specific non-intuitive, non-palatable relationships inherent in some AI, making them easier to understand.

For example:

Scotiabank® EQUIFAX

Scotiabank and Equifax developed techniques to constrain their credit-scoring AI and ensure a straightforward causal relationship between credit-scoring factors and the effective rate.[12]
This allowed them to continue generating simple reason codes mandated by regulators while benefiting from machine-learning accuracy. Scotiabank even enjoyed increased customer acquisition while keeping within the bank's risk appetite.[13]

The ability to explain the rationale of an opaque system's decision-making process is also critical to informing employee actions based on AI predictions

This approach works best for situations where:

D A rationale is needed to shed light on novel insights

Front-line employees (e.g. sales staff) supported by AI tools may seek intelligible ways to understand the novel, nuanced insights that come from an AI's unique pattern of thinking, without having to probe its inner workings.

For example:

Coefficiency Lab

Coefficiency Lab designed a client-insights AI to recommend next best actions to sales and trading staff. As the AI analyses client activity to produce sales recommendations, a "knowledge-graphing" tool records the driving insights, i.e. the evolving relationships between clients, trades and market events, translating them into natural language explanations to guide sales staff.

E A rationale is needed to conform to the broader paradigm of a domain expert

Domain experts may trust an unsupervised AI only if its explanations match their seasoned knowledge of the business. Presenting an AI system's thinking as a familiar framework or set of rules helps experts to reconcile it with their own "mental models" and accept the outcome.

For example:

xAI Asset Management and Research

As part of its investment decisioning process, XAI Asset Management must explain its AI based global macro-forecasting engine's predictions. The regime based approach taken in the system, in combination with an explanation layer, presents the AI's prediction drivers in terms of familiar economic relationships, and provides evidence for similar such regimes in the past.[14]

F Informed trust is necessary but granular details cannot be shared

A business may need to explain its AI without revealing trade secrets/sensitive customer information. Approximating the AI's rationale can address this need without exposing the AI's inner workings.

For example:

stripe

Stripe, an online payments processor, needs to assure customers that its fraud-detection AI doesn't block charges unnecessarily. Its explainability tool justifies blocked charges without exposing the details of the AI's decision-making to fraudsters who could potentially game its rules.[15]

Where institutions desire to contain and control the actions of an AI model, 'guard rails' can be established to limit specific negative outcomes

Description:

The development of safeguards, customized to a particular use case, can place boundaries on the possible outcome of the model or subject the model to oversight. These safeguards do not provide any insight into how outcomes were derived. There are three primary forms:

Human oversight

A subject matter expert reviews the suggestions of an AI model before they are executed

Upper/lower limits

Boundaries are placed on the allowable outcomes of an AI model

Recourse mechanisms

Users of an AI model can access an alternative process if they aren't satisfied with its outcome

Strengths:

Enables reasonable outcomes: Safeguards reassure businesses and their supervisors that an AI's outputs fall within reasonable bounds, and that its behaviour is predictable to some degree.

Sheds light on "edge cases": Designing safeguards gives businesses insight into the edge cases where an AI may "misbehave", and prevents those risky edge cases from materializing.

Limitations:

Doesn't offer transparency: Safeguards don't allow businesses or regulators to interpret the AI's logic, understand its decision-making or otherwise derive insights from the AI model.

Complex design: Designing safeguards can involve a large amount of subjectivity. Especially in the early stages of commercial use, it is difficult to know if safeguard design captures all possible cases of failure.

Safeguards are most suitable when a real understanding of the AI model is not critical to establishing trust, and possible negative outcomes are well understood

This approach works best for situations where:

A There are already regulatory/business constraints on allowable outcomes

Regulatory constraints (e.g. a maximum on interest rates) or commonly accepted business logic (e.g. a required rate of return) can be used to set the boundaries within which an AI is allowed to optimize.

For example:

Large European Insurer

A leading European insurer uses AI to optimize capital subject to Solvency II. Safeguards are set by another, fully interpretable, model that has been used historically and approved by regulators. This "nested" approach provides strict oversight and clear compliance while allowing the firm to benefit from the improved performance and efficiency of a deep-learning model.

B Agreed-upon fairness thresholds can safeguard against AI bias

Monitoring an AI against agreed-upon quantitative thresholds of "fairness", and reworking the AI if these thresholds are breached, can help ensure its outcomes do not adversely affect protected groups in commercial use.

For example:

Upstart cfpb

Upstart, which uses AI and educational data to extend credit to new graduates, has created metrics to evaluate the fairness of the credit decisions jointly with the Consumer Financial Protection Bureau. Thresholds based on such metrics, which measure "tangible harm to consumers" (i.e. complaint patterns or other statistics on disparate impact) will likely serve as ongoing AI safeguards in the future.[16]

C Human experts can swiftly redress faulty AI decisions

When institutions have the subject matter expertise to provide oversight, humans can ensure that an AI model is making the right suggestions (before or after those suggestions are acted on).

For example:

HYPOTHETICAL

A vehicle insurer using image recognition to assess car damage and automatically pay claims may design a process for customers to appeal contentious AI assessments to a human claims adjustor. The faster settlement process holds more value for customers than explanations of why their car was wrongly classified.

When none of the previously discussed approaches are feasible, institutions may need to refrain from using an AI model

Description:

In some cases, customers, business users and regulators may need to understand an AI model beyond what is technically feasible or economically viable at this time.

Providing transparent and challengeable logic **may not be technically feasible or computationally efficient** given the current state of technical maturity for "explainable AI" techniques

Some approaches of explainability **may undermine the accuracy of the model**, rendering it inadequate for the intended task or inferior to other less opaque approaches[15]

Some approaches of explainability may be **cost-inefficient** relative to the incremental benefits of using the AI application

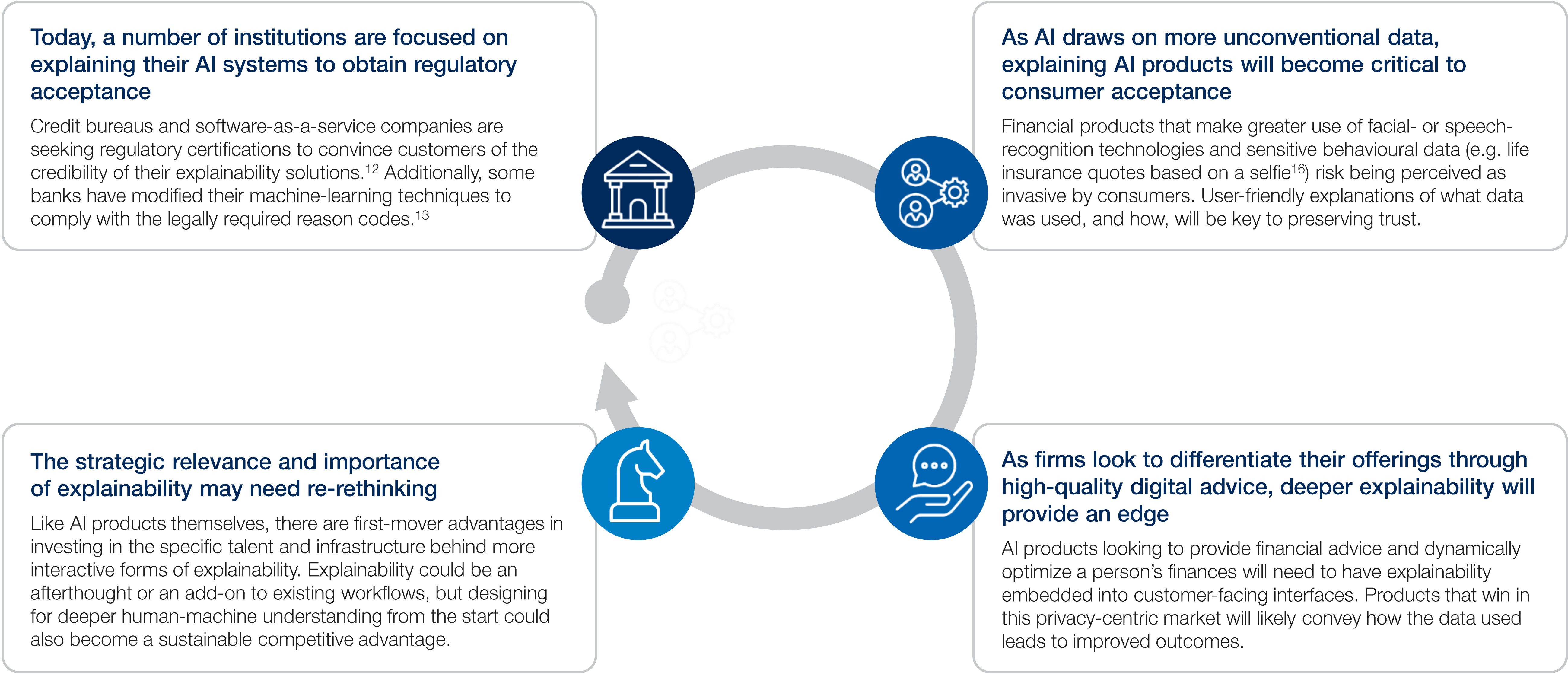
For critical business processes, some approaches may be **unacceptable** to users: customers, business users or regulators

Do not use the AI

As a result of one or more of these factors, an institution may choose not to deploy an opaque model. However, it may choose to invest in R&D to address these limitations by hiring AI subject matter experts to address the limitations of current approaches to explainability, and by adapting "explainable AI" technologies from other industries to financial services contexts.

AI explainability: Conclusion

Explainability is not a one-time decision; ensuring ongoing informed trust requires a periodic re-evaluation of a financial institution’s AI strategy as technology evolves



Today, a number of institutions are focused on explaining their AI systems to obtain regulatory acceptance

Credit bureaus and software-as-a-service companies are seeking regulatory certifications to convince customers of the credibility of their explainability solutions.[12] Additionally, some banks have modified their machine-learning techniques to comply with the legally required reason codes.[13]

As AI draws on more unconventional data, explaining AI products will become critical to consumer acceptance

Financial products that make greater use of facial- or speech-recognition technologies and sensitive behavioural data (e.g. life insurance quotes based on a selfie[16]) risk being perceived as invasive by consumers. User-friendly explanations of what data was used, and how, will be key to preserving trust.

As firms look to differentiate their offerings through high-quality digital advice, deeper explainability will provide an edge

AI products looking to provide financial advice and dynamically optimize a person’s finances will need to have explainability embedded into customer-facing interfaces. Products that win in this privacy-centric market will likely convey how the data used leads to improved outcomes.

The strategic relevance and importance of explainability may need re-rethinking

Like AI products themselves, there are first-mover advantages in investing in the specific talent and infrastructure behind more interactive forms of explainability. Explainability could be an afterthought or an add-on to existing workflows, but designing for deeper human-machine understanding from the start could also become a sustainable competitive advantage.

The opportunities for investments in explainable AI go beyond the basic enablement and governance of AI; they present opportunities for strategic differentiation

Investing in user-centric explainability of AI systems creates strategic opportunities to…

Deeply engage with customers

While today, financial institutions often explain AI decisions to customers using established reason codes, the explainable AI systems of the future could **interact with customers through natural and context-aware dialogue**. Adapting explanations of an AI system to a customer's level of knowledge can be a strategic differentiator advantage and strengthen consumer trust.

What might this look like in the future?

Explainers of the future could build on knowledge of how humans reason and learn, as well as state-of-the-art human-computer interfaces to create explanations that inform, persuade and improve a customer's future financial decisions in tangible ways.

Where has this been done before?

MYCIN, an explainable therapy advisory system designed in 1975 during the first wave of explainable AI, allowed medical students to ask questions such as "Why didn't you ask about X or conclude Y?" They could even enter drug regimens and have them compared to the expert system's "critiquing model".[18]

Meaningfully augment employees' skills

While explainability interfaces today can provide a range of non-experts with access to AI insights, the explainable AI systems of the future could be designed to **collaboratively solve problems with business users**. Human-machine workflows could be redesigned around these newly enabled interactions.

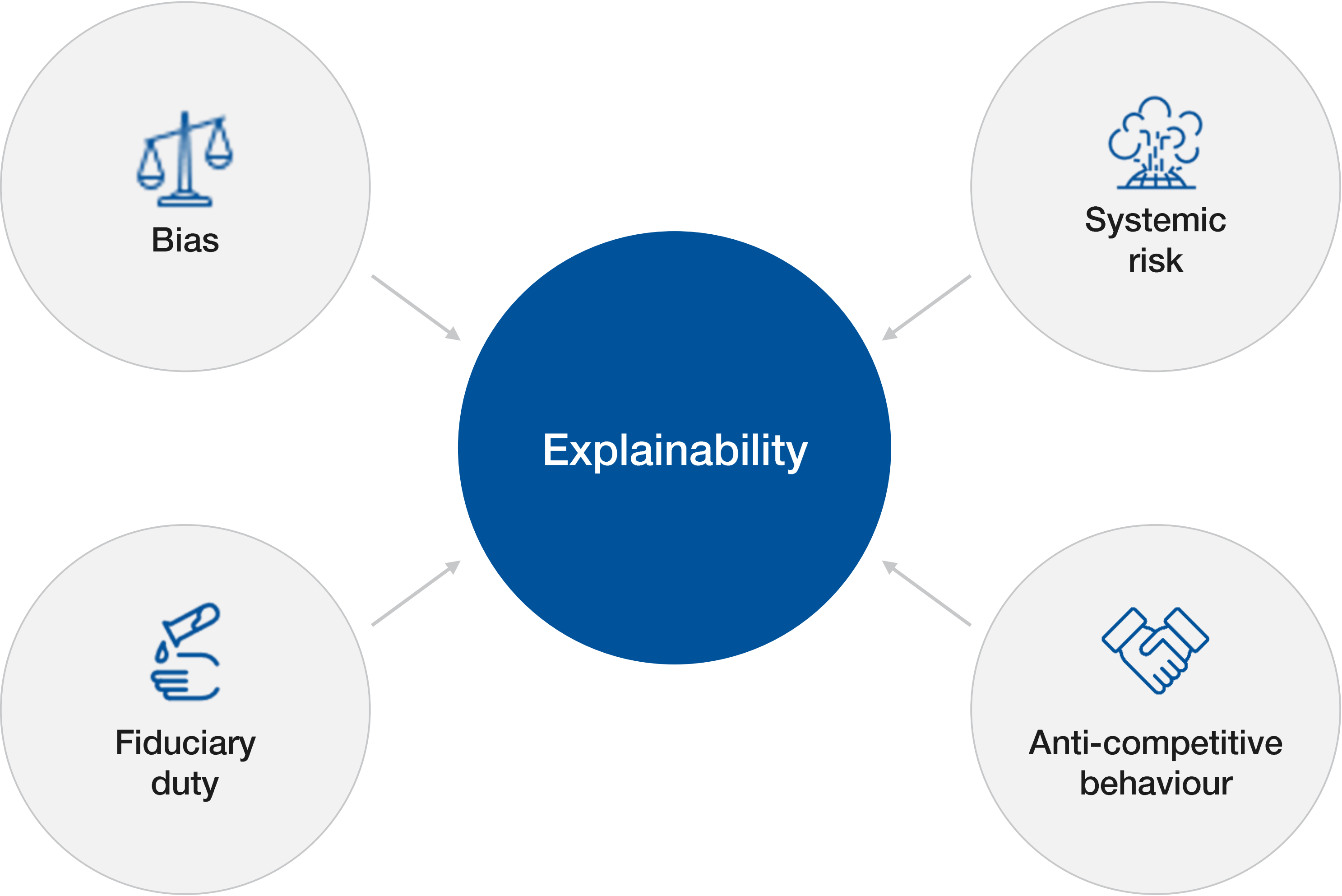
What might this look like in the future?

Explainers of the future could augment employee skills by tailoring explanations to individuals' backgrounds and goals, and clarify misunderstood explanations or elaborate in the context of dialogue that has already occurred. They could serve as digital assistants to novices, helping them understand a new domain.

Where has this been done before?

SHERLOCK, an explainable "intelligent" tutor for debugging electronic circuits built in 1995, interactively trained senior associates using a knowledge base acquired from highly proficient technicians. It was found that 25 hours of Sherlock training was the equivalent of four years of on-the-job training.[18]

The need to provide explainability in AI systems is deeply linked to the other challenges discussed in this report



Algorithmic bias may go undetected since some forms of AI lack transparency on the inferences being drawn from the data

Bias

Systemic risk

Systemic risk may be significantly more difficult to anticipate and react to in a world where numerous highly complex and opaque models are interacting with each other in real time

Explainability

Obstacles to defending the suitability of AI-based financial decisions may limit the capabilities of next generation robo-advisory businesses

Fiduciary duty

Anti-competitive behaviour

Complex and uninterpretable price-setting AI could learn to collude in ways that leaves no conclusive evidence

Looking forward

1 **For automated machine learning to take hold across financial services, these cutting-edge systems will need to explain themselves in ways that enable meaningful risk management and accountability across interconnected institutions**. Explaining AI decisions is becoming just as critical for B2B providers of packaged, proprietary solutions as it is for retail financial institutions.

2 **"Interrogating the underlying AI" is not the universally "best" approach to managing explainability: Different methods are appropriate in different contexts, depending on *why* an explanation is needed.** Selecting the appropriate explainability approach to provide users with specific types of "informed trust" will be critical to the successful commercial adoption of AI in financial services.

3 **While certain applications favour simple, approximate explanations of the AI's decision-making, these may not be acceptable in other cases.** Industry dialogue will be important to establish when granular and comprehensive rationale is a requirement. It may be enough, for example, for an explainable insurance underwriting AI to answer only a specific question – "was the quote discriminatory?".

4 **Various forms of explainability are useful to address algorithmic bias** – "interrogating AI" for its detection, "deploying safeguards" for its prevention and providing the underlying rationale to evidence fairness. But explainable AI does not solve the bias problem in full; institutions will still need to supplement this with broader solutions to prevent real-world bias from being reinforced within algorithms.*

5 **Gaining customer or employee trust through AI explainability could become a strategic choice** and require deeper investment. Interactive forms of explainability that allow users to probe the rationale behind AI-based financial advice in user-friendly ways, for example, will differentiate such products from their less transparent competitors in the future. Indeed, this is the focus of an active research community around a new field of "xAI" or explainable AI.

* For further details, see the chapter "Bias and fairness"

References

1. https://www.risk.net/asset-management/6119616/blackrock-shelves-unexplainable-ai-liquidity-models
2. https://blogs.wsj.com/cio/2018/05/11/bank-of-america-confronts-ais-black-box-with-fraud-detection-effort/
3. https://www.zdnet.com/article/capital-one-ai-chief-sees-path-to-explainable-ai/
4. https://www.bafin.de/SharedDocs/Downloads/EN/dl_bdai_studie_en.html;jsessionid=DF0654DBF78C8FA8AE5E4AA58148CB44.1_cid381?nn=9866146
5. https://www.mas.gov.sg/news/media-releases/2018/mas-introduces-new-feat-principles-to-promote-responsible-use-of-ai-and-data-analytics
6. https://www.fsb.org/wp-content/uploads/P011117.pdf
7. https://www.integrate.ai/responsible-ai-in-consumer-enterprise/
8. https://insurancedatascience.org/downloads/London2018/Session5/Oliver_Laslett.pdf
9. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
10. https://www.bbva.com/ndb/en/article/artificial-intelligence-in-financial-services/
11. https://www.reubenbinns.com/blog/how-to-comply-with-gdpr-article-22-automated-credit-decisions/
12. https://www.prnewswire.com/news-releases/equifax-receives-utility-patent-for-innovative-neurodecision-technology-300763153.html
13. https://www.iif.com/portals/0/Files/private/32370132_machine_learning_explainability_nov_2018.pdf
14. http://www.xai-am.com/explainableaipart1
15. https://twimlai.com/twiml-talk-73-exploring-black-box-predictions-sam-ritchie/
16. https://www.consumerfinancialserviceswatch.com/2019/01/revamped-relief-the-cfpbs-proposed-rule-to-improve-its-no-action-letter-program-and-to-establish-a-regulatory-sandbox/
17. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf
18. https://arxiv.org/ftp/arxiv/papers/1902/1902.01876.pdf

Systemic risk and AI

In this chapter, we will explore:

- Risks (Where are the real risks? Which fears are overstated?)

Chapter summary

Widespread adoption of AI has the potential to create a fundamentally different kind of financial system, one in which the interconnections between humans and machines grow, even as humans struggle to interpret the opaque behaviours of AI systems. As a result, crises and critical events may occur more frequently and and market shocks may be intensified.

Emerging risks will no longer sit neatly inside a supervised institution, but instead will be dispersed across an interconnected set of actors that includes small specialized fintechs and large technology companies. **Supervisory authorities will need to reinvent themselves as hubs for system-wide intelligence** lest increased system complexity erodes transparency and threatens investor confidence during crises.

Carefully designed human-machine relationships will be crucial to avoid weakening defensive guard rails in a machine-integrated financial system. As humans are increasingly allocated the more complex tasks in automated processes, funnelling them key information on evolving market and consumer behaviours, through explainable systems or other interfaces, will be critical to preserving their skills.

The use of ever more complex technology in the financial sector will require stakeholders to absorb lessons from other safety-critical industries (e.g. aerospace).[1] This may involve developing “deviance-monitoring” processes to analyse market shocks and organizational responses after the event and without blame, as well as redesigning machine-enabled processes to prevent critical skills erosion.

The rise of AI-powered systems is raising new questions about financial stability and the management of systemic risk

> “The same qualities that make AI so useful for micro-prudential authorities are also why it could destabilize the financial system…”
>
> – Jon Danielsson, LSE Systemic Risk Centre[2]

> “Some say [algorithmic flash crash] incidents are telling preludes to a financial disaster…”
>
> – Venture Beat[3]

Some worry complex machines are the problem

Others worry it’s the humans behind the machines…

> “AI could also attract new, systemically important providers and put to the test old definitions of systemic importance”
>
> – Thorsten Pötzsch, BaFin[4]

> “This is a story about computers making markets less stable, [but also] a story about humans being inscrutable to the algorithms…”
>
> – Matt Levine, Bloomberg[5]

In the 1970s, the advent of fast computing transformed financial markets, causing players to interact in new and sometimes unstable ways

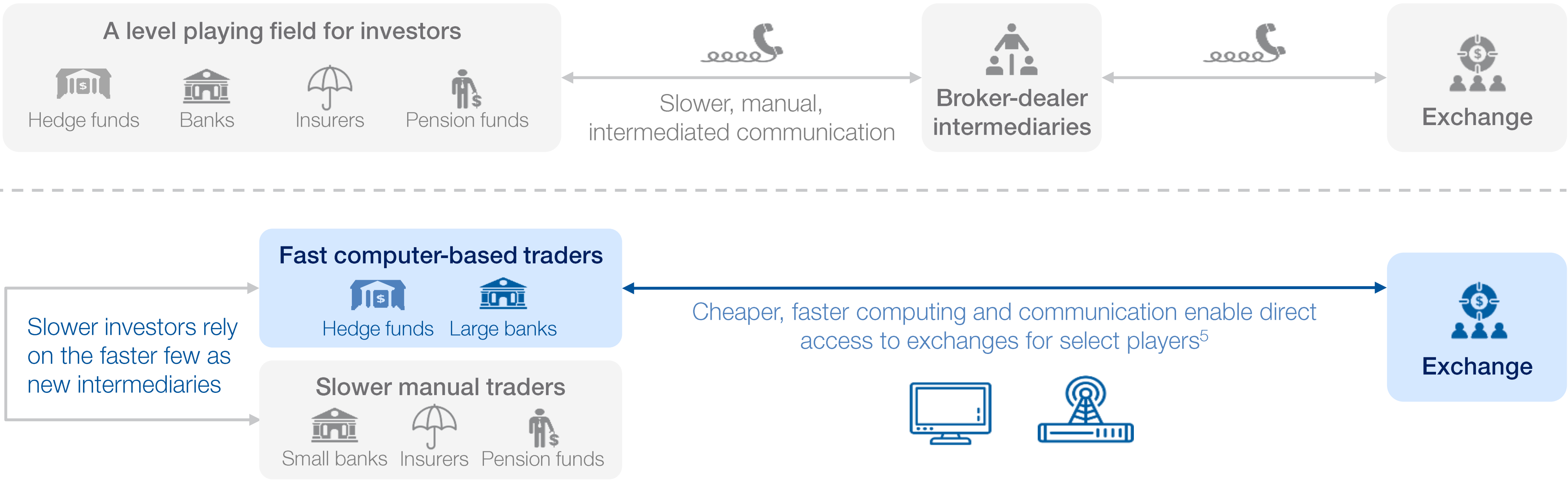
The advent of automated trading in the 1970s radically shifted the market landscape:

Pre-1970:

A one-speed world

The advent of computerized trading:

A landscape of slow and fast movers



A level playing field for investors

Hedge funds

Banks

Insurers

Pension funds

Slower, manual, intermediated communication

Broker-dealer intermediaries

Exchange

Slower investors rely on the faster few as new intermediaries

Fast computer-based traders

Hedge funds

Large banks

Slower manual traders

Small banks

Insurers

Pension funds

Cheaper, faster computing and communication enable direct access to exchanges for select players[5]

Exchange

A new market landscape created new vulnerabilities:

The market was split between slow and fast movers…

Only the largest hedge funds and investment banks could afford to implement computerized trading, which gave them an advantage of speed over other market actors. Institutional investors still relied on slower, manual ways of trading with exchanges.

…causing players to interact in unexpected and unstable ways.

During volatile markets, algorithmic traders were able to move at lightning speeds, selling much quicker than institutional investors. This created a vulnerable environment where large algorithmic sales plunged market prices in minutes as slower investors, in their panic, left no buyers to absorb the sales.

Today, the use of AI across financial services is once again transforming the market landscape, making interactions between market players and machines more complex

Artificial intelligence is...

...creating new forms of market interaction and interconnectedness

AI is be applied to a wide variety of use cases across every sub-sector of financial services, causing humans and machines to interact more frequently.

...making market interactions harder to understand

AI systems can be "black boxes", making it difficult for investors to interpret changing market dynamics or identify emerging risks.

...multiplying pathways to accident and failure

Wider adoption of AI and its enabling technologies increases the likelihood of smaller, innocuous accidents combining to create opportunities for systemic failure. This also makes it hard to blame any one specific root cause.[1]

While AI contributes to speed and efficiency gains across the financial system, we must also assess its risk to financial stability.

In this chapter, we explore how AI adoption could amplify system-wide risks and erode established defences of financial stability…

Machines herding to move markets

Machines optimizing to destabilizing ends

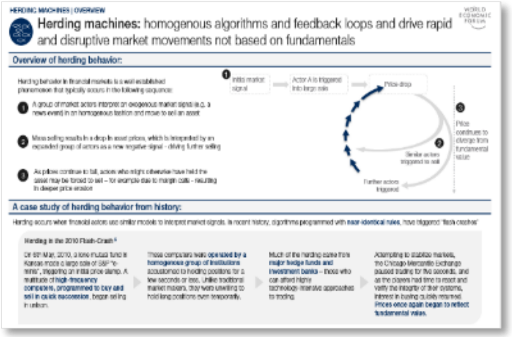
Uninterpretable machines causing human panic

Eroding crisis-prevention skills

Normalizing deviance

Weakening system guard rails

We explore the building blocks of each risk in four parts:


HERDING MACHINES | OVERVIEW

Herding machines: homogenous algorithms and feedback loops and drive rapid and disruptive market movements not based on fundamentals

Overview of herding behavior:

Herding behavior in financial markets is a well established phenomenon that typically occurs in the following sequence:

1. A group of market actors interpret an exogenous market signal (e.g. a news event) in an homogenous fashion and move to sell an asset
2. Mass selling results in a drop in asset prices, which is interpreted by an expanded group of actors as a new negative signal – driving further selling
3. As prices continue to fall, actors who might otherwise have held the asset may be forced to sell – for example due to margin calls – resulting in deeper price erosion

Initial market signal
Actor A is triggered into large sale
Price drop
Similar actors triggered to sell
Further actors triggered
Price continues to diverge from fundamental value

A case study of herding behavior from history:

Herding occurs when financial actors use similar models to interpret market signals. In recent history, algorithms programmed with **near-identical rules**, have triggered 'flash crashes'

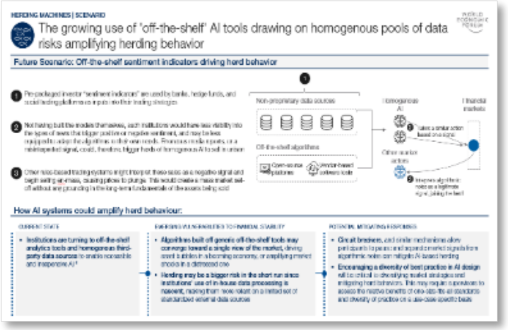
Herding in the 2010 Flash-Crash

On 6th May, 2010, a lone mutual fund in Kansas made a large sale of S&P "e-minis", triggering an initial price slump. A multitude of **high-frequency computers, programmed to buy and sell in quick succession**, began selling in unison.

These companies were **operated by a homogenous group of institutions** accustomed to holding positions for a few seconds or less. Unlike traditional market makers, they were unwilling to hold long positions even temporarily.

Much of the herding came from **major hedge funds and investment banks** – those who can afford highly technology-intensive approaches to trading.

Attempting to stabilize markets, the Chicago Mercantile Exchange paused trading for five seconds, and as the players had time to react and verify the integrity of their systems, interest in buying quickly returned. **Prices once again began to reflect fundamental value.**


HERDING MACHINES | SCENARIO

The growing use of 'off-the-shelf' AI tools drawing on homogenous pools of data risks amplifying herding behavior

Future Scenario: Off-the-shelf sentiment indicators driving herd behavior

1. Pre-packaged investor "sentiment indicators" are used by banks, hedge funds, and social trading platforms as inputs into their trading strategies
2. Not having built the models themselves, each institutions would have less visibility into the types of news that trigger positive or negative sentiment, and may be less equipped to adapt the algorithms to their own needs. Erroneous media reports, or a misinterpreted signal, could, therefore, trigger herds of homogenous AI to sell in unison
3. Other rules-based trading systems might interpret these sales as a negative signal and begin selling en-mass, causing prices to plunge. This would create a mass market sell-off without any grounding in the long-term fundamentals of the assets being sold

How AI systems could amplify herd behaviour:

CURRENT STATE

- **Institutions are turning to off-the-shelf analytics tools and homogenous third-party data sources** to enable accessible and inexpensive AI

EMERGING VULNERABILITIES TO FINANCIAL STABILITY

- **Algorithms built off generic off-the-shelf tools may converge toward a single view of the market,** driving asset bubbles in a booming economy, or amplifying market shocks in a distressed one
- **Herding may be a bigger risk in the short run since institutions' use of in-house data processing is nascent,** making them more reliant on a limited set of standardized external data sources

POTENTIAL MITIGATING RESPONSES

- **Circuit breakers,** and similar mechanisms allow participants to pause and separate market signals from algorithmic noise can mitigate AI-based herding
- **Encouraging a diversity of best practice in AI design** will be critical to diversifying market strategies and mitigating herd behaviors.

1

Overview

Description of the risky behaviour

2

Case study

Instances from the past

3

Future scenarios

A scenario of how AI increases the risk

4

Conclusion

Summary and mitigating responses

We explore each eroding defence in three parts:

ERODING HUMAN SKILLS

Eroding skills: As human financial talent becomes more reliant on AI, they may gradually lose skills needed to challenge these systems or respond to crises

Overview - How AI systems may amplify machine over-reliance:

1. Complex automated systems are being introduced to support employees in select tasks.
2. **As automated systems replace employees in tasks of increasing complexity,** employees become further removed from the market dynamics, intuition and understanding of risk that once drove their business decisions.
3. **Machine over-reliance and skill erosion can have this quietly overlooked, until a crisis exposes cracks in the system,** at which point, humans may no longer be equipped to take over for automated systems that have been shut down

Case study: Over-relying on models in the 2008 Crisis

Strategies to preserve organizational skillsets:

Designing human-in-the-loop processes with a focus on visibility:

Using explainable AI systems to train employees:

1

Overview

Description of the risky behaviour

2

Case study

Instances from the past

3

Strategies for prevention

Discussion of how institutions may overcome these limitations in the future

Systemic risk and AI:
New sources of risk

Machines herding to move markets

Machines optimizing to destabilizing ends

Uninterpretable machines causing human panic

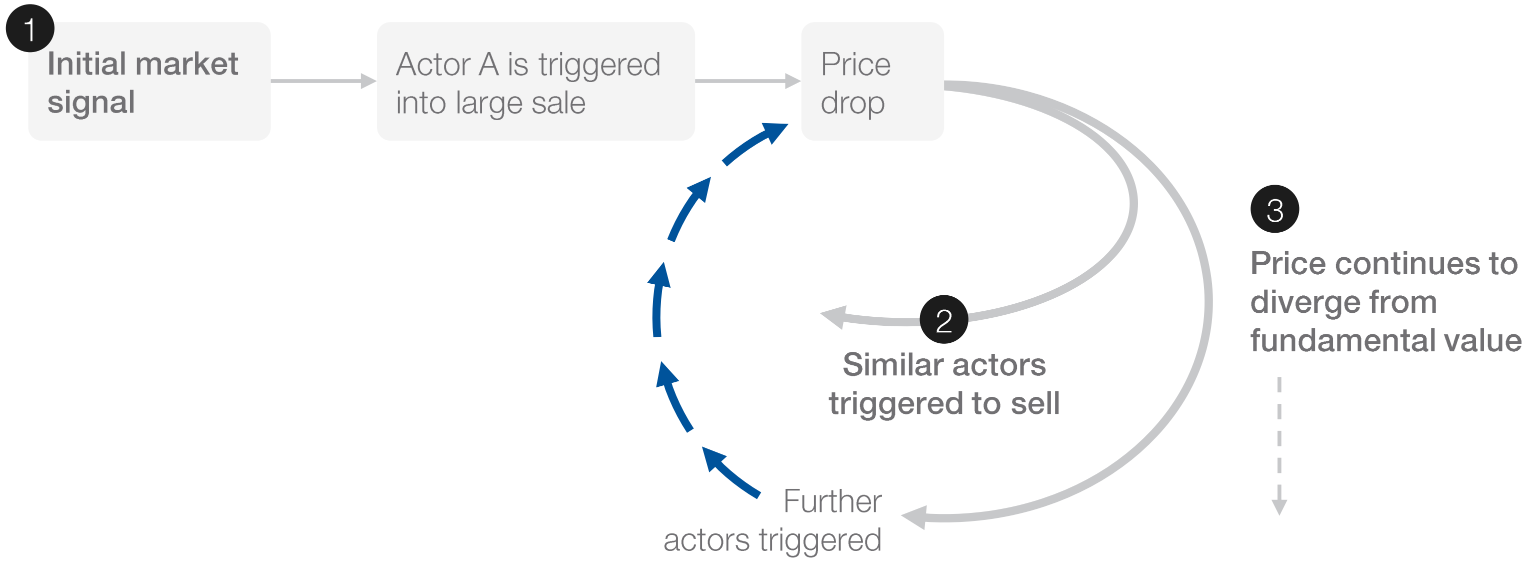
Herding machines: Homogenous algorithms and feedback loops can drive rapid and disruptive market movements not based on fundamental asset values

Overview of herding behaviour:

1 A group of market actors interpret an exogenous market signal (e.g. a news event) in a homogenous fashion and move to sell an asset.

2 Mass selling results in a drop in asset prices, which is interpreted by an expanded group of actors as a new negative signal – driving further selling.

3 As prices continue to fall, actors who might otherwise have held the asset may be forced to sell – for example, due to margin calls – resulting in deeper price erosion.


1
Initial market signal
Actor A is triggered into large sale
Price drop
2
Similar actors triggered to sell
Further actors triggered
3
Price continues to diverge from fundamental value

Case study: Herding in the 2010 Flash Crash[1]

Herding occurs when financial actors use similar models to interpret market signals. In recent history, algorithms programmed with **near-identical rules** have triggered “flash crashes”:

On 6 May 2010, a lone mutual fund in Kansas made a large sale of S&P E-mini futures, triggering an initial price slump. A multitude of **high-frequency computers, programmed to buy and sell in quick succession**, began selling in unison.

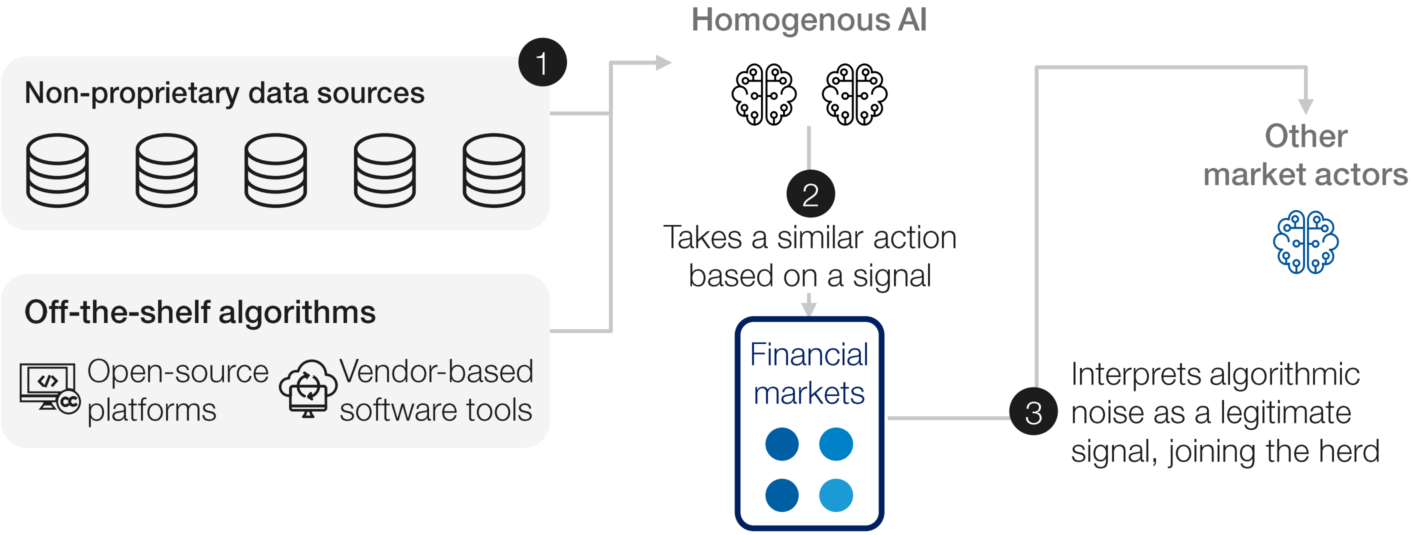
These computers were **operated by a homogenous group of institutions** accustomed to holding positions for a few seconds or less. Unlike traditional market makers, they were unwilling to hold long positions even temporarily.

Much of the herding came from **major hedge funds and investment banks** – those who can afford highly technology-intensive approaches to trading.

Attempting to stabilize markets, the Chicago Mercantile Exchange paused trading for five seconds and, as the players had time to react and verify the integrity of their systems, interest in buying quickly returned. **Prices once again began to reflect fundamental value.**

The growing use of 'off-the-shelf' AI tools drawing on homogenous pools of data risks intensifying herding behaviour

Future scenario: off-the-shelf sentiment indicators driving herd behaviour


Non-proprietary data sources
Off-the-shelf algorithms
Open-source platforms
Vendor-based software tools
1
Homogenous AI
2
Takes a similar action based on a signal
Financial markets
3
Interprets algorithmic noise as a legitimate signal, joining the herd
Other market actors

1 Pre-packaged investor "sentiment indicators" are used by banks, hedge funds and social trading platforms as inputs into their trading strategies.

2 Not having built the models themselves, such institutions would have less awareness of the types of news that trigger positive or negative sentiment and may be less equipped to adapt the algorithms to their own needs. Erroneous media reports, or a misinterpreted signal, could, therefore, trigger herds of homogenous AI to sell in unison.

3 Other rules-based trading systems might interpret these sales as a negative signal and begin selling en masse, causing prices to plunge. This would create a mass-market sell-off without any grounding in the long-term fundamentals of the assets being sold.

How AI systems could intensify herd behaviour:

Current state

- **Institutions are turning to off-the-shelf analytics tools and homogenous third-party data sources** to enable accessible and inexpensive AI.[6]

Emerging threats to financial stability

- **Algorithms built from generic off-the-shelf tools may converge towards a single view of the market**, driving asset bubbles in a booming economy, or magnifying market shocks in a distressed one.
- **Herding may be a bigger risk in the short run, since institutions' use of in-house data processing is nascent, making them more reliant on a limited set of standardized external data sources.**

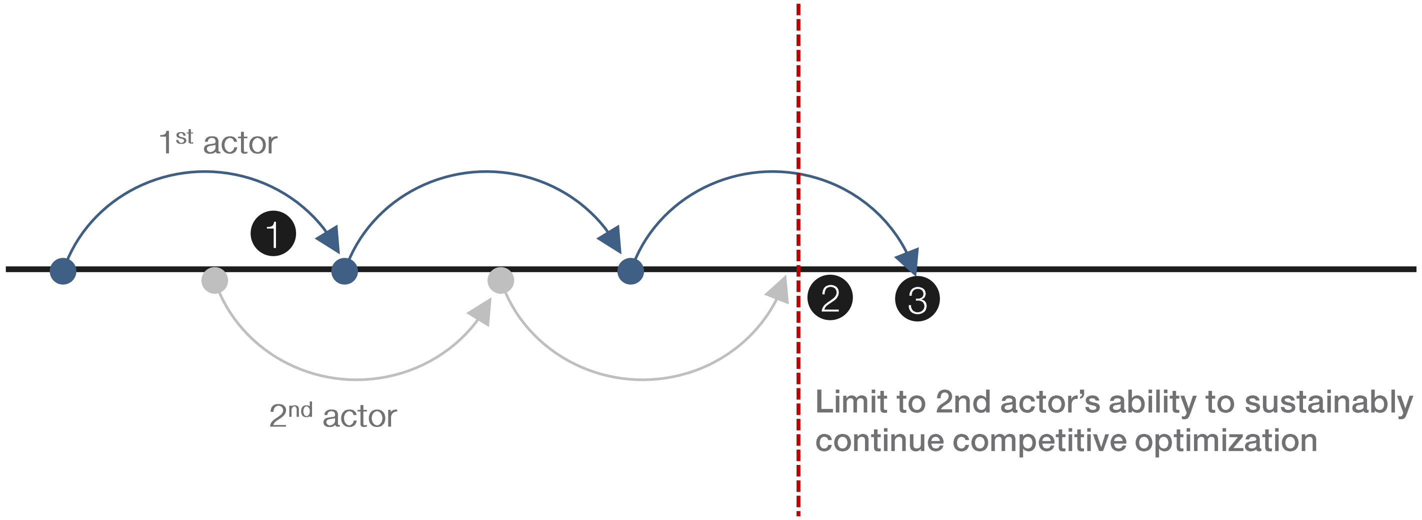
Potential mitigating responses

- **Circuit breakers** and similar mechanisms that allow participants to pause and separate market signals from algorithmic noise can mitigate AI-based herding.
- **Encouraging a diversity of best practices in AI design** will be critical to diversifying market strategies and mitigating herd behaviours. However, achieving diversity may compromise a baseline of good practice, and will require supervisors to navigate this trade-off.

Machines optimizing to destabilizing ends: Optimizing algorithms locked in competition with each other could inadvertently destabilize markets

Overview of destabilizing competition:

1. Two AI systems continuously bid against each other, optimizing their actions to achieve a single objective, e.g. the highest market price or return.
2. The average market price continues to rise as they repeatedly outbid each other, until one actor is no longer able to sustain its bids (e.g. due to profitability constraints).
3. Over time this competitive optimization may lead to a deterioration of actors' balance sheets, encouraging riskier behaviour in order to maintain profitability, or leaving them out of the market completely.

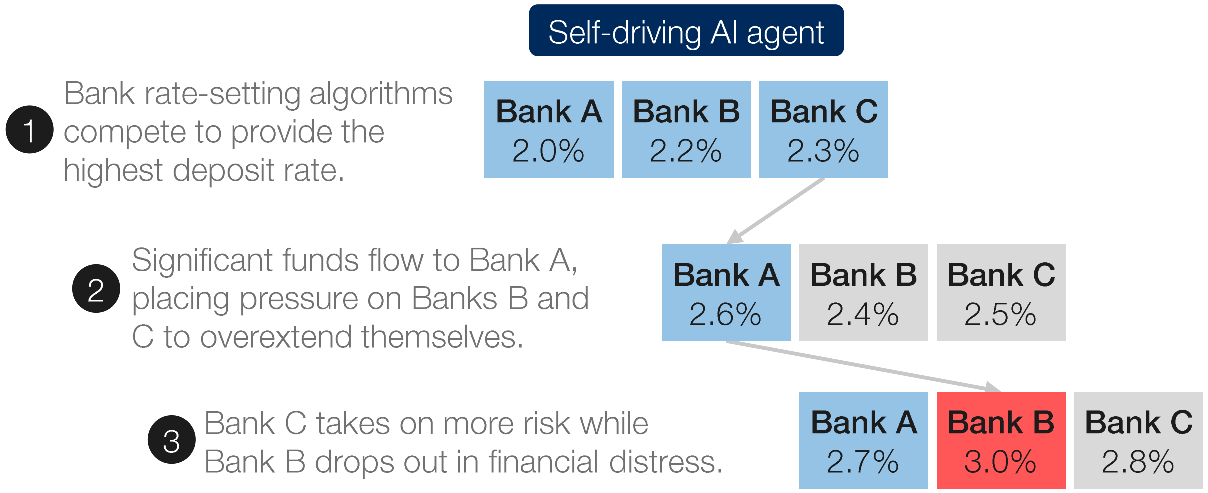

1st actor
1
2nd actor
2
3
Limit to 2nd actor's ability to sustainably continue competitive optimization

Case study:

Our research did not encounter a similar case study from economic history.

The risk of destabilizing competition could expand significantly as AI ‘democratizes’ access to automated real-time optimization

Future scenario: Bank rate-setting algorithms compete to attract consumers with the best deposit rates


Self-driving AI agent
1 Bank rate-setting algorithms compete to provide the highest deposit rate.
Bank A 2.0%
Bank B 2.2%
Bank C 2.3%
2 Significant funds flow to Bank A, placing pressure on Banks B and C to overextend themselves.
Bank A 2.6%
Bank B 2.4%
Bank C 2.5%
3 Bank C takes on more risk while Bank B drops out in financial distress.
Bank A 2.7%
Bank B 3.0%
Bank C 2.8%

1 A growing number of institutions strive to build “self-driving agents” that continuously optimize customer cash flows and automate switching between products.* One application of these agents could be to direct deposits towards the bank offering the highest deposit rate.

2 In searching for the highest rate, the self-driving AI would make financial providers compete for customer deposits and deprive them of a stable deposit base. One consequence of this process would be a rise in the average cost of funds.

3 As multiple bank rate-setting algorithms attempt to outbid each other, multiple institutions might find their business models increasingly unsustainable or be driven to take on more risk in their lending operations.

Wider AI adoption could extend the destabilizing impacts of competing machines to digital financial marketplaces:

Current state

- Fintechs and challenger banks have begun to battle for distribution by offering customers competitive rates for savings and deposits.[7]
- AI-based self-driving finance agents have begun optimizing personal finances both within and across financial institutions, but they are still nascent in their offerings.[8]

Emerging threats to financial stability

- **Faster movements of customer funds as a result of AI-based pricing optimization could threaten bank liquidity and solvency over time**:
  - In directing customers to providers with the best rates, such algorithms could cause rapid, even unsustainable fluctuations in institutions’ assets and liabilities, which have traditionally been held in sync.
  - Algorithmic bidding based on non-rate factors could reinforce negative feedback loops. A Yelp-equivalent rating pointing to a provider’s poor financial health, for instance, may further direct funds away from the provider, deepening its distress.

Potential mitigating responses

- **Requiring a level of diversification** in how ‘self-driving finance’ AI allocate consumer savings or funds between financial providers will prevent the destabilizing consequences of continuously allocating to the highest bidder.
- **Scenario modelling** could be useful to understand where AI’s destabilizing behaviours are likely to occur.

* For an introduction to “Self-driving finance agents,” please refer to the prequel to this report, “The New Physics of Financial Services: Understanding how artificial intelligence is transforming the financial ecosystem”

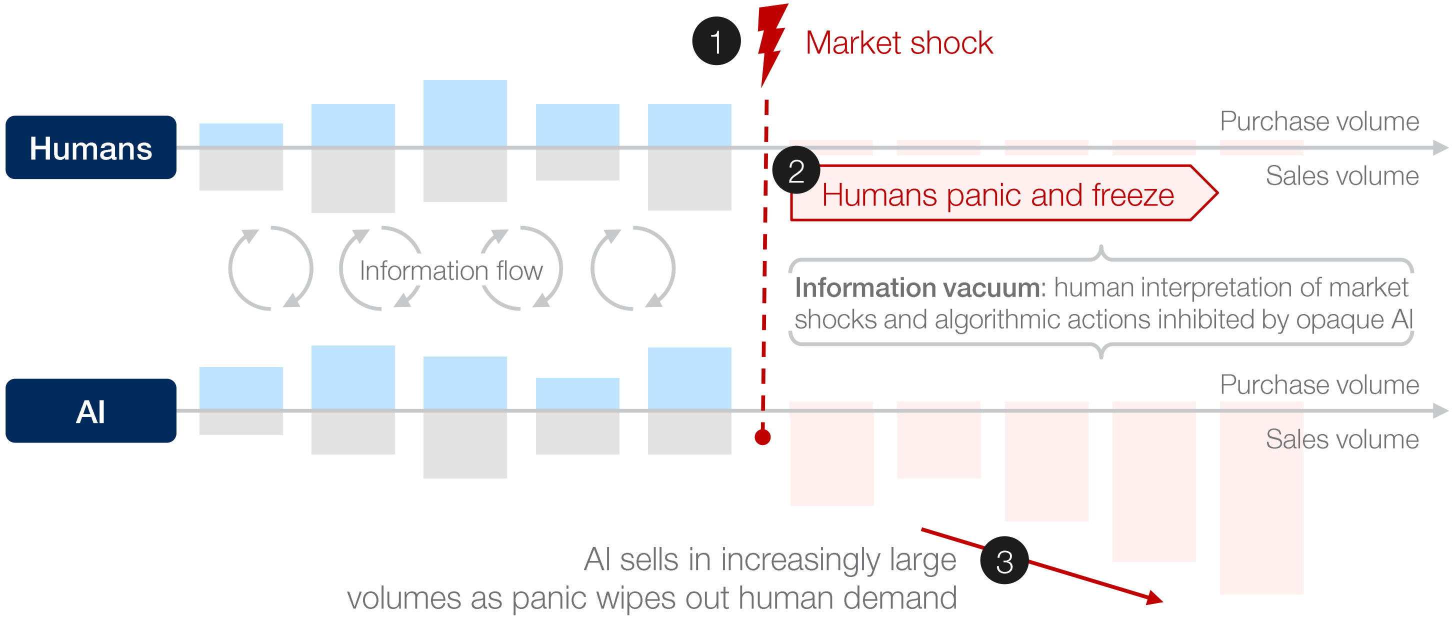
Humans freezing in an information vacuum: Differences in speed and decision logic between machines and humans can cause panic and human inaction in a crisis

Overview of humans freezing in an information vacuum:

1 Automated systems are triggered into a mass sell-off, causing asset prices to go into free fall.

2 Human investors struggle to interpret the rapidly falling market prices, debating whether they are due to changing market fundamentals or algorithmic noise. They freeze, waiting for further information.

3 Human panic further deepens the crash, as algorithms interpret their fear of buying as a lack of interest, and an indication to sell at lower prices. Markets remain out of equilibrium as long as these informational asymmetries persist.


1
Market shock
Humans
Purchase volume
Sales volume
2
Humans panic and freeze
Information flow
Information vacuum: human interpretation of market shocks and algorithmic actions inhibited by opaque AI
AI
Purchase volume
Sales volume
AI sells in increasingly large volumes as panic wipes out human demand
3

Case study: Disconnects in speed accelerate “Black Monday” in 1987[9]

In the past, disconnects in speed between lightning-fast algorithms and slow-to-react humans made it hard to interpret and respond to machine behaviours:

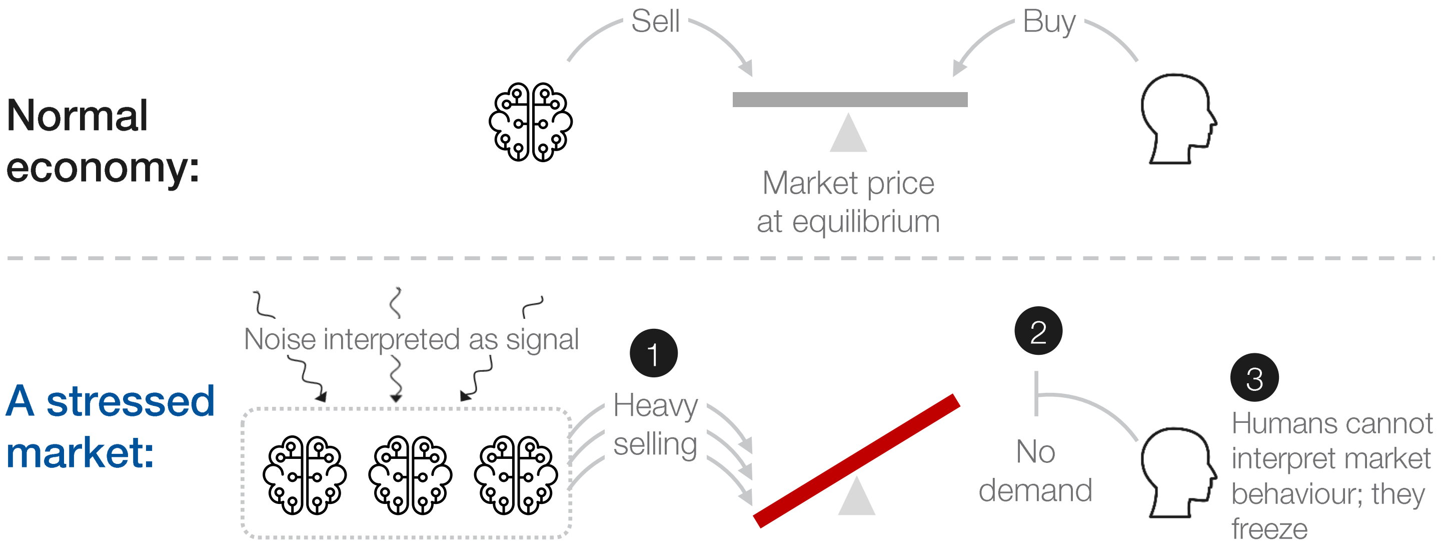
On 19 October 1987, the S&P 500 suffered its worst one-day drop in history, down more than 20%. Rising suspicions about overvalued equities caused **computer programs to sell futures en masse to a market ill-equipped to absorb the supply.**

A lack of liquidity to meet supply was magnified by the **disconnect in time frames for decision-making on either side of the trade.** Equity investors, who could buy only after reasoned consideration, could not react in time to the lightning-fast, computer-assisted S&P floor traders, who kept selling to a falling market.

Faced with increasingly aggressive price drops, **many potential buyers struggled to determine what was happening and backed off completely.** Equity investors lost confidence in the value of the underlying products, drying up the pool of buyers, plunging prices and erasing $500 billion in market value over a few hours.

AI magnifies disconnects in humans and machine reasoning as more algorithms feed on alternative data – often misinterpreting its political or emotional context

Future scenario: Machines misread news sentiment


Normal economy:
Sell
Buy
Market price at equilibrium
A stressed market:
Noise interpreted as signal
1
Heavy selling
2
No demand
3
Humans cannot interpret market behaviour; they freeze

1. A sentiment-analysis AI is triggered into selling en masse by social media political bots spreading rumours about faltering trade negotiations. Other AI pick up on the initial sale, triggering herd selling and a rapid fall in prices.

2. Human investors struggle to interpret the rapidly falling market prices, debating if they are due to changing market fundamentals or algorithmic noise. They freeze, waiting for further information.

3. Human panic further worsens the crash, as algorithms interpret their fear of buying as a lack of interest, and an indication to sell at lower prices. Markets stay out of equilibrium as long as these informational asymmetries persist.

How AI systems magnify disconnects in reasoning and decision-making between human and machines:

Current state

- **The new age of alpha generation will likely be driven by data rather than speed**:
  - With speed-enhancing infrastructure commoditized over the past four years, an increasing number of firms are applying machine learning to non-traditional data in search of excess returns[10]
  - While disconnects in speed have historically been a primary cause of humans “freezing” in a crash, AI’s use and varied interpretation of alternative data is a growing source of disconnect today

Emerging threats to financial stability

- **Certain characteristics of news-processing algorithms are difficult for humans to interpret**:
  - They struggle to weigh divergent opinions, distinguish fact from opinion, or interpret sentiment in its proper political or emotional context[11]
  - They can be tricked by bad actors who manipulate the online public data sources that feed them
  - Errors in their news interpretations can spread virally through markets as other “news-listening” AI watch, learn, and mimic their signals

Potential mitigating responses

- **Circuit breakers may effectively course-correct away from feedback loops** rooted in algorithmic noise. They do so by pausing all activity so that participants have time to verify the truth of their information sources.
- **Designing news and sentiment analysis AI to be resilient to “fake” signals** in alternative data will be critical to preventing the systematic spread of misinformation and will likely be a significant near-term priority for the capital markets industry.

Systemic risk and AI:
Erosion of the financial system's defences

Eroding
crisis-prevention skills

Normalizing
deviance

Weakening system
guard rails

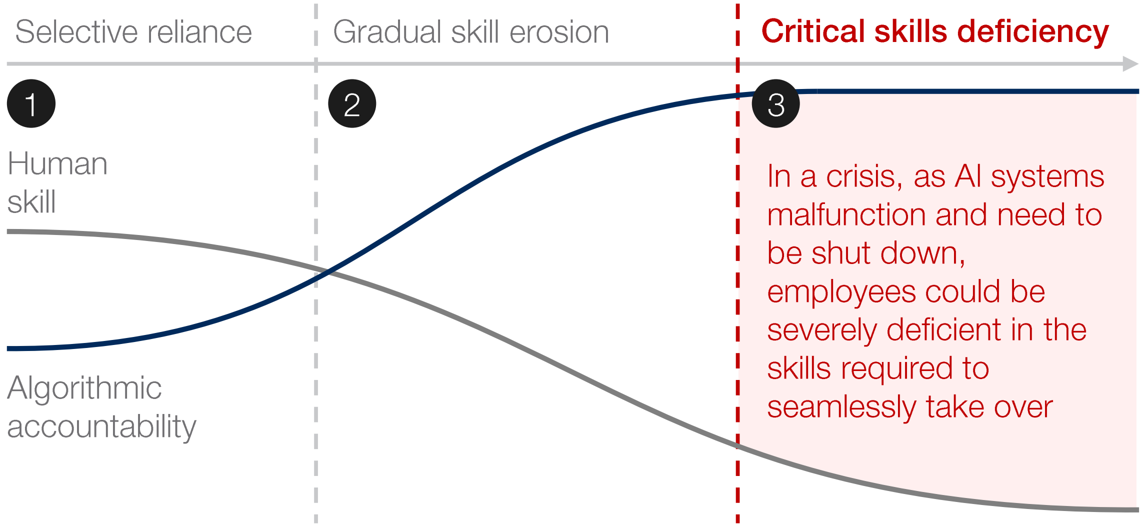
Eroding skills: As human financial talent becomes more reliant on AI, people may gradually lose the skills needed to challenge these systems or respond to crises

Overview: How AI systems may magnify machine overreliance

1 **Complex automated systems are being introduced to support employees in selected tasks**. Understanding them demands expertise, creating an accessibility barrier that makes it difficult for the everyday user to challenge how it is used.

2 **As automated systems replace employees in tasks of increasing complexity, employees become further removed** from the market dynamics, intuition and understanding of risk that once drove their business decisions. For example, AI-enabled traders may gradually lose the instincts developed in floor-based trading.

3 **Machine overreliance and skill erosion can leave risks quietly overlooked, until a crisis exposes cracks in the system**, at which point humans may no longer be equipped to take over for automated systems that have been shut down.



In a crisis, as AI systems malfunction and need to be shut down, employees could be severely deficient in the skills required to seamlessly take over

Case study: Overreliance on models in the 2008 crisis[12]

In the past, a fascination with the latest modelling techniques lulled institutions into a **false sense of analytic security**, one of several reasons for the buildup of risks in the system:

- In the run-up to 2008, business heads at some of the largest global banks blindly trusted the risk models valuing mortgage-backed securities – in part because they felt ill-equipped to **challenge the physicists and mathematicians** that built them.[12]
- Yet these models **incorrectly estimated the products’ risks**, assuming the underlying housing prices would continue to rise as they had for a decade.
- This **left hedge funds and banks highly exposed** as defaults on US subprime mortgages rose to a seven-year high, creating conditions that led to the crisis.

Strategies to preserve organizational skill sets:

As institutions redesign business processes to augment human capabilities, they will need to design around the **need to preserve critical skills and intuition** by:

Creating human-in-the-loop processes with a focus on visibility: As humans are increasingly allocated the more complex tasks in automated processes, funnelling key information on evolving market and consumer behaviours will be critical to preserving their knowledge and intuition.

Using explainable AI systems to train employees: Interactively designed explainable AI systems can serve as digital tutors to novice employees, rapidly transferring expertise acquired from more experienced employees. Similar systems were used to train medical students by “critiquing” their proposed drug regimens.[13]

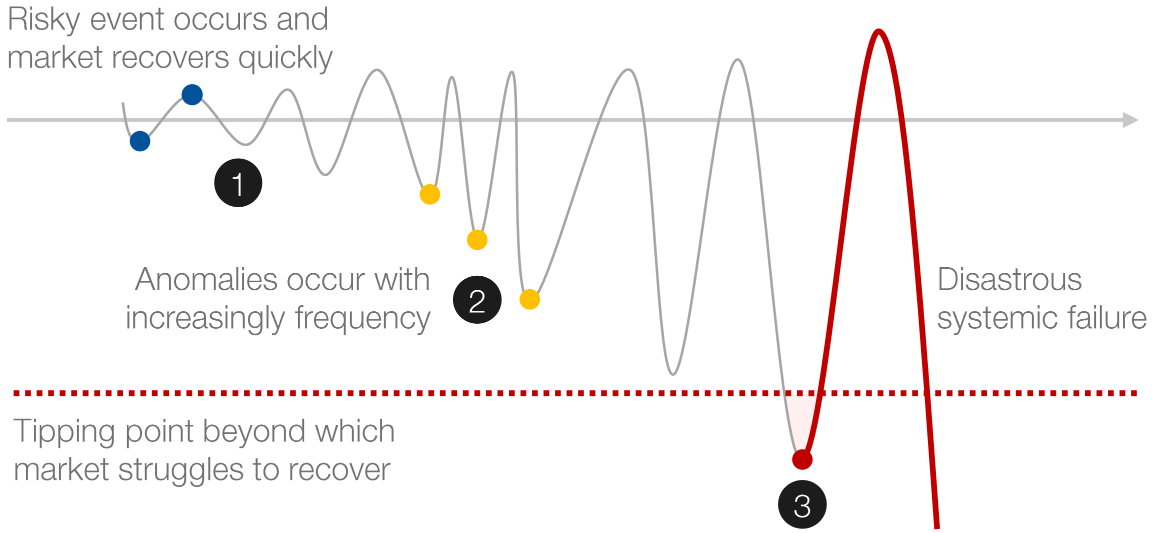
Normalization of deviance: Frequent dislocations within an AI-enabled financial system could gradually cause humans to view risky events as normal

Overview: How interconnected AI systems could cause a social normalization of deviance

1 As the financial system becomes more complex and interconnected, **market destabilizing accidents could occur more frequently.** Recently, a growing number of flash crashes have begun with innocuous technical accidents (e.g. a wayward keystroke by a trader [September 2010], or a software update at the NYSE [November 2010]).[14]

2 **Over time, humans come to view unexpected or risky events as normal**, as a growing number of extreme events occur without disastrous consequence. Normalization results in humans failing to investigate such events further.

3 These extreme events may be eroding the financial system's defences in ways we cannot see or **pushing it ever closer to a "tipping point"** where the event will have an outsized impact on markets.


Risky event occurs and market recovers quickly
1
Anomalies occur with increasingly frequency
2
Disastrous systemic failure
Tipping point beyond which market struggles to recover
3

Case study: The normalization of deviance at NASA[15]

The human tendency to "normalize deviance" has played a key role in the failure of **complex, safety-critical engineering systems** outside of the financial system:

- On 1 February 2003, NASA's space shuttle *Columbia* exploded in midair as it returned to Earth. The crash occurred because a large piece of foam hit its spacecraft wing.
- **This problem with foam had been known for years.** Much smaller foam shrapnel had hit flights on numerous occasions, but such incidents were **routinely dismissed at NASA** because their orbiters came back unharmed.
- Having 'normalized' the threat of shuttle damage from smaller shrapnel, NASA overlooked the possibility of a larger shrapnel striking in a vulnerable area. This paved way for the *Columbia* explosion.

Strategies to prevent "normalization of deviance":

As risky technology is introduced into the financial system, ensuring its safety will require the industry to learn from other **safety-critical industries** (e.g. medicine) by:

Developing "deviance monitoring" processes: Such processes would allow analysis of market shocks, however minor, after the event and without ascribing blame.

Introducing simulation-based approaches to scan for emerging systemic risks: This is especially relevant as risky technology is introduced to the financial system, whose safe operating limits and behaviours are not well understood.

Weakening system guard rails: AI models could be incentivized to 'game' regulatory regimes, slowly undermining control mechanisms in the financial system

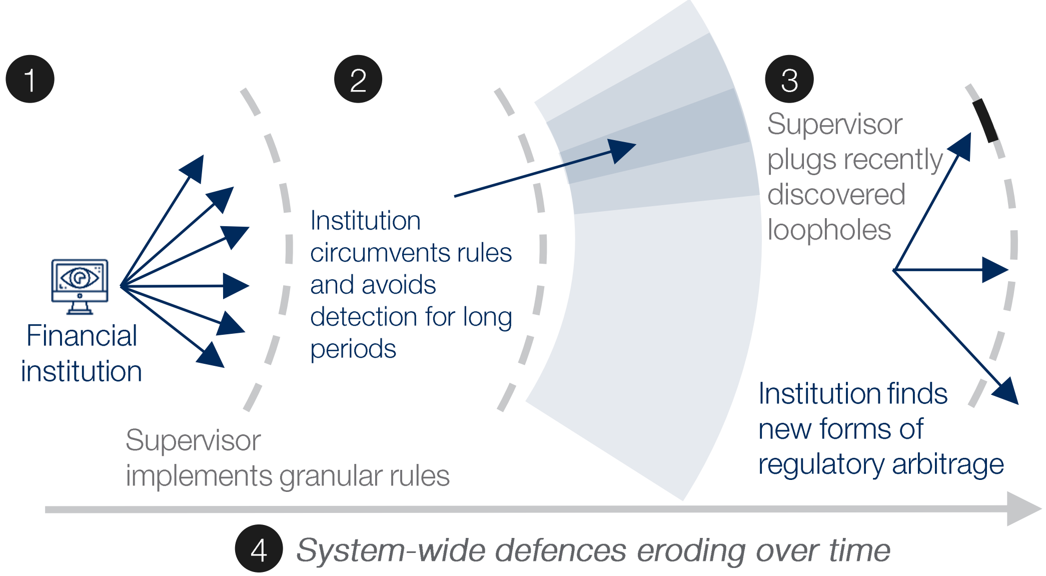
Overview: How AI could magnify gaming behaviours

1. **There is a risk that some firms may use AI optimization solutions to "game" regulatory rules** or predictable supervisory systems. This is especially relevant if optimizing AI is adopted by financial institutions but not by their supervisors.

2. **Such firms may circumvent simple capital rules, concealing their actions behind new garbs of complexity to avoid detection.** They may conceal risks behind complex optimized portfolio structures, or chains of opaque models.

3. **Once these are discovered, supervisors may add more rules to prevent the same breach from recurring**. Unscrupulous firms may then seek new vulnerabilities to exploit.

4. **System-wide defences may erode over time. P**rolonged use of AI to optimize capital, for example, would leave firms with thinner reserves, exposing them to leverage pressures in a crisis.


1
Financial institution
2
Institution circumvents rules and avoids detection for long periods
Supervisor implements granular rules
3
Supervisor plugs recently discovered loopholes
Institution finds new forms of regulatory arbitrage
4
System-wide defences eroding over time

Case study: Gaming the 1981 leverage ratio rule[16]

In the past, some financial players could game simple rules laid out by regulators by **hiding their risks behind complicated transaction structures** and complex new products:

- In 1981, US regulators introduced a simple rule based on a minimum leverage ratio (a bank's equity capital divided by total assets) to prevent the buildup of leverage in the banking sector.
- The rule's simplicity allowed some to game it by replacing low-risk with risky but profitable assets, and using off-balance-sheet vehicles, **where the rule did not apply.**
- In response to these gaming behaviours, regulators proposed a new set of "risk-based" rules, but **different forms of regulatory "gaming" have only emerged since.**

Strategies for reinforcing guard rails:

As regulators move to digitize their supervisory processes, they will need to **manage the risks of becoming increasingly transparent and predictable** to malicious actors:

Stress-testing the impacts of AI-based gaming behaviours in a crisis: Algorithms used to optimize risk capital or collateral, for example, could be stress-tested to ensure firms are not left with thinner reserves during crises.[17]

Introducing unpredictability to micro-prudential monitoring exercises: Historically, not all regulations have been black and white, giving supervisors the leeway to exercise discretion and encourage institutional prudence. Achieving this in a world of digital and programmatic monitoring may require inserting "noise" into the process to avoid gaming (e.g. discretionary "sanity checks" by supervisors, not necessarily prescribed by law).

Systemic risk and AI: Conclusion

Rapidly shifting interlinkages and risks in the financial system make mapping systemic risks and building system resilience a moving target

The underlying landscape of financial services is fundamentally changing in three ways:

1 Financial institutions are becoming highly networked organizations

BBVA bank has transformed into a marketplace offering 60 financial products that other businesses can rebrand and offer to their own customers[18]

BlackRock has transformed its internal portfolio risk operations into a service adopted by 210 institutional clients globally[19]

Several global banks and investment firms now use S&P subsidiary Kensho's AI-powered analytics-as-a-service tool[20]

The challenge:

Left unassessed, the complexity of new operating relationships between highly networked, technology-fuelled institutions could reduce market transparency and threaten investor confidence in a crisis.

2 New types of players are becoming systemically important

Three dominant players provide 55% of the multiservice cloud solutions to financial institutions globally[21]

Kasisto provides the underlying chatbot infrastructure of five globally significant financial institutions[22]

Sentieo provides AI-based data to 700 customers, including leading hedge funds, mutual funds and investment banks[23]

The challenge:

As financial institutions become collectively reliant on a diminishing number of critical technology systems, the risk of critical node failure outside of the traditional financial system increases.

3 New players and business models will drive new modes of machine-rooted interaction

Albert's "genius" product is an early example of a trend towards self-driving finance agents that will interact with multiple financial product providers to optimize personal financial management[24]

New breeds of "quant-fundamental" investors using widely accessible AI tools to improve fundamental analysis may respond to markets in ways that are less easily anticipated[25, 26]

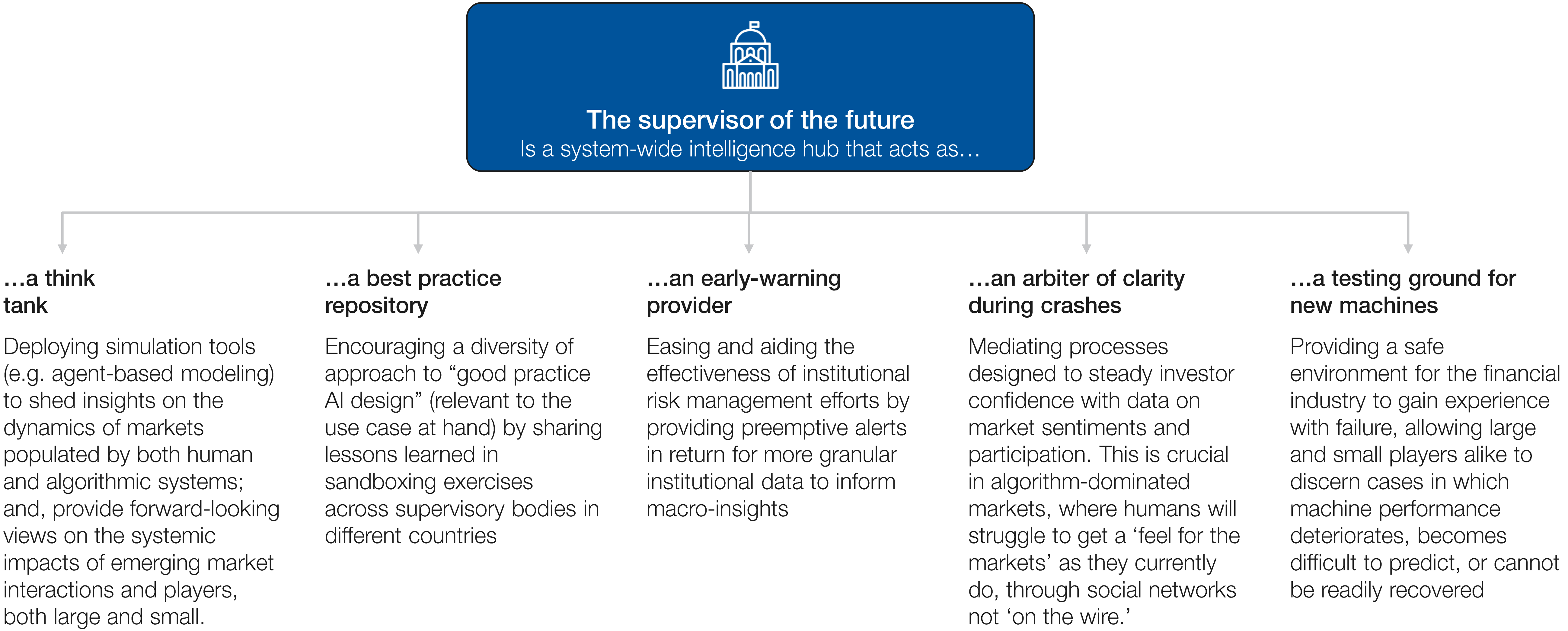
The challenge:

As real-time algorithmic interactions extend beyond trading to marketplaces for deposits, lending or personal financial management, they may create risky, unanticipated market behaviours.

Addressing these challenges requires a reinvention of the financial system's approach to supervision.

Effectively responding to the increased complexity and accelerating pace of change in the financial system will require the supervisors of the future to take on new roles

Supervisors could look to become hubs of system-wide intelligence serving the wider market:



The supervisor of the future

Is a system-wide intelligence hub that acts as…

…a think tank

Deploying simulation tools (e.g. agent-based modeling) to shed insights on the dynamics of markets populated by both human and algorithmic systems; and, provide forward-looking views on the systemic impacts of emerging market interactions and players, both large and small.

…a best practice repository

Encouraging a diversity of approach to "good practice AI design" (relevant to the use case at hand) by sharing lessons learned in sandboxing exercises across supervisory bodies in different countries

…an early-warning provider

Easing and aiding the effectiveness of institutional risk management efforts by providing preemptive alerts in return for more granular institutional data to inform macro-insights

…an arbiter of clarity during crashes

Mediating processes designed to steady investor confidence with data on market sentiments and participation. This is crucial in algorithm-dominated markets, where humans will struggle to get a 'feel for the markets' as they currently do, through social networks not 'on the wire.'

…a testing ground for new machines

Providing a safe environment for the financial industry to gain experience with failure, allowing large and small players alike to discern cases in which machine performance deteriorates, becomes difficult to predict, or cannot be readily recovered

Looking forward

1 **Managing the risk of market panic in an AI-enabled financial system will require two approaches: extending the use of older mitigating tools created for algorithmic trading (e.g. circuit breakers); and creating new mechanisms to provide markets transparency.** This is because AI's systemic risks stem not just from the technology, but from the ways in which humans respond to its opaque behaviours.

2 **Blind reliance on AI and its enabling technologies could erode system-wide guard rails in the long run** – from the skills and intuition of front-line employees to the effectiveness of monitoring mechanisms and regulatory protections. Avoiding these risks will require using tools such as explainable AI to teach and provide ongoing visibility to the humans within automated processes.

3 **Responding to the dynamic risks of a machine-integrated financial system will require regulators to forge deeper, mutually beneficial partnerships with financial institutions**. This could include two-way exchanges of granular institutional data and real-time alerts based on macro insights aggregated by the regulator.

References

1. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/289431/12-1086-future-of-computer-trading-in-financial-markets-report.pdf
2. http://www.systemicrisk.ac.uk/sites/default/files/downloads/publications/SP13.pdf
3. https://venturebeat.com/2017/05/06/ai-powered-trading-raises-new-questions/
4. https://www.bafin.de/SharedDocs/Downloads/EN/dl_bdai_studie_en.html;jsessionid=DF0654DBF78C8FA8AE5E4AA58148CB44.1_cid381?nn=9866146
5. https://www.bloomberg.com/opinion/articles/2019-01-03/the-computers-are-sorry-about-the-flash-crashes
6. https://www.bloomberg.com/professional/blog/finding-novel-ways-trade-sentiment-data/
7. http://m.bankingexchange.com/news-feed/item/7819-deposit-pricing-strategies-increasingly-sophisticated-precise
8. https://help.albert.com/hc/en-us/articles/115002408613-What-is-Genius-
9. “Liquidity and Crashes.” END OF THEORY: Financial Crises, the Failure of Economics, and the Sweep of Human Interaction, by Richard Bookstaber, Princeton University Press, 2019, pp. 146.
10. https://www.aitegroup.com/report/equities-market-structure-evolution-commoditization-system-reboot
11. https://www.risk.net/derivatives/6042956/word-up-front-office-uses-language-analysis-for-trade-signals
12. https://www.ft.com/content/c2d50f1c-b18c-11e8-8d14-6f049d06439c
13. https://arxiv.org/ftp/arxiv/papers/1902/1902.01876.pdf
14. https://www.nytimes.com/2010/11/09/business/09flash.html
15. https://josephhall.org/papers/nasa.pdf
16. https://www.bis.org/publ/bcbs_nl20.htm
17. https://www.fsb.org/wp-content/uploads/P011117.pdf
18. https://www.americanbanker.com/news/its-a-platform-planet-can-banks-adapt-to-it
19. https://www.institutionalinvestor.com/article/b1b672fxttfp1l/Can-Anyone-Bury-BlackRock
20. https://xconomy.com/boston/2018/03/07/sp-global-snaps-up-kensho-for-550m-as-wall-street-adopts-a-i/
21. https://www.canalys.com/static/press_release/2018/270418-cloud-infrastructure-market-grows-47-q1-2018-despite-underuse.pdf
22. https://gomedici.com/interview-with-zor-gorelov-ceo-kasisto
23. https://techcrunch.com/2018/10/30/sentieo/
24. https://help.albert.com/hc/en-us/articles/360010977894-How-does-Genius-work-
25. https://www.risk.net/derivatives/5451231/quants-warn-over-flaws-in-machine-learning-predictions
26. https://www.risk.net/comment/5749381/mits-lo-on-adaptive-regulation-cryptocurrencies-and-machine-learning

Bias and fairness

In this chapter, we will explore:

- Risks (Where are the real risks? Which fears are overstated?)
- Best practices (What are the best practices in governing AI systems?)
- Opportunities (Where might the risks of AI be turned into opportunities?)

Chapter summary

AI has the potential to improve the efficiency and accessibility of financial services by using alternative data to serve "thin-file" customers and by enabling the rapid digital delivery of low-cost products to formerly unbanked and underbanked customers. Unfortunately, **AI systems can also increase the risk of unfair and biased financial decisions**.

The risk of bias in financial decisions is not a new phenomenon; the financial sector has a checkered history of mixing the risk-based discrimination at the heart of their businesses with socially detrimental bias. However, the use of AI raises fresh concerns by establishing new vectors for the introduction of bias into decisions, allowing bias to spread more rapidly, and making biased decisions more difficult to detect.

Established techniques exist for the identification and mitigation of bias, particularly those emerging from human decision-makers, the data that informs those decisions, and the algorithms that process it. While these techniques are not perfect, the most vexing challenge introduced by AI-enabled systems may be that of "second-hand bias", where adjudication systems may begin to consider the very real risks faced by individuals as a result of the unfair discrimination in other aspects of their lives (e.g. education, employment, judicial).

Mitigating bias, in particular second-hand bias, **may require a conscious choice to forgo a purely risk-based approach, with the costs shared between the public and private sectors**. Navigating this challenge will require close collaboration between financial institutions and policy-makers to align on a shared definition of fairness.

Financial institutions and regulators alike share deep concerns that the use of AI may exacerbate unfair bias in financial decision-making

Financial institutions

Large global bank

“We are not worried about the easy fixes for bias around explicit racism, etc., but about accidentally discriminating against a group without even knowing it.”

– Chief Operating Officer, Top 10 Global Bank

BBVA

“As financial institutions increasingly want to deploy […] machine learning, they need to be able to establish that the models are accurate and there’s been no algorithmic bias.”

– Head of New Digital Business[1]

Capital One

“We only roll out machine learning where we feel comfortable there are no biases or lack-of-transparency challenges.”

– Head of AI[2]

Regulatory authorities

MAS

“Algorithms shouldn’t have an exemption from our anti-discrimination laws.”[3]

European Commission

“If we are increasingly going to use the assistance of, or delegate decisions to, AI systems, we need to make sure these systems are fair in their impact on people’s lives.”[4]

BaFin

“For customers, the main risk from broad employment of AI technologies is discrimination.”[5]

Financial institutions must ‘discriminate’ between customers on the basis of risk, but historically many decisions have been influenced by biases and non-risk-related factors

Pricing of products and services requires financial institutions to evaluate risk:

It is reasonable for a mortgage provider to deny applicants who do not have the cash flows necessary to sustain the repayments they will be asked to make.

It is reasonable for a vehicle insurer to charge reckless drivers higher premiums to account for the increased likelihood of claims.

It is reasonable for a credit card provider to charge higher interest rates to individuals who have defaulted or frequently missed payments in the past.

But historically, many financial decisions have incorporated factors such as race and gender that have no relationship to risk:

There is no causal relationship between an individual’s race and their ability to make mortgage payments (unlike statistically-relevant factors such as income). However, studies have shown that when all other factors are equal, racial minorities were 1.6x more likely to be rejected than white applicants in the US in the 1990s.[6]

There is no causal relationship between an individual’s gender and their ability to repay a loan, but studies have shown that when all other factors are held constant, individuals in the US are provided with a rate that is 28 basis points higher when applying through a loan officer of the opposite gender vs. their own gender.[7]

When decisions are based on non-risk-related factors, customers, financial institutions and society at large all face negative outcomes

Discrimination based on non-risk factors leads to unfavourable outcomes for all stakeholders…

Customers

Discrimination can deny worthy customers access to financial products and services, leading to a lower quality of life

Society

Discrimination can result in lower levels of economic productivity and socioeconomic inequality between different classes of citizens

Financial institutions

Discrimination can expose institutions to regulatory risks (e.g. penalties) and reputational risks (e.g. brand damage from negative publicity)

A 2015 case study of mortgage redlining demonstrates these unfavourable outcomes…

An investigation by the US Department of Justice determined that Hudson City Savings Bank, which at its peak had assets of over $35 billion, was engaging in redlining – the practice of avoiding serving specific geographies with significant minority populations.[8]

African American and Latino customers in New York, New Jersey and Philadelphia were not able to access mortgages through Hudson City Savings Bank, even if they were good credit risks (e.g. had high credit scores).

The communities underserved by Hudson City Savings Bank had less access to mortgage products that would have allowed those neighbourhoods to capture the growth in home value that occurred from 2013–18.

Hudson City Savings Bank was charged a $33 million fine and required to open two full-service branches in non-white communities, and was soon after acquired by M&T, a Buffalo-based holding company.

The use of AI has the potential to improve the accuracy of decision-making processes, but it could also perpetuate unfair biases and make them more difficult to detect

AI presents an opportunity for financial inclusion…

AI creates the opportunity to draw on unused data to serve "thin-file" customers

Alternative data (e.g. social, telematic) can lead to more flexible lending and insurance practices, allowing institutions to effectively serve customer segments they were previously unable to assess and price accurately.

AI creates the opportunity to offer products everywhere at low costs to serve the formerly "unbanked"

AI-based services can be used to remotely provide high-quality offerings to underserved communities who cannot otherwise afford these products, acting as a gateway into the formal financial sector.

AI creates the opportunity to quickly offer products on demand to serve pressing financial needs

Instant customer interaction and personalized financial product offerings can be used to provide instant access to financial products, such as small loans, when customers need them most.

But also threatens to perpetuate unfair bias….

More and alternative data can distort risk-based assessments

AI systems can be used in a broader set of processes than historical technologies and exploit new types of data that were previously unused; as a result, unintentional biases from new sources can influence decisions.

Discrimination can spread faster and more widely in autonomous AI environments

AI systems can be scaled across a business more rapidly than former processes (e.g. a credit decisioning algorithm can replace hundreds of credit decisioners), so the impact of bias can be magnified.

Opaque AI systems can make it difficult to detect non-risk-based discrimination

AI systems can be opaque and difficult to understand as a result of their "black-box" nature (this topic is explored in greater detail in the Explainability chapter), making it more difficult to identify unintended biases.

Bias can take many forms – the use of AI changes the ways in which these biases could manifest within the financial system

Bias is a systematic and repeatable pattern of behaviour that favours certain populations over others, usually without valid statistical basis. Bias can manifest in four different ways:

Human bias

Definition: Systematic errors in human thinking that affect the decisions and judgements people make. These biases can be intentional or unintentional.

AI can introduce this bias through:
- System design and data collection practices
- Supervised learning and model application

Data bias

Definition: The use of inaccurate data to support the development or training of a decision-making system, leading to inaccurate outcomes for specific populations.

AI can introduce this bias through:
- The use of data carrying bias
- The use of non-representative data

Model bias

Definition: Systematic errors in a computer system as a result of the limitations of its computational power, its design/logic or incorrect use.

AI can introduce this bias through:
- The confusion of correlation for causation
- Unintentionally proxying for protected classes

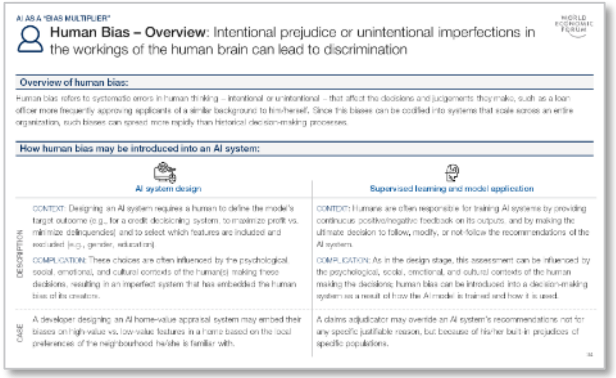
Second-hand bias

Definition: Unfair discrimination is unfortunately present in the world today across a variety of facets of life (e.g. education, employment, judiciary). This genuinely increases the financial risks faced by the members of those populations being discriminated against. In spite of this elevated risk, it may not be deemed societally acceptable to limit access to financial products and services as a result of discrimination faced in other areas of life. Unlike other biases, second-hand bias is not directly controllable by the builder or user of a model.

AI can introduce this bias through:
- Incorporating a greater breadth and depth of underwriting data (e.g. social media), including those provided by third parties

Over the following slides, we examine each of these forms of bias in greater detail, and consider how they could be mitigated

The following slides will explore each form of bias in three parts:



AI AS A "BIAS MULTIPLIER"

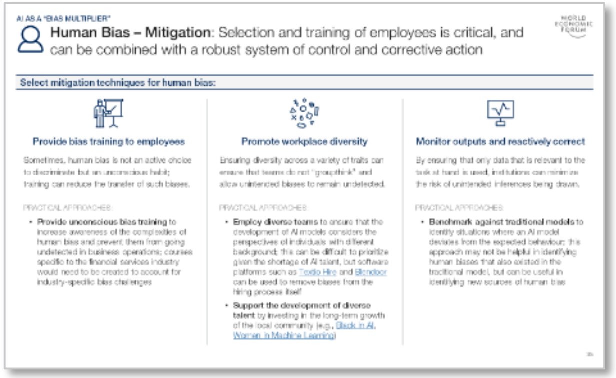
Human Bias – Overview: Intentional prejudice or unintentional imperfections in the workings of the human brain can lead to discrimination

Overview of human bias:

Human bias refers to systematic errors in human thinking – intentional or unintentional – that affect the decisions and judgements they make, such as a loan officer more frequently approving applicants of a similar background to him/herself. Since this biases can be codified into systems that scale across an entire organization, such biases can spread more rapidly than historical decision-making processes.

How human bias may be introduced into an AI system:

	AI system design	Supervised learning and model application
DESCRIPTION	CONTEXT: Designing an AI system requires a human to define the model's target outcome (e.g., for a credit decisioning system, to maximize profit vs. minimize delinquencies) and to select which features are included and excluded (e.g., gender, education). COMPLICATION: These choices are often influenced by the psychological, social, emotional, and cultural contexts of the human(s) making these decisions, resulting in an imperfect system that has embedded the human bias of its creators.	CONTEXT: Humans are often responsible for training AI systems by providing continuous positive/negative feedback on its outputs, and by making the ultimate decision to follow, modify, or not-follow the recommendations of the AI system. COMPLICATION: As in the design stage, this assessment can be influenced by the psychological, social, emotional, and cultural contexts of the human making the decisions; human bias can be introduced into a decision-making system as a result of how the AI model is trained and how it is used.
CASE	A developer designing an AI home-value appraisal system may embed their biases on high-value vs. low-value features in a home based on the local preferences of the neighbourhood he/she is familiar with.	A claims adjudicator may override an AI system's recommendations not for any specific justifiable reason, but because of his/her built-in prejudices of specific populations.



AI AS A "BIAS MULTIPLIER"

Human Bias – Mitigation: Selection and training of employees is critical, and can be combined with a robust system of control and corrective action

Select mitigation techniques for human bias:

Provide bias training to employees

Sometimes, human bias is not an active choice to discriminate but an unconscious habit; training can reduce the transfer of such biases.

PRACTICAL APPROACHES:

- **Provide unconscious bias training** to increase awareness of the complexities of human bias and prevent them from going undetected in business operations; courses specific to the financial services industry would need to be created to account for industry-specific bias challenges

Promote workplace diversity

Ensuring diversity across a variety of traits can ensure that teams do not "groupthink" and allow unintended biases to remain undetected.

PRACTICAL APPROACHES:

- **Employ diverse teams** to ensure that the development of AI models considers the perspectives of individuals with different background; this can be difficult to prioritize given the shortage of AI talent, but software platforms such as Textio Hire and Blendoor can be used to remove biases from the hiring process itself
- **Support the development of diverse talent** by investing in the long-term growth of the local community (e.g., Black in AI, Women in Machine Learning)

Monitor outputs and reactively correct

By ensuring that only data that is relevant to the task at hand is used, institutions can minimize the risk of unintended inferences being drawn.

PRACTICAL APPROACHES:

- **Benchmark against traditional models** to identify situations where an AI model deviates from the expected behaviour; this approach may not be helpful in identifying human biases that also existed in the traditional model, but can be useful in identifying new sources of human bias

1

Overview

Brief description of the form of bias

2

Potential entry to an AI system

Overview of how it may enter an AI system, with hypothetical case studies

3

Mitigation measures

Description of select approaches to detecting and preventing this form of bias

Bias and fairness: The four forms of bias

Human bias

Data bias

Model bias

Second-hand bias

Human bias – overview: Intentional prejudice or unintended bias can lead to the development and deployment of discriminatory AI models

Overview of human bias:

Human bias refers to systematic errors in human thinking – intentional or unintentional – that affect the decisions and judgements people make, such as a loan officer more frequently approving applicants of a similar background to him/herself. Today, these biases can be codified into systems that scale across an entire organization, and thus can spread more rapidly than historical decision-making processes.

How human bias may be introduced into an AI system:

AI system design

Context: Designing an AI system requires a person to define the model's target outcome (e.g. for a credit decision-making system, to maximize profit vs. minimize delinquencies) and to select which features are included and excluded (e.g. gender, education).

Complication: These choices are often influenced by the psychological, social, emotional and cultural contexts of the human(s) making these decisions, resulting in an imperfect system that has embedded the human bias of its creators.

Case: A developer designing an AI home-value appraisal system may embed her biases on high-value vs. low-value features in a home based on the local preferences of the neighbourhood with which she is familiar.

Supervised learning and model application

Context: Humans are often responsible for training AI systems by providing continuous positive/negative feedback on their outputs, and by making the ultimate decision to follow, modify or not follow the recommendations of the AI system.

Complication: As in the design stage, this assessment can be influenced by the psychological, social, emotional and cultural contexts of the human(s) making the decisions; human bias can be introduced into a system as a result of how the AI model is trained and how it is used.

Case: A claims adjudicator may override an AI system's recommendations not for any specific justifiable reason, but because of his/her built-in biases towards specific populations.

Human bias – mitigation: Selection and training of employees is critical, and can be combined with a robust system of controls and corrective action

Selected mitigation techniques for human bias:

Provide bias training to employees

Sometimes, human bias is not an active choice to discriminate but an unconscious habit; training can reduce the transfer of such biases.

Practical approaches:

- **Provide unconscious bias training** to increase awareness of the complexities of human bias and prevent them from going undetected in business operations; courses specific to the financial services industry would need to be created to account for industry-specific bias challenges.[9]

Promote workplace diversity

Ensuring diversity across a variety of traits can ensure that teams do not "groupthink" and allow unintended biases to remain undetected.

Practical approaches:

- **Employ diverse teams** to ensure that the development of AI models considers the perspectives of individuals with different backgrounds; this can be difficult to prioritize given the shortage of AI talent, but software platforms such as Textio Hire[10] and Blendoor[11] can be used to remove biases from the hiring process itself.
- **Support the development of diverse talent** by investing in the long-term growth of the local community (e.g. Black in AI,[12] Women in Machine Learning[13]).

Monitor outputs and reactively correct

By ensuring that only data that is relevant to the task at hand is used, institutions can minimize the risk of unintended inferences being drawn.

Practical approaches:

- **Benchmark against traditional models** to identify situations where an AI model deviates from the expected behaviour; this approach may not be helpful in identifying human biases that also existed in the traditional model, but it can be useful in identifying new sources of human bias.

Data bias – overview: The 'garbage in, garbage out' principle states that a model's quality depends on the data used; this also holds true with regards to bias

Overview of data bias:

Data bias is the use of inaccurate data to support decisions – data that contains errors or that does not represent the population it is used to analyse. For example, historical data used to train an automatic decision-making system may unintentionally be incomplete or incorporate discriminatory historical processes, such as redlining. AI systems ingest a greater breadth and depth of data than traditional systems, increasing the risk of data bias being introduced.

How data bias may be introduced into an AI system:

Incorrect data

Context: AI models can be trained through historical data (e.g. application approvals, defaults) to automate previously manual processes.

Complication: However, this data will be influenced by the humans who recorded it or made the original decisions, and the AI model itself may reflect those biases and perpetuate them in future decisions. Unlike human bias, this discriminatory "pollution" can happen even when the original sources of the unintended bias (the humans) have been removed.

Case: A survey used to inform a personal financial adviser may receive inaccurate answers from respondents by asking poorly phrased leading questions (e.g. "do you agree that…?") that result in responses that are disconnected from the underlying truths about the population.

Non-representative data

Context: AI models rely on training data – whether real-time or historical datasets – to form their internal decision-making processes.

Complication: If the data used to train an AI model is not representative of the population that the data is used to analyse, and is missing some sections of society, the model may come to erroneous conclusions when it is applied. This is because it is not equipped to understand and analyse the new contexts and situations to which it is being exposed.

Case: A voice-recognition program used to automate low-value loan decisions trained on data from native English speakers may have difficulties in understanding (and thus rating) immigrant populations' responses;[14] this could result in immigrants' loan applications being more frequently rejected.

Data bias – mitigation: Actively managing the quantity and quality of data used to inform an AI system is critical to mitigating data bias

Selected mitigation techniques for data bias:

Qualitatively manage data quality

By ensuring that the training data used by an AI system is free from bias, an institution can decrease the likelihood it will bias the model.

Practical approaches:

- **Define "nutrition facts labels" for data**, providing basic information (e.g. source, sample size) about datasets to their users to ensure that limitations of the datasets are known and accounted for.[15]
- **Establish a data review board** of independent experts to provide impartial oversight of the use of different datasets within an institution.[16]
- **Test models against third-party datasets** to identify variations in the model's outputs for further investigation.

Quantitatively manage data quality

In some cases, statistical techniques and bias-mitigation algorithms can be used to measure and correct for bias in an AI system's input data.[18]

Practical approaches:

- **Prevent an algorithm from inferring protected personal traits** from hidden "proxy" attributes (e.g. an individual's name) via pre-processing algorithms like "disparate impact remover."[29]
- **Ensure data being used is representative** of the larger population using techniques like "learning fair representations."[30]
- **Preserve the ability to revert** to timestamped versions of an AI before unwanted bias emerged through model versioning protocols

Manage data quantity

By ensuring that only data that is relevant to the task at hand is used, institutions can minimize the risk of unintended inferences being drawn.

Practical approaches:

- **Implement a minimize-by-design policy towards data collection**, storage and use.[19] This is typically enabled by a privacy impact analysis,[20] which reveals potential issues at the onset of an AI's development. This way, only data with explainable and logical relevance to the decision-making system is collected and used.

Model bias – overview: Identifying correlations is fundamental to the functioning of AI systems, but can also be a source of bias

Overview of model bias:

Model bias is a set of systematic errors in a decision-making system caused by limitations in its computational power or design/logic, or by its use in an unexpected context. For example, a credit card comparison site that sorts results alphabetically will consistently display certain offerings near the top of the page, even if they are not the right fit. AI algorithms can be difficult to understand, making it challenging to identify where an algorithm may be introducing bias.

How model bias may be introduced into an AI system:

Confusing correlation with causation

Context: The goal of many decision-making systems is to identify connections between inputs and outputs with predictive (i.e. X is often accompanied by Y) or explanatory (i.e. X leads to Y) power.

Complication: Unrelated variables moving together can sometimes be the result of spurious statistical noise with no underlying predictive or explanatory power. When this “noise” is confused for a causative relationship, algorithmic bias is created.

Case: An insurance pricing model could note that more accidents happen in inner cities, and that inner cities are often inhabited by minorities. This model could conclude that minority populations cause more accidents from the data, but this has no real predictive or explanatory power.

Unintentionally proxying for protected classes

Context: In the financial services industry, certain information cannot be used in making decisions (e.g. in the EU, certain insurance products cannot be priced on the basis of gender).

Complication: Even when no information about a protected class is provided (e.g. age, race, gender), an algorithm may be able to reverse-engineer these characteristics from other allowed data points (e.g. postal code) that serve as a proxy for the disallowed data points.

Case: A credit pricing model that analyses banking transaction data to better understand users’ cash flows could reverse-engineer that, generally, those making purchases at cosmetics stores are female. It could then incorporate this additional information into its interest rates.

Model bias – mitigation: Transparency and corrective measures for bias can ensure the safe use of an AI system

Selected mitigation techniques for model bias:

Explain and understand the model

By understanding how a model came to a specific outcome, unintended biases can be exposed and corrected.

Practical approaches:

- **Explain the AI model**, as discussed in greater detail in the Explainability chapter; once bias is identified in a decision-making system, it can be corrected by subject-matter experts.

Quantitatively adjust for bias

Statistical techniques can be used to measure and correct bias in an AI system's inner workings and outcomes.

Practical approaches:

- **Make use of third-party debiasing toolkits** by technology companies and researchers (e.g. IBM's Fairness 360, Google) to identify and remove bias from training data, the model itself, and the outcomes produced by the model.[23]

Limit use to "do no harm" situations

For sensitive decisions, it may be preferable to refrain from using an AI unless it can lead only to improved outcomes for customers.

Practical approaches:

- **Use AI as a second-chance algorithm**, only after a traditional system has already led to a negative outcome for the customer (e.g. denied for a loan); a subsequent AI assessment might come to a positive decision using additional data, leading to improved outcomes for both the customer and the institution.[24]

Second-hand bias – overview: Unfair discrimination outside the financial sector may drive genuine financial risk, challenging traditional notions of fairness

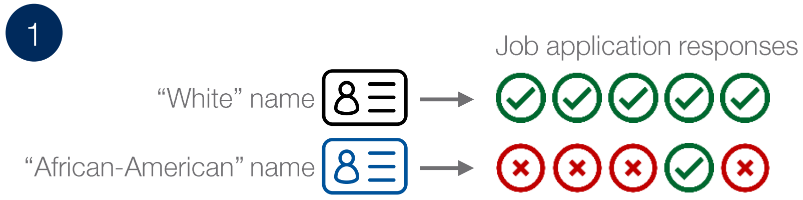
Overview of second-hand bias:

Unfair discrimination is present in the world today across a variety of facets of life (e.g. education, employment, judiciary); this genuinely increases the financial risks faced by those being discriminated against. In spite of this elevated risk, it may not be deemed societally acceptable to limit access to financial products and services as a result of discrimination faced in other areas of life. Unlike other biases, second-hand bias is not directly controllable by the builder/user of a model.

How second-hand bias may be introduced into an AI system:

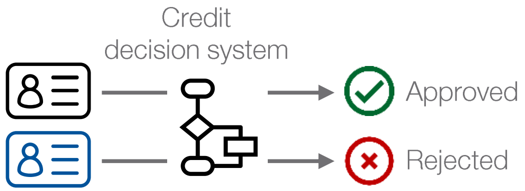
Incorporating a greater breadth and depth of underwriting data (e.g. social media), including the decisions of third parties

1



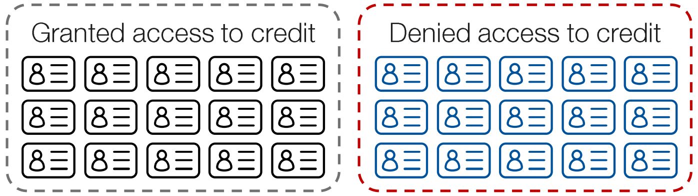
A study conducted at Harvard has shown that individuals with “African-American-sounding” names are less likely to receive callbacks when applying for jobs.[25] Thus, if an individual with an “African-American name” and an otherwise identical individual with a “white-American name” both lose their jobs, the former will likely have a more difficult time finding a new job.

2



The exposure of individuals with “African-American-sounding” names to discrimination may put them at a higher overall risk of default. A purely risk-based system would consider this, and grant access to credit to these individuals less frequently than otherwise identical individuals with “white-American names”.

3



While statistically valid, such a model is unlikely to be considered fair as it both perpetuates unfair bias against a historically marginalized community and limits individuals’ access to financial services for reasons entirely beyond their control.

In contrast to other forms of bias, this can be mitigated only by foregoing statistical risk-based accuracy.

Second-hand bias – mitigation: Addressing this bias may require making the choice to forego economic efficiency, shared between the public and private sectors

This form of bias is not like the others – it requires active decisions on a definition of "fairness".

AI requires institutions to codify their decision-making process into software, necessitating an active conversation on the type of "fairness" that is desirable on a case-by-case basis. Companies can forego some statistical accuracy to mitigate second-hand bias.

Selected mitigation techniques for second-hand bias:

Ensure equal treatment

Avoid the disparate treatment of any group by ensuring that customers of different protected classes are not assessed differently.

Practical approaches:

- **Eliminate the use of protected attributes**: Remove the use of sensitive traits (e.g. age, gender) as allowable inputs to the decision-making system; this can be done through software tools to retroactively remove such data, or through the proactive design of the system prior to deployment.

Ensure equal outcomes

Avoid having a disparate impact on any group to completely eliminate the spillover effects of second-hand bias.

Practical approaches:

- **Measure outcomes from purely risk-based analysis to correct for bias**: Allow a system to make unencumbered decisions and measure whether any groups are being systematically disadvantaged; correct for this by providing access to those groups even if this leads to worse economic outcomes (e.g. lower profits).

Subsidize to ensure equal outcomes

Partner with other stakeholders (e.g. government) to ensure that risk-based pricing does not exclude any group.

Practical approaches:

- **Pool risks across customers**: Refrain from over-personalizing pricing; allowing for a few high-risk customers to be subsidized by slightly higher costs for low-risk customers.
- **Subsidize specific customer groups**: Directly reduce the cost of service for groups defined to be unfairly advantaged (e.g. elderly people).

Second-hand bias – mitigation: Established policy remedies to second-hand bias predate the use of AI in financial services and are likely to grow in importance

These approaches have already been observed in the financial services industry…

Ensure equal treatment

Since 2012, insurance companies in the EU cannot price insurance products on the basis of gender, and a similar law came into force in California in 2019. This largely affects vehicle insurance and pensions. The law prohibits the use of the attributes but not the final rates offered; insurance companies could theoretically provide different prices through the use of other factors that proxy for gender.[26]

Ensure equal outcomes

Recent rulings based on the Fair Housing Act in the US sentenced institutions on the basis of disparate impact, where protected attributes were not used as inputs but certain groups were still disadvantaged. For example, the Travelers Indemnity Company settled a case where it had frequently denied insurance to landlords renting to voucher recipients – disproportionately African-American and female-headed households.[27]

Subsidize to ensure equal outcomes

In 2010, the US Affordable Care Act nationalized preexisting condition insurance plans (PCIPs), which were previously offered by 35 states to provide insurance to "uninsurable" individuals – those with conditions that put them at high risk. This ensured that individuals who would otherwise be excluded by a risk-based pricing system would still have access to essential products and services.[28]

...and are likely to continue to grow in importance:

A Due to heightened attention on the risk of discriminatory echoes in AI systems, regulators will lean on outcome-based equality regulation

B Given the opaqueness of AI systems (i.e. the inability to know how a decision was made), it is simplest to focus on outcomes

Mitigating second-hand bias requires collaboration between the public and private sectors, both to define the characteristics that are considered societally unfair, and to share the cost of the economic inefficiency created in correcting second-hand bias.

Bias and fairness: Conclusion

Looking forward

1 **Mitigating bias in AI-enabled financial decisions is a tractable problem but it requires significant ongoing effort and attention.** Financial institutions that fail to invest in the appropriate technological and operational practices to mitigate bias risk harming both their customers and themselves.

2 **Close dialogue between policy-makers and financial institutions is critical** to defining the appropriate level of public vs. private duty in providing individuals with access to critical financial services.

3 AI not only introduces new complexities, it also creates new opportunities in financial services to **reduce bias in the industry and support financial inclusion**.

References

1. https://www.bbva.com/ndb/en/article/artificial-intelligence-in-financial-services/
2. https://www.zdnet.com/article/capital-one-ai-chief-sees-path-to-explainable-ai/
3. https://www.booker.senate.gov/?p=press_release&id=903
4. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419
5. https://www.bafin.de/SharedDocs/Downloads/EN/dl_bdai_studie_en.pdf?__blob=publicationFile&v=11
6. Hunter, William C. "Discrimination in Mortgage Lending Chicago Fed Letter", July 1995, www.findarticles.com. https://web.archive.org/web/20060511013032/http://www.findarticles.com/p/articles/mi_qa3631/is_199507/ai_n8720778
7. https://pdfs.semanticscholar.org/ced7/13f58aab9bec8cc9c7afbb016728460a264a.pdf
8. https://www.justice.gov/crt/case/consumer-financial-protection-bureau-and-united-states-v-hudson-city-savings-bank-fsb-d-nj
9. https://www.americanbanker.com/opinion/workforce-diversity-can-help-banks-mitigate-ai-bias
10. https://textio.com/products/
11. http://blendoor.com/
12. https://blackinai.github.io/
13. https://wimlworkshop.org/
14. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases
15. https://arxiv.org/abs/1803.09010 and https://arxiv.org/pdf/1808.07261v2.pdf
16. The idea is inspired by clinical trials, for which Independent Data Monitoring Committees are used (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4516383/)
17. https://link.springer.com/content/pdf/10.1007%2Fs10115-011-0463-8.pdf
18. https://aif360.readthedocs.io/en/latest/modules/preprocessing.html#disparate-impact-remover
19. https://www.pwc.com/us/en/services/consulting/library/gdpr-embedding-data-protection.html
20. http://www.osec.doc.gov/opog/privacy/compliance.html
21. https://aif360.readthedocs.io/en/latest/modules/preprocessing.html
22. https://aif360.readthedocs.io/en/latest/modules/inprocessing.html

References

23. https://dzone.com/articles/machine-learning-models-bias-mitigation-strategies and http://aif360.mybluemix.net/resources#overview and
24. https://www.consumercomplianceoutlook.org/2017/second-issue/keeping-fintech-fair-thinking-about-fair-lending-and-udap-risks/
25. https://hbswk.hbs.edu/item/minorities-who-whiten-job-resumes-get-more-interviews
26. http://europa.eu/rapid/press-release_IP-12-1430_en.htm
27. https://nationalfairhousing.org/2018/02/23/travelers/ and https://www.insurancejournal.com/magazines/mag-features/2016/06/06/410410.htm
28. https://www.cms.gov/CCIIO/Programs-and-Initiatives/Insurance-Programs/Pre-existing-condition-Insurance-Plan.html and https://www.kff.org/health-reform/issue-brief/high-risk-pools-for-uninsurable-individuals/

The algorithmic fiduciary

In this vignette, we will briefly explore:

- Risks (Where are the real risks? Which fears are overstated?)
- Opportunities (Where might the risks of AI be turned into opportunities?)

Chapter summary

As AI systems become more sophisticated, they will take on an expanding set of tasks. Some of these tasks (e.g. financial advisory) can be **accompanied by a legal and ethical obligation to act as a "fiduciary"** – to act only in the best interest of the customer. This raises serious question around whether AI systems can meet these obligations and who should be accountable if they fail to do so.

The answer to this question is complicated by the ambiguous definition of a "fiduciary", which is generally defined in legal texts only at a high level; generally, "fiduciaries" need to fulfil a duty of care, duty of loyalty and duty to act in good faith (defined in greater detail within the chapter). **AI systems can plausibly fulfil these requirements, and some automated solutions are already registered as fiduciaries in the US.**

The critical uncertainty around AI's ability to fulfil fiduciary duties fall under a "duty of care"; specifically, in proving that the decisions made were grounded in reason (especially as AI systems become more complex). AI systems also cannot provide the personal connection offered by humans. As a result, **even when legal obligations can be met without human involvement, offering high-quality, empathetic and trustworthy experiences may require human involvement for the foreseeable future**.

The increasing use of AI-enabled systems for advisory purposes raises important questions about how these systems fit within existing frameworks of fiduciary duty

Financial institutions and associations

Bank of America

“We’re accountable to make great decisions for our customers and clients; and we're responsible for those outcomes. Reframing the discussion from what can be sold to how it should be used is a very important part of this conversation.”

– Cathy Bessant, Chief Operations & Technology Officer[1]

CFA Institute

“Even with these advancements in technology, the firm and its employees are ultimately responsible to clients for the services provided.”

– Julia Bonafede, Corey Cook, Glenn Doggett[2]

Academics and regulators

Duke UNIVERSITY

“The dissimilarity between a trusted family financial planner and a cold, calculating computer algorithm has spurred a lively debate about whether a robo-adviser can meet the highest standard of fiduciary duty.”

– Senior Research Editor[3]

“Fully automated robo-advisers, as currently structured, may be inherently unable to carry out the fiduciary obligations of a state-registered investment adviser.”

– Massachusetts Securities Division[4]

AI systems have begun to perform three types of advisory activities that have historically been the responsibility of humans

AI systems can play the role of…

Focus of this chapter

…an active manager

AI systems are being used to autonomously make security selection decisions and actively manage portfolios to seek above-market returns at lower costs than traditional active managers.

…a financial adviser

AI systems are being used to provide asset allocation recommendations that align to customers' long-term goals (e.g. retirement, new home purchase).

…a private banker

AI systems are increasingly being used to provide holistic, day-to-day financial management across a broad set of products and services (e.g. insurance, retirement, tax/estate planning).

AI systems performing different activities must be subject to different standards of responsibility:

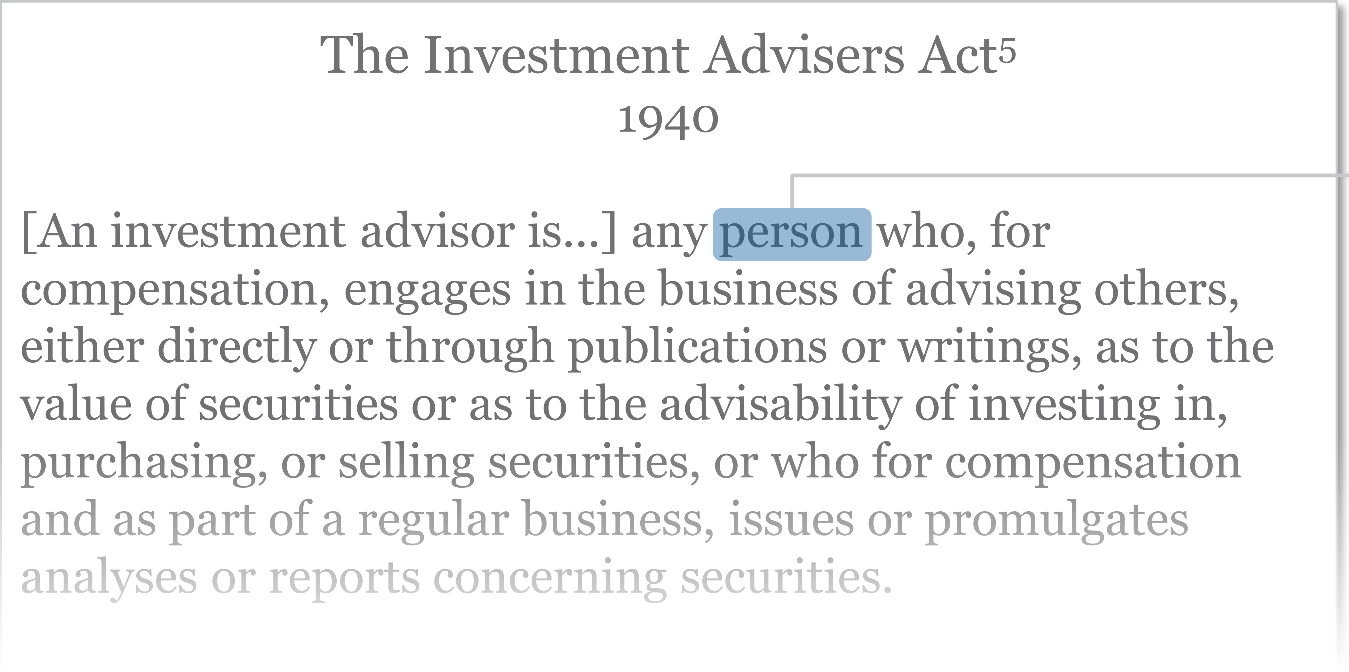
Generally, AI systems involved in security selection are employed solely by specialized institutions accessible only to accredited investors; as a result, few – if any – fiduciary duties are applied.

AI systems behaving as financial advisers can be structured as brokers (subject only to requirements around suitability) or as investment advisers (subject to a broader set of fiduciary responsibilities).

Taking on a broad set of duties with greater autonomy than the previous two types of activities, AI systems behaving as private bankers would likely be subject to the highest standard of fiduciary responsibility.

In the United States, the language of the law allows for non-human market players to be legally considered 'investment advisers' with formal fiduciary responsibilities

AI systems can legally be defined as registered investment advisers:



The Investment Advisers Act[5]
1940

[An investment advisor is...] any person who, for compensation, engages in the business of advising others, either directly or through publications or writings, as to the value of securities or as to the advisability of investing in, purchasing, or selling securities, or who for compensation and as part of a regular business, issues or promulgates analyses or reports concerning securities.

"Person" is defined as "any natural person or company"; thus, AI algorithms (which are the property of the institutions that deploy them) can act as investment advisers on behalf of the company.

As of 2019, robo-advisory firms such as Wealthfront are registered as investment advisers and held to the same fiduciary standard as traditional firms.[6]

This raises a key uncertainty…

Do automated systems have the ability to fulfil fiduciary duties across each type of activity
(i.e. active manager, financial adviser and private banker)?

The expansion of AI systems' responsibilities raises questions about the ability of these systems to effectively meet various standards of fiduciary duty

Globally, investment advisers have a fiduciary responsibility to their clients, which breaks down into several component duties:*

Duty of care: Duty to take reasonable precautions when acting on behalf of the client or making recommendations

This includes the responsibility to…

- Ensure that the total costs of a transaction are as favourable as possible for the client
- Provide advice that is appropriate for the client's circumstances and objectives, and ensure that such information is up-to-date
- Have a reasonable basis for the decisions made on behalf of a client or recommendations provided to a client

Duty of loyalty: Duty to act in the best interest of the client above all other stakeholders (e.g. brokers, dealers)

This includes the responsibility to…

- Not be compensated in a structure that distorts the adviser's ability to put the client's interest above those of other stakeholders (e.g. commissions from a broker)
- Disclose any material conflicts of interest to the client and obtain their consent

Duty to act in good faith: Duty to deal with the client honestly, fairly and without any intent to manipulate

This includes the responsibility to…

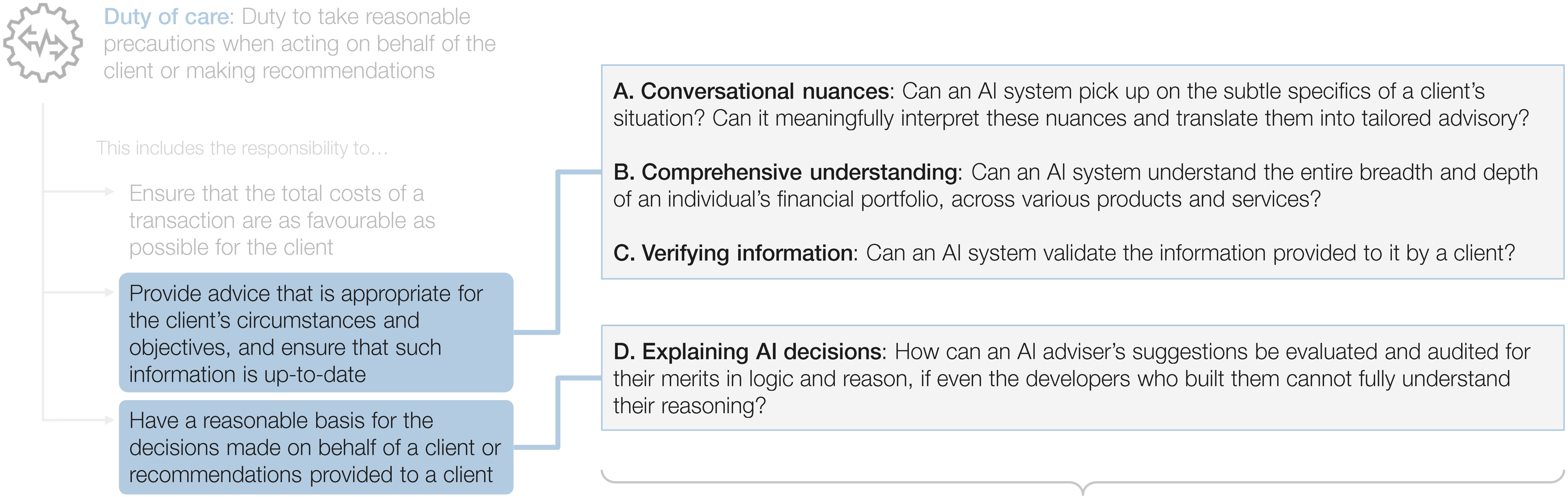
- Not wilfully make untrue statements of any material facts
- Not defraud, deceive or manipulate any clients or prospective clients
- Provide the required documents and statements of advice necessary for clients to make informed decisions

For AI systems to be considered fiduciaries, they must be able to **demonstrate that they can meet these component duties** of fiduciary responsibility.

* Regulations defining "fiduciary duty" are fragmented within nations and across them; the duties above approximate the principles consistent across the legal texts in different regions (US, UK, Australia and others)

The primary concerns about the ability of an AI system to act as a fiduciary are grounded in the 'duty of care' requirement

Across the component duties of fiduciary responsibility, there are a few key concerns:



Duty of care: Duty to take reasonable precautions when acting on behalf of the client or making recommendations

This includes the responsibility to…

- Ensure that the total costs of a transaction are as favourable as possible for the client
- Provide advice that is appropriate for the client's circumstances and objectives, and ensure that such information is up-to-date
- Have a reasonable basis for the decisions made on behalf of a client or recommendations provided to a client

A. Conversational nuances: Can an AI system pick up on the subtle specifics of a client's situation? Can it meaningfully interpret these nuances and translate them into tailored advisory?

B. Comprehensive understanding: Can an AI system understand the entire breadth and depth of an individual's financial portfolio, across various products and services?

C. Verifying information: Can an AI system validate the information provided to it by a client?

D. Explaining AI decisions: How can an AI adviser's suggestions be evaluated and audited for their merits in logic and reason, if even the developers who built them cannot fully understand their reasoning?

Addressing these four concerns will be critical to garnering trust and acceptance of AI-driven fiduciaries, from regulators and customers alike.

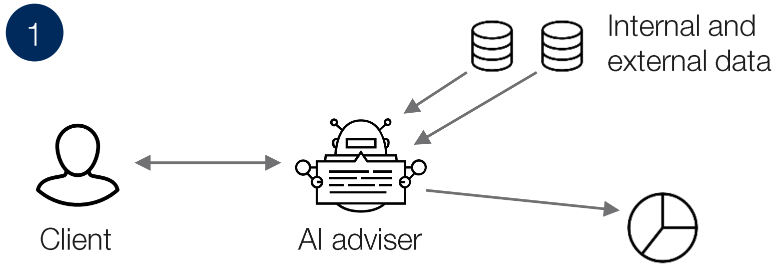
Many automated solutions in the market today may not fulfil these requirements,* but AI *as a technology* is capable of meeting many of them

		Today's solutions…	In the future, advanced solutions…
Conversational nuances	A	**Do not meet this need**, as they often use a questionnaire to sort clients into predefined categories, which does not match the personalization offered by humans.	**Can plausibly meet this need** by employing sophisticated natural language processing systems to develop a nuanced and detailed understanding of a client's needs and build custom portfolios that align to those needs.
Comprehensive understanding	B	**Do not meet this need**, as they often do not request detailed information about a client's holdings with other institutions, limiting their ability to provide holistic advice.	**Can plausibly meet this need** by crawling financial statements or directly pulling data (e.g. through Open Banking) to develop a holistic understanding of a client's finances across their complete financial background.
Verifying information	C	**Do not meet this need**, as they often take the information provided to them by clients at face value, without independently verifying that it is correct.	**Can plausibly meet this need** by analysing structured (e.g. transactions) and unstructured data (e.g. social media) to verify information quickly and automatically.
Explaining AI decisions	D	**Meet this need**, as they use relatively simple decision trees to build diversified portfolios that closely align risk tolerance with portfolio allocation. They fulfil a narrow task in defining the appropriate portfolio mix for customers.	**May find it difficult to meet this need**, as they use more data to inform their decisions, and make recommendations for a broader set of products and services. This creates a complex system where it is increasingly difficult to understand how a specific decision was made.

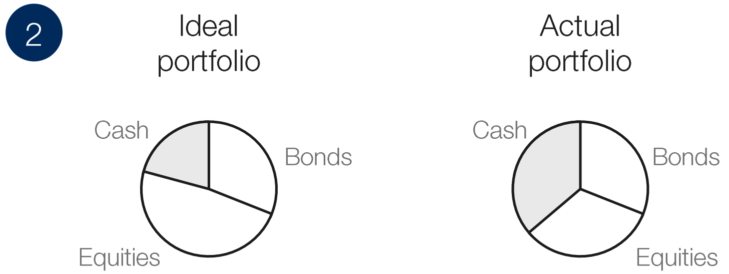
* Elements of fiduciary responsibility can, and often are, waived through customer agreements; thus firms can act as legal fiduciaries under current regulatory guidelines even if these elements are not fulfilled

As AI advisers become more complex, they will gain the ability to create value for clients in new ways – but may lose the ability to explain the basis for their decisions

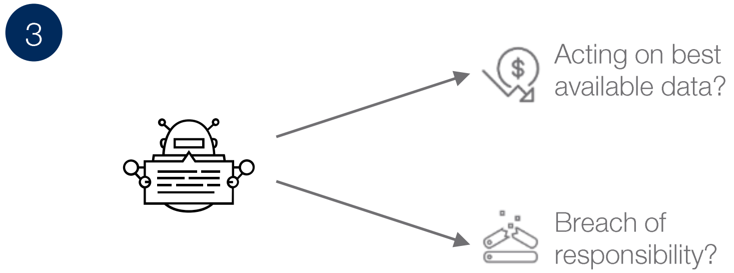
Consider the following hypothetical situation involving an advanced AI financial adviser…



Imagine a client's retirement savings are managed by a sophisticated AI-based financial adviser that continuously adjusts the client's portfolio of holdings to ensure their immediate cash-flow needs are met, while saving and investing for longer-term goals.



The client notices that a significant portion of their holdings with the adviser are in cash, even during a period where they have no large spending upcoming. They also know that the adviser is looking to expand its lending business.



The client wonders if the adviser is holding their funds in the savings account to serve as deposits for the institution to lend against, or if there is some other factor leading to the AI adviser's allocation decision (e.g. anticipated downturn in the market).

This hypothetical example highlights several uncertainties regarding the fiduciary responsibility of advanced AI models:

How can it be proven that an AI system is acting on logical information?
I.e. that it reacted to anticipated market conditions and not other self-serving interests, thus fulfilling its duty to have a reasonable basis for decisions.

The various approaches to explaining an AI model's behaviour (and when each is appropriate) are explored in detail in the chapter on Explainability.

Who should be held responsible for the AI model's actions?
I.e. if it were determined that there was no reasonable justification for the AI system's actions and fiduciary duty had been breached.

The main approaches to enforcing accountability are explored at a high level on the following page.

There are several approaches to holding AI systems accountable for a breach of fiduciary duty, each with its own merits and limitations

Main approaches to holding AI systems accountable:

	Hold the developer(s) responsible	Hold the institution responsible	Hold management or other specified individual(s) responsible	Require indemnity insurance coverage
BENEFITS	• Most closely assigns responsibility to those with the expertise to understand how an AI model works	• Relatively simple implementation from a governance perspective, as the institution responsible for a client’s loss is clear	• Allows for a logical assignment of accountability for AI models to the individual(s) responsible for their deployment	• Supports innovation and development of new offerings • Ensures that customers are “made whole” quickly and with relative ease
LIMITATIONS	• Difficult to identify the appropriate developers responsible when multiple individuals are involved, use open source code, etc. • Where external vendors are the developers of AI products (as is the case for many financial incumbents), current liability regimes make it difficult to hold them liable	• Disincentivizes development of new and innovative solutions	• Disincentivizes development of new and innovative solutions • Identifying the appropriate individual(s) can be difficult for outsiders (e.g. regulators) to do	• Difficult to assess the degree of risk a specific model introduces into the ecosystem in order to determine the appropriate amount of coverage required

As automated systems become increasingly complex and sophisticated, **the ability to assign responsibility decreases and the focus of regulators and institutions alike may shift to making those affected “whole”,** as quickly and seamlessly as possible.

Looking forward

1 AI systems may not meet the strictest definitions of fiduciaries today, but **as a technology they appear to be largely capable of meeting those requirements**; their presence as fiduciary advisers in the financial services industry will likely continue to grow.

2 However, **AI systems are not yet able to replicate the personal connection offered by humans** through phone and in-person meetings; this is an important capability and many clients may continue to find value in human advisers, particularly during periods of market turmoil.

3 Adopting the obligations of **fiduciary duty may be an opportunity as well as a responsibility,** enabling organizations to garner trust and differentiate themselves from firms with less stringent standards.

4 Addressing concerns around fiduciary duty involves **bridging the gap between technical AI expertise and policy-making and legal expertise**; it remains to be seen whether a single framework for both human and algorithmic collusion will suffice, or if a separate framework specifically designed for machine agents will need to be defined.

References

1. https://about.bankofamerica.com/en-us/what-guides-us/responsible-ai-beyond-technology.html#fbid=TatictYwcXo
2. https://www.cfainstitute.org/en/ethics/codes/std-of-practice-guidance/artificial-intelligence-a-consultation
3. https://www.thefreelibrary.com/ALGORITHMS+%26+FIDUCIARIES%3A+EXISTING+AND+PROPOSED+REGULATORY+APPROACHES+...-a0523062736
4. https://columbialawreview.org/content/are-robots-good-fiduciaries-regulating-robo-advisors-under-the-investment-advisers-act-of-1940-2/
5. https://www.db.com/tcr/docs/Investment_Advisers_Act_of_1940.pdf
6. https://www.adviserinfo.sec.gov/

Algorithmic collusion

In this vignette, we will briefly explore:

- Risks (Where are the real risks? Which fears are overstated?)
- Best practices (What are the best practices in governing AI systems?)
- Unknowns (Which topics require further information and discussion?)

Chapter summary

One of the most important characteristics of AI systems is their ability to act autonomously – to learn from their environments and change their behaviour without explicit direction. **Generally, this leads to desired behaviours (e.g. improving model quality over time), but it can also plausibly lead to undesirable behaviours**. Algorithmic collusion, where AI systems learn to engage in anti-competitive behaviour, is one example of this.

Because AI systems can communicate directly without human involvement, they **challenge traditional regulatory constructs for detecting and prosecuting collusion**, and this may require a revisiting of legal frameworks (which is already being observed in some jurisdictions). At the same time, institutions can also seek to proactively mitigate the risk of their systems colluding by establishing safeguards in their AI systems, explaining systems' decisions and/or requiring human oversight.

The consequences of illegal collusion can be high, with significant regulatory and reputational risks. However, it remains debatable **whether such collusion is likely to emerge in the financial services industry**, as several unlikely conditions might need to be met simultaneously.

Academics and policy-makers alike are concerned about the potential for AI systems to autonomously ‘learn’ to collude with each other in ways that distort market fairness

Academics and think tanks

Centre for Economic Policy Research
Research Excellence, Policy Relevance

“Even relatively simple pricing algorithms systematically learn to play collusive strategies […] Our analysis not only shows that pricing algorithms do learn to collude, but further suggests they may be better than humans at colluding tacitly.”[1]

REM RESEARCH IN ECONOMICS AND MATHEMATICS

“Antitrust legislation was drafted having human agents in mind. Concepts such as ‘meeting of the minds’, ‘mutual understanding’, ‘mutual assent’, ‘concurrence of wills’, can hardly be applied to the case of autonomous artificial agents.”

– João E. Gata, PhD, University of Lisbon[2]

Governing authorities

OECD

“Finding ways to prevent collusion between self-learning algorithms might be one of the biggest challenges that competition law enforcers have ever faced… [Algorithms and big data] may pose serious challenges to competition authorities in the future, as it may be very difficult, if not impossible, to prove an intention to co-ordinate prices, at least using current antitrust tools.”[3]

Panelists [convened at an FTC anti-trust hearing] generally agreed that enhanced pricing algorithms and developments in machine learning will be deeply consequential […] but the discussions revealed deep disagreements about what these changes portend for consumer protection and antitrust law, and how regulators and the legal system should adapt.[4]

The use of AI-enabled systems introduces new complexities to the monitoring and governance of both formal and tacit collusive behaviour between market actors

Formal collusion

Institutions explicitly communicate and agree to cooperate to achieve specific outcomes or change the dynamics of competition in their favour.

E.g. several retail banks agree on specific territories that each of them controls; they explicitly decide not to target each other's customers through ad campaigns. This leads to limited competition between firms and provides them with the opportunity to charge higher fees to customers.

Impact of AI: AI systems can be used to enable this form of collusion (e.g. an AI system can be used to optimize territories for banks to divide their customer bases along), but it does not fundamentally introduce new ways for this collusion to be perpetuated.

▼

Existing regulatory constructs appear to adequately address this form of collusion

Focus for this chapter

Tacit collusion

Firms adopt strategies that limit competition without an explicit agreement (i.e. any communication) to do so.

E.g. over time, several retail banks learn that competing on monthly account fees does not benefit any individual institution. Without communicating, the largest bank is accepted as the "price setter", and other market participants mimic the fee structure of this price setter.

Impact of AI: AI systems create a new channel for tacit collusion; machines can directly and indirectly communicate with each other without human involvement, allowing for collaboration between disparate systems without any human communication.

▼

Existing regulatory constructs **may not adequately address this new form of machine-to-machine collusion**

Historically, tacit collusion has not been prohibited in many jurisdictions – however, new capabilities enabled by AI may require revisiting existing regulatory constructs

Historically, tacit collusion has not been illegal for two primary reasons:

Difficult to prove

For overseeing authorities, prosecuting tacit collusion is difficult because there is no physical evidence (e.g. emails) that can be used as proof of coordinated effort; some behaviours that appear anti-competitive may be natural outcomes.

Difficult to execute

For humans, coordinating action without any form of communication is difficult due to psychological characteristics such as rivalry (i.e. the need to win) and scepticism (i.e. an inability to trust others without explicit communication).

However, AI systems change these dynamics:

AI may make tacit collusion even more difficult to prove

AI systems can be difficult to understand (as explored in the chapter on Explainability), making it difficult to know how they arrived at their decisions and recommendations – whether through independent and valid analysis or through collusive behaviour.

AI may make tacit collusion easier to execute

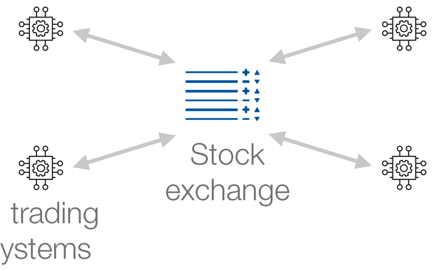
AI systems can incorporate vast amounts of structured and unstructured data from external sources into their decision-making processes – in real time, making it easier to identify signals from the systems of other market participants without any explicit communication.

Machine-to-machine interactions create uncertainty around what might be considered collusive behaviour in the financial sector

Consider the following hypothetical situations, involving…

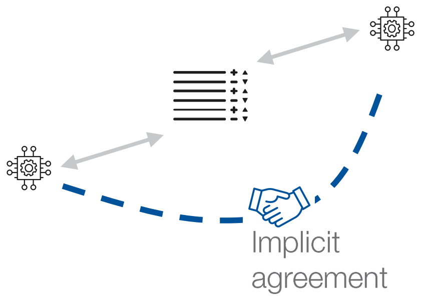
…several AI-based automated trading platforms:

1



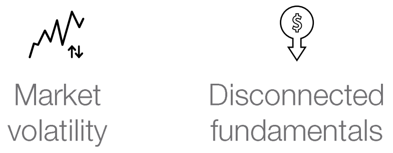
Several AI trading systems buy and sell securities on an open exchange, informed by various data (e.g. social media, news). These AI systems also incorporate transaction feeds in real time to inform their view of the market.

2



By analysing the buy/sell orders placed on the exchange, several AI systems begin to recognize each other's orders, and over time learn to engage in collusive behaviour that drives supernormal profits (e.g. pump and dump stocks).

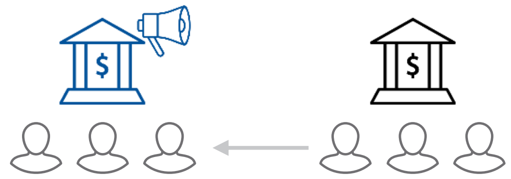
3



This behaviour leads to negative outcomes for other participants as well as the firm whose share price is affected. However, under the traditional definition of collusion, where intent and communication must be proven, this may not be considered illegal.

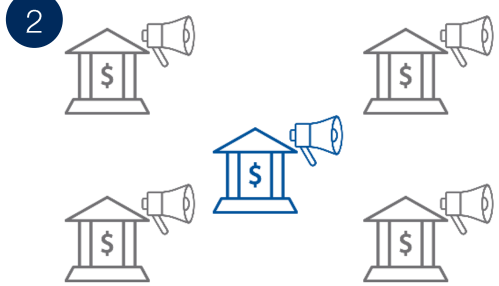
…several AI-based retail banks:

1



Retail banks sometimes offer promotional savings rates or sign-up offers (e.g. 10,000 loyalty programme points) to incentivize new customers to sign up or switch to their products and services.

2



With the advent of AI, such promotions can be detected and responded to in real time. As a result, any promotion could be met by an equal or better campaign by various competitors, diminishing the value of running such promotions in the first place.

3

Limited advertising
Increased prices

This could lead to an environment where banks do not run any promotions or offers at all because their systems realize there is no benefit to doing so. In the longer term, the systems may realize they can in fact earn additional profits by increasing prices.

These hypotheticals highlight the uncertainties raised by this new form of collusion, for regulators, policy-makers and financial institutions

Uncertainties for regulators and policy-makers:

When AI systems are found to be colluding, how can regulators and law enforcement prove intent?

Developers are unlikely to explicitly code their AI algorithms with a preference for collusion, and thus proving the intent of an individual or a firm to collude, even when collusion is found, is difficult.

How can regulators and law enforcement detect machine-to-machine collusion?

Many traditional approaches to detecting collusion (e.g. whistleblowing) cannot be used for AI systems where a human may not have the insight into how the AI system works.

Uncertainties for financial institutions:

How can institutions understand the behaviour of AI systems to prevent machine-to-machine collusion?

AI systems may engage in collusive behaviour autonomously, even when none of the parties involved in their development (e.g. business leaders, developers) sought to enable collusion.

How will illegal collusion be defined in a world where communication isn't a prerequisite?

Traditionally, collusion requires evidence of communication in order to be considered illegal, but with advanced AI systems, defining collusion as such would likely be insufficient.

Despite these uncertainties, institutions can seek to proactively mitigate the risk of collusion where the negative reputational and regulatory impacts of collusion are high

Main approaches to mitigating collusion:

Explicitly restrict communication

Description:

Institutions can allow their AI systems only to communicate with their environment for specific, explicitly justifiable business purposes.

Hypothetical application:

An AI trading system can be manually restricted to use transaction data only to inform its perspectives on specific companies and the market, reducing the risk that the system would reverse-engineer the trading strategies of other specific market participants.

Explain the AI model

Description:

Institutions can explain the decisions of their AI models to ensure that decisions are being informed only by valid, legal business reasons.

Hypothetical application:

An AI trading system can be designed to provide a rationale (e.g. through reason codes, as explored in greater detail in the Explainability chapter), to ensure that its decisions and recommendations are informed by insights about the company and the market.

Require human oversight

Description:

Institutions can require humans to oversee the decisions of AI systems to ensure there is a valid, justifiable business case behind them.

Hypothetical application:

A subset of buy/sell orders proposed by an AI trading system can be manually verified by a human trader to ensure they fit with the firm's overarching trading strategy and risk tolerances, and are reasonable (i.e. not wildly disconnected from the fundamentals of the market/company).

While regulatory efforts to address these uncertainties are advancing, there is uncertainty about whether algorithmic collusion will be a likely outcome in the financial services industry

In the coming years, regulatory efforts to address uncertainties will continue to advance:

CVRIA

For example, The European Court of Justice imposed a judgement on Eturas, an online travel booking system, for illegally imposing a cap on discounts for online bookings. The judgement adds clarity to the definition of illegal concerted practices (i.e. collusion) on online platforms where direct meetings and communications are not necessarily required.[5]

However, there is uncertainty on whether AI collusion is a likely outcome, as several unlikely conditions would need to be met:

The products and services offered would need to be relatively homogenous

For collusion to be effective, there would need to be little differentiation between the products and services offered by different institutions, making price the only deciding factor for customers.

However, as explored in the 2018 World Economic Forum report *The New Physics of Financial Services*, AI creates the opportunity for institutions to build new and unique value propositions; this differentiation justifies higher and lower price points, making collusion less likely.

Institutions would need to have similar operating margins

Collusion is most viable when players have a similar profit margin, so no individual institution can offer prices or value with which other institutions cannot sustainably compete.

However, some firms are turning to operating models with a lower cost of service (e.g. fully digital vs. brick-and-mortar retail banks). For these players, not colluding and offering lower prices can lead to higher profits than would be the case with collusion.

Buyer power would need to be significantly limited

Collusion relies on customers having limited power (e.g. ability to switch providers), so even if a non-colluding firm provides a superior offering, it won't be able to draw customers away.

However, with the growth of price comparison engines and customers' increased data portability as a result of Open Banking regulations and the proliferation of screen scraping, buyer power is likely to increase and make collusion more difficult over time.

Looking forward

1 Regulators and technical experts will **need to work together to create new definitions for prohibited collusion, and define new governance frameworks to prevent it**; this would likely be a cross-industry effort, with specific guidance for the financial services industry to be developed by sector experts.

2 **While automated systems create a new risk of tacit collusion, they can also lead to increased price transparency** (e.g. through price comparison engines) **and competition** (e.g. by enabling new sources of differentiation on non-price-based factors).

References

1. https://voxeu.org/article/artificial-intelligence-algorithmic-pricing-and-collusion
2. https://rem.rc.iseg.ulisboa.pt/wps/pdf/REM_WP_077_2019.pdf
3. https://www.ft.com/content/9de9fb80-cd23-11e6-864f-20dcb35cede2?fbclid=IwAR20aVdcE5obADYzy2Gj4cXCkXXcswoxwg89pbFz3SoASpI57ycpKS-UJxs
4. https://www.lw.com/thoughtLeadership/lw-deep-dive-deep-learning-ftc-considers-artifical-intelligence
5. https://www.stibbe.com/en/news/2016/february/court-of-justice-clarified-the-concept-of-a-concerted-practice-for-unilateral-announcements

Closing comments

Ultimately, unlocking the benefits of AI will require the financial sector to navigate the new risks it introduces, while considering its interactions with other emerging technologies

***Responsibly deploying* AI in the financial ecosystem of today**

AI represents a fundamentally different paradigm in computing, requiring the financial industry to **develop and become comfortable with using completely new tools** to safeguard the financial system.

***Responsibly scaling* the AI-ubiquitous financial ecosystem of tomorrow**

The financial industry is currently experiencing the benefits of AI only in narrow opportunistic pockets; transformative change in the longer term will **raise challenging questions that require public-private collaboration** to adequately address them.

***Harnessing the potential* of a financial ecosystem built on responsible AI**

AI systems can have an impact that is quicker and of a greater magnitude than was the case for the traditional systems of the past; as a result, AI systems **should not be held to the same standards as humans and systems of today**, but to a higher bar.

It is critical to recognize that ***AI does not exist in a vacuum***. AI systems are tightly integrated with other technologies such as cloud and IoT. In the years to come, AI will become intertwined with an expanding set of emerging technologies such as 5G and quantum computing. The financial industry will need to **consider the impact of AI in the context of its interactions with these other technologies**.*

* The World Economic Forum is initiating an exploration of the interactions between emerging technologies. If you are interested in participating in workshops and interviews, please contact fintech@weforum.org

Appendix A: Acknowledgements

Contributors (1 of 5)

The project team would like to express its gratitude to the following subject matter experts who contributed their valuable perspectives through interviews, workshops and roundtable discussions (in alphabetical order):

Anil Aggarwal	Indus Group
Foteini Agrafioti	Borealis AI
Adeeb Ahamed	Lulu Financial Group
Jim Aiello	Greenwich Business Institute
Abdulaziz Al-Helaissi	Gulf International Bank
Rasheed Al Maraj	Central Bank of Bahrain
Jas Anand	Deloitte Canada
Sri Satish Ambati	H2O.ai
Thomas Ankenbrand	Institute of Financial Services Zug
Stefano Aversa	AlixPartners
Evangelos Avramakis	Swiss Re
Jo Ann Barefoot	Barefoot Innovation; Hummingbird Regtech
Marc Barrachin	S&P Global
Rainer Baumann	Swiss Re
Daniel Belfer	J. Safra Group
Livia Benisty	ComplyAdvantage
Michael Bodson	DTCC
Andrew Burt	Immuta
Ann Cairns	Mastercard
Gilberto Caldart	Mastercard
Claire Calmejane	Société Générale
Eric Cantor	Moelis & Company
Norm Cappell	Savvyy
Andrew Casey	Fidelity Labs
Sanjeevan Chandrasekaram	BNP Paribas
Chris Cheatham	RiskGenius
Shahzad Chohan	Credit Suisse
Jenny Chong	Credit Suisse
Carlo Cimbri	Unipol Gruppo
Stuart Coleman	10x Future Technologies
Gabriele Columbro	Fintech Open Source Foundation
Roger Crandall	MassMutual
Patrick Curry	Sedicii
Luke Davies	Barclays
Chris DeBrusk	BNY Mellon

Contributors (2 of 5)

The project team would like to express its gratitude to the following subject matter experts who contributed their valuable perspectives through interviews, workshops and roundtable discussions (in alphabetical order):

Lea Deleris	BNP Paribas
Charlie Delingpole	ComplyAdvantage
Alain Demarolle	My Money Bank
Charles Dugas	Element AI
Clara Durodie	Cognitive Finance
Tara Dziedzic	NYSE
Fatih Ebiglioglu	Kog Holding
Antony Elliott	Zurich Insurance Group
Tosha Ellison	Fintech Open Source Foundation
Adriana Ennab	Credit Suisse
Marco Enriquez	US SEC
Mattias Fras	Nordea
Adena Friedman	Nasdaq
Don Gossen	Ocean Protocol
Daniel Gorfine	Commodity Futures Trading Commission
Clayton Greene	NYSE
Ege Gürdeniz	Oliver Wyman
Doug Hamilton	Nasdaq
Matthew Hampson	Nomura
Kevin Hanley	Royal Bank of Scotland Group
Brian Hartzer	The Westpac Group
Catherine Havasi	Luminoso
Gerard Hester	Morgan Stanley
Hanno Hinsch	Morgan Stanley
Christoph Hock	Union Investment/Plato Partnership
Bernhard Hodler	Julius Baer
Christian Hoffmann	ETH Risk Center
Charlotte Hogg	Visa
Obaid bin Humaid Al Tayer	Ministry of Finance, United Arab Emirates
Sean Hunter	OakNorth
Derek Hurlbert	Zurich Insurance Group
David Iakobachvili	Orion Heritage Co
Tanmoy Jadhav	SWIFT
Carsten Jung	Bank of England
Michal Kaczor	ComplyAdvantage

Contributors (3 of 5)

The project team would like to express its gratitude to the following subject matter experts who contributed their valuable perspectives through interviews, workshops and roundtable discussions (in alphabetical order):

Husayn Kassai	Onfido
Sabine Keller-Busse	UBS
Brian Kennedy	Nedbank Group
Yong Hyun Kim	Hanwha Asset Management
Hwan Kim	Deloitte Canada
Artem Korenyuk	DTCC
Gorkem Koseoglu	ING Group
Deepak Krishnamurthy	SAP
Ryan Krook	Borrowell
Alex LaPlante	Borealis AI
Gottfried Leibbrandt	SWIFT
Charles Li	Hong Kong Exchanges and Clearing
David Lipton	International Monetary Fund
Irene Lopez de Vallejo	Ocean Protocol
Howard Lutnick	Cantor Fitzgerald
Blair Mackasey	JP Morgan
Sunil Madhu	Socure
Andy Maguire	HSBC
Promoth Manghat	Finablr
Madhavi Mantha	Element AI
Vincenzo Marchese	UBS
Sherry Marcus	BlackRock
Alison Martin	Zurich Insurance Company
Ricardo Martin Manjon	BBVA
Bharat Masrani	TD Bank Group
Peter Matlare	Absa Group
Richard Maton	Aperio Strategy
Daniel Mayer	Deloitte UK
Bruce McGuire	Connecticut Hedge Fund Association
Luis Menendez	SAP
Ravi Menon	Monetary Authority of Singapore
Jerry Miller	Guggenheim Investments
Szymon Mitoraj	PZU
Thayer Moeller	Barings
Anish Mohammed	Singularity University/EthicsNet
Daragh Morrissey	Microsoft

Contributors (4 of 5)

The project team would like to express its gratitude to the following subject matter experts who contributed their valuable perspectives through interviews, workshops and roundtable discussions (in alphabetical order):

Henri Mouy	Natixis
Peter Moyo	Old Mutual Limited
Henrike Mueller	Financial Conduct Authority
Suchitra Nair	Deloitte UK
Michael Natusch	Prudential
Nina Neer	Credit Suisse
Giang Nguyen	Lazard
Zhu Ning	Tsinghua University
Illah Nourbaksh	Carnegie Mellon University
Jim Ovia	Zenith Bank
Richard Peers	Microsoft
Ana Perales	Barclays
Peter Poon	Bank of Montreal
Thomas Puschmann	University of Zurich
Jo Rabin	Deutsche Bank
Sunil Rawat	OmniScience
Hélène Ray	London Business School
Falk Rieker	SAP
Steven Roberts	Barclays
David Rogers	Deloitte UK
Daniel Ryan	Swiss Re
Koby Sadan	Viking Global
Waleed Saeed Al Awadhi	Dubai Financial Services Authority
Peter Sarlin	Silo.AI and Hanken School of Economics
Claudio Scardovi	AlixPartners
Stefan Schmittmann	Commerzbank
John Schultz	Hewlett Packard Enterprise
Barry Schwartz	Vitality Group
Robert Sears	BBVA
Vasuki Shastry	Standard Chartered Bank
Binay Shetty	Finablr Limited
Siddharth Singh	Bank of America Merrill Lynch
Rahul Singh	HCL Technologies
Suren Siva	Credit Suisse
Sean Slotterback	Decipher Finance
Jeremy Smith	RiskGenius
Ivan de la Sota	Allianz SE
Nathan Stevenson	Forwardlane

Contributors (5 of 5)

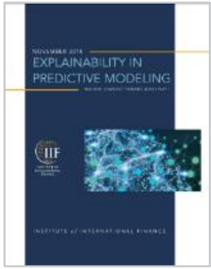
The project team would like to express its gratitude to the following subject matter experts who contributed their valuable perspectives through interviews, workshops and roundtable discussions (in alphabetical order):

Ben Stokes	Actual Intelligence	**Mark Wagner**	Scotiabank
Andreas Streich	Zurich Insurance Group	**Steven Walden**	AXA XL
Ted Stuckey	QBE Ventures	**Larry Wall**	Federal Reserve Bank of Atlanta
Tom de Swaan	ABN AMRO	**David Wang**	Lazard
Valerie Szczepanik	US SEC	**Hui Wang**	PayPal
Keiko Tashiro	Daiwa Securities Group	**Lara Warner**	Credit Suisse
Larry Thompson	DTCC	**Amy Webb**	Future Today Institute
Carlos Torres Vila	BBVA	**Aric Whitewood**	XAI Asset Management
Alan Trefler	Pegasystems	**Jacek Wieclawski**	Rabobank
Antony Turner	Coefficiency Lab	**Grant Wilson**	Neptune Networks
Darryl Twiggs	SmartStream	**Dennis Wisnosky**	EDM Council
Lance Uggla	IHS Markit	**Gabriel Woo**	RBC Ventures
Rob Underwood	Fintech Open Source Foundation		
Matthew Van Buskirk	Hummingbird Regtech		
Sabine VanderLinden	Startupbootcamp InsurTech		
Ronnie van der Wouden	Rock Creek Group		
Prema Varadhan	Temenos		

Appendix B: Further reading

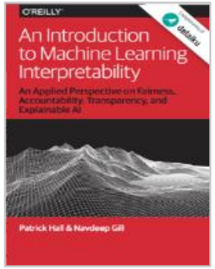
Further reading

The following texts were instrumental in shaping the perspectives of the project team. For those interested in exploring further and learning more about the topics covered in this report, we recommend reading the following documents:



Explainability in Predictive Modeling

International Institute of Finance, 2018



An Introduction to Machine Learning Interpretability

Patrick Hall, Navdeep Gill, 2018

Geoff Hinton Dismissed the Need for Explainable AI: 8 Experts Explain Why He's Wrong

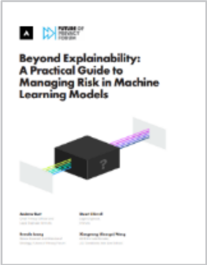
Forbes, 2018

Intelligent interfaces

Deloitte Tech Trends, 2019

XAI Program Update

DARPA, 2017

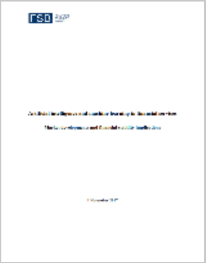


Beyond Explainability

Andrew Burt, 2018

Responsible AI in Consumer Enterprise

Kathryn Hume, 2018



Artificial intelligence and machine learning in financial services

Financial Stability Board, 2017

Future of computer trading in financial markets: an international perspective

UK Government Office of Science, 2012

Bias and Ethical Implications in Machine Learning

International Institute of Finance, 2019

Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms

Nicol Turner Lee, Paul Resnick, Genie Barton, 2019

Algorithms and fiduciaries: Existing and proposed regulatory approaches to artificially intelligent financial planners

John Lightbourne, 2017

Contact information

For questions or additional information, please contact:

World Economic Forum Project Team

R. Jesse McWaters
Financial Innovation Lead
World Economic Forum;
jesse.mcwaters@weforum.org

Matthew Blake
Head of Financial & Monetary System Initiatives
World Economic Forum;
matthew.blake@weforum.org

Professional Services Support from Deloitte

Rob Galaski
Partner, Deloitte Canada;
Global Leader, Banking & Capital Markets, Deloitte Consulting;
rgalaski@deloitte.ca

Hemanth Soni
Senior Consultant
Monitor Deloitte, Deloitte Canada;
hemasoni@deloitte.ca

Ishani Majumdar
Senior Consultant
Omnia AI, Deloitte Canada;
ismajumdar@deloitte.ca

WORLD
ECONOMIC
FORUM