On the use of artificial intelligence in financial regulations and the impact on financial stability*

Jon Danielsson
London School of Economics

Andreas Uthemann
Bank of Canada
Systemic Risk Centre, London School of Economics

February 2024

First version September 2023

Abstract

Artificial intelligence (AI) can undermine financial stability because of malicious use, misaligned AI engines and since financial crises are infrequent and unique, frustrating machine learning. Even if the authorities prefer a conservative approach to AI adoption, it will likely become widely used by stealth, taking over increasingly high-level functions, driven by significant cost efficiencies and its superior performance on specific tasks. We propose six criteria against which to judge the suitability of AI use by the private sector for financial regulation and crisis resolution and identify the primary channels through which AI can destabilise the system.

*Corresponding author Jon Danielsson, J.Danielsson@lse.ac.uk. We thank Charles Goodhart, Gudmundur Kristjansson, Eva Micheler, Robert Macrae, Inaki Aldasoro, Leonardo Gambacorta, Vatsala Shreeti and Bruno Tissot for valuable comments. Updated versions of this paper can be downloaded from from `modelsandrisk.org/appendix/AI`. We thank the Economic and Social Research Council (UK) [grant number ES/K002309/1] for their support. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the Bank of Canada.

Electronic copy available at: https://ssrn.com/abstract=4604628

1 Introduction

Artificial intelligence (AI) is transforming the financial system, promising improved efficiency, robustness and impartiality at much lower costs than existing arrangements. AI also threatens financial stability when AI vulnerabilities viciously interact with economic fragilities.

While there is no single notion of what AI is, it is helpful to see it as a computer algorithm performing tasks usually done by humans. AI differs from machine learning (ML) and traditional statistics in that it not only provides quantitative analysis but also gives recommendations and makes decisions.[1] Today's AI excels at extracting patterns from large, unstructured datasets for use in classification and prediction tasks. But they do not reason in the same way humans do. Once AI achieves proficiency at a given task — has been trained — it can advise human decision makers or make decisions autonomously. Norvig and Russell (2021) list a number of possible definitions of AI. Among these, AI as a rational maximising agent resonates with the economic notion of utility maximising agents and, hence, is particularly helpful in the analysis of AI use in the financial system.

A financial authority entrusted with regulating the financial system has two objectives. The first is microprudential regulation, or micropru, concerned with day-to-day issues such as risk management, consumer protection and fraud. AI is generally beneficial for micropru as data is ample, the rules are mostly fixed on the timescale decisions are made and the cost of mistakes is small.

The second policy objective is macroprudential regulations, or macropru, focusing on broad-picture issues relating to financial stability. The emphasis here is decidedly long run, both the avoidance of large losses and systemic financial crises years and decades into the future, as well as the resolution of crises when they occur. Macropru is more difficult to execute and less accurate than micropru, as events to be controlled are infrequent and often unique. That implies considerable uncertainty about the ability of AI and the dangers arising from its use, as noted by Danielsson, Macrae and Uthemann (2022).

It is helpful to think of AI's application to financial policy on a spectrum. At one end, we have a problem with ample data and clear and simple objectives, playing to the strength of AI and it's learning algorithms. As the frequency of events drops and rules are increasingly mutable, the AI advantages erode until we reach once-in-a-working-lifetime events, such as the resolution of a systemic financial crisis, where

[1] For general overview of AI Russel (2019) is particularly useful. For a more technical review of the underlying ML techniques, see Murphy (2023) and Prince (2023).

Electronic copy available at: https://ssrn.com/abstract=4604628

AI, though very valuable, can threaten the system's stability.

The use of AI in the financial system is growing rapidly. Private sector financial institutions apply it to tasks such as risk management, credit allocation, fraud detection and regulatory compliance. Even if many traditional banks have been reticent in its adoption, some employ large AI teams, and challenger banks have been particularly enthusiastic about its use. Significant cost savings and efficiency improvements in a highly competitive landscape will likely rapidly drive the further adoption of AI in the private sector.

The financial authorities need to respond, and most are formulating AI policy (see, e.g. Kiarelly et al., 2024; Moufakkir, 2023). While some authorities might prefer a slow, deliberative and conservative approach to AI, that will not be tenable for three reasons. First, an authority that persists in regulating with traditional methods may find itself outmanoeuvred. Second, a policy of not using AI for high-level decisions will likely be undermined by AI being adopted by stealth. Finally, AI may be considered essential for advising on a critical task. Even if an authority wanted to limit AI to basic advisory roles, there might not be much difference between AI making decisions and AI providing crucial advice, particularly if its internal representation of the system is not intelligible to its human operators.

The usefulness of AI for the financial authorities is directly affected by five conceptual economic challenges that can viciously interact with AI vulnerabilities. The first, perhaps paradoxically, is data. After all, the financial system may appear to be the ideal use case for AI because it creates vast amounts of data, leaving plenty for AI to train on. However, financial data collection is incomplete, and recorded data are often inconsistently or even inaccurately measured, leading to non-representative samples. What is more, financial crises are rare, meaning there is little data available on the most important events for macropru.

The second challenge arises from the uniqueness of crises. While major financial crises share common fundamental features, every crisis is unique because the driving factors are specific to the institutional structure and the particular regulatory and political regimes in place. Rarity and uniqueness in a sparsely monitored system imply that systemic crises can be seen as "uncertain" in Frank Knight's (1921) classification and thus do not fit well into the conventional ML paradigm.

AI's third conceptual challenge relates to how the financial system responds to control, echoing Goodhart's (1974) law and the Lucas (1976) critique. New regulations change the structure of the financial system in ways that are hard to predict, making it difficult for AI to learn from historical data about the possible consequences of regulatory actions.

Electronic copy available at: https://ssrn.com/abstract=4604628

The fourth conceptual problem stems from a lack of clarity on what objectives the AI should pursue. This is particularly problematic for macropru, where the regulators' problem is often described in abstract high-level terms such as "ensure financial stability". The objectives of the most important macropru actions only emerge in times of crisis, in a process led by the political leadership.

The final conceptual challenge relates to difficulties in aligning AI's incentives with the authorities' objectives. The purpose of financial regulations is to align the private sector's incentives with society at large. The interactions between financial supervisors and private sector decision-makers can be seen as a difficult principal-agent (PA) relationship because of incomplete contracting over a latent variable, risk, in a world with explicit and implicit government guarantees. With AI, the one-sided financial institution-regulator problem becomes two-sided, amplifying the already significant existing PA problem in financial regulation.

The literature on AI has identified particular risks that AI engines pose; see, for example, Barnichon et al. (2022), and several researchers, such as Weidinger et al. (2022), Bengio et al. (2023) and Shevlane et al. (2023) identify societal risks arising from AI. When studying how these interact with the economic conceptual challenges listed above, we find four channels for how AI use may destabilise the financial system.

The first channel is the malicious use of AI by its human operators, a particular concern in the financial system because it is replete with highly resourced profit-maximising economic agents not too concerned about the social consequences of their activities. Such agents can bypass controls and change the system in a way that benefits them while being difficult for competitors and regulators to detect. They may even deliberately create market stress, which is highly profitable for those forewarned. These agents either directly manipulate AI engines or use them to find loopholes to evade control. Both are easy in a financial system that is effectively infinitely complex. Such activities can be socially undesirable and even be against the interests of the institution employing the operator of the AI engine. We expect the most common malicious use of AI will be by employees of financial institutions, careful to stay on the right side of the law. AI will likely also facilitate illegal activities, such as rogue traders and criminals, as well as terrorists and nation-states aiming to create social disorder.

The second channel is due to the users of AI being both misinformed about its abilities and strongly dependent on it. That is most likely when data-driven algorithms, such as those used by AI, are asked to extrapolate to areas where data is scarce and objectives unclear, which is very common in the financial system. AI engines are

Electronic copy available at: https://ssrn.com/abstract=4604628

designed to provide advice even when they have very low confidence about the accuracy of their answer. They can even make up facts or present arguments that sound plausible but would be considered flawed or incorrect by an expert, both instances of the broader phenomenon of AI hallucination. The risk is that AI will present confident recommendations about outcomes they know little about. To overcome that, the engines will have to provide an assessment of the statistical accuracy of their recommendations. Here, it will be helpful if the authorities overcome their frequent reluctance to adopt consistent quantitative frameworks for measuring and reporting on the statistical accuracy of their data-based inputs and outputs.

The third channel emerges from the difficulties in aligning the objectives of AI with those of its human operators. While we can instruct AI to follow our instructions, there is no guarantee it will actually do so. It is impossible to pre-specify all the objectives AI has to meet, a problem since AI is very good at manipulating markets and being incentivised by high-level objectives such as profit maximisation is not concerned with the ethical and legal consequences of its actions unless explicitly instructed. The superior performance of AI can destabilise the system even when it is only doing what it is supposed to do. That is particularly problematic in times of extreme stress when the objective of financial institutions, and hence the AI working for them, is survival, amplifying existing destabilising behaviour, such as flights to safety, fire sales and investor runs. More generally, AI will find it easy to evade oversight because it is very difficult to patrol a nearly infinitely complex financial system. The authorities have to contend with two opposing forces. AI will be very helpful in keeping the system stable but, at the same time, aids the forces of instability. We suspect the second factor dominates. The reason is that AI attempting to evade control only has to find one loophole to misbehave, while the supervisors not only need to find all the weak points but also monitor how AI interacts with each of them and then effectively implement corrective measures. That is a very difficult computational task, made worse by the private sector having access to better computational resources than the authorities. The more we use AI, the more difficult the computational problem for the authorities becomes.

The final channel emanates from the business models of those companies designing and running AI engines exhibits increasing returns to scale. AI analytics businesses depend on three scarce resources: compute, human capital and data, pushing the AI industry towards an oligopolistic market structure dominated by a few large vendors. The end result is amplified procyclicality and more booms and busts. Suppose the authorities also depend on the same AI engine for its analytics, which seems likely. In that case, they may not be able to identify the resulting fragilities until it is too late. In other words, the oligopolistic nature of the AI analytic business increases

Electronic copy available at: https://ssrn.com/abstract=4604628

systemic financial risk. It is a concern that neither the competition nor the financial authorities appear to have fully appreciated the potential for increased systemic risk due to oligopolistic AI technology in the recent wave of data vendor mergers.

Along the way, the use of AI in both the public and private sectors will need to be regulated. The reliance of both the regulated and regulators on the same AI vendors, along with challenges in data sovereignty and attribution of mistakes, demands special attention from the architects of future AI regulations. Meanwhile, the public sector's use of AI in regulations raises tough questions. Who is accountable when a regulator's AI makes decisions or provides crucial inputs for human decisions, and how can a regulated entity challenge decisions? The regulatory AI may not be able to explain its decision in terms that are intelligible to the subjects of regulatory actions, the AI's human owner or the judges involved in arbitration processes. To ensure the quality and safety of AI performance, authorities might find benchmarking AI against defined tasks a particularly fruitful way forward.

We do not want to overemphasise these issues. We suspect that the benefit of AI will be overwhelmingly positive in the financial system. However, the authorities must be alive to these threats and adapt regulations to meet them. The ultimate risk is that AI becomes both irreplaceable and a source of systemic risk before the authorities have formulated the appropriate response.

The organisation of the paper is as follows. After the introduction, we present the conceptual challenges in Section 2 and follow that with a discussion on artificial intelligence and machine learning in Section 3. Section 4 is focused on the use of AI in the financial system, while Section 5 contains the five channels for how AI can destabilise the financial system. We then apply those to the regulation of AI in Section 6. Section 7 concludes.

2 Economic conceptual challenges

> *"Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."*
>
> Charles Goodhart's (1974) Law.

Any attempt at controlling the financial system is frustrated by the presence of highly resourced financial market participants operating in very uncertain social environments that are subject to frequent structural changes. The rules of the game are usually unclear, and the actors often do not know each other's objectives. The

Electronic copy available at: https://ssrn.com/abstract=4604628

participants can even change the rules to their advantage in a way that others only partially observe. Agents interact by competing, cooperating or even colluding with each other, cheating and deceiving along the way. They learn from interactions with the system and are continually changing their beliefs and, hence, how they will react in similar situations in the future.

Any authority, be it AI or human, government regulators or internal controllers will have to content with five economic conceptual challenges that get in the way of their mission. However, the inherent characteristics of AI make AI controllers particularly affected by these challenges.

2.1 Usefulness of data

Data should play to AI's advantage as the financial system generates many petabytes of it daily. Every transaction is recorded, all decisions are documented, decision makers are monitored and recorded, and we can track processes over their lifetime. Financial institutions must report some of this data to the financial authorities, and the authorities can demand almost all of it later. One might expect this ocean of data would make it easy for AI to study the financial system in detail and identify all the causal relationships. That is true for most microprudential problems, but not the macroprudential.

Start with the basic measurement. The standards for recording financial data are inconsistent, so different stakeholders might not record the same activity in the same way, leading to complex matching problems. Identification coding and database design can differ significantly. Financial institutions have a lot of legacy systems not set up with data collection and sharing in mind, rendering data collection, especially in a format that is standard across the industry and necessary for the authorities, costly and error prone. Fortunately, while real today, these problems are rapidly being overcome, not the least with the help of AI.

A bigger challenge is all the silos in the regulatory structure that hinder data sharing. Most data stays within a financial institution and is not shared. Even when shared, the financial institution might retain copyright, allowing it to control who sees the data so that it might be available for compliance but not for broader objectives such as financial stability. Furthermore, financial system data are collected by authority silos where data sharing is limited. There might be restrictions on data sharing between the supervisory and statistical unit of a central bank, between authorities in the same country or between jurisdictions. These problems were made clear in the crisis in 2008, where nobody had an overview of the aggregate market for structured

Electronic copy available at: https://ssrn.com/abstract=4604628

credit. The situation has improved somewhat since then due to mandatory trade reporting for many derivatives transactions and the increasing availability of data on cash and repo transactions for bond markets. However, silos persist and digesting the large amount of data remains a challenge.

Finally, when it comes to the most serious events, systemic financial crisis, the events under consideration are, fortunately not frequent. The typical OECD country only suffers a systemic crisis one year out of 43, according to the crisis database maintained by Laeven and Valencia (2018). However, that fortunate low frequency of crises frustrates the data driven analytics AI depends on.

These three issues, data quality, silos and rarity of events, are usually not all that important for microprudential regulations. Not so for macropru, where they amplify each other, negatively impacting the design of macroprudential regulations, enforcement, and crisis resolution.

2.2 Unknown-unknowns

The second conceptual problem arises from most crises being unknown-unknown events that are both unique and infrequent. It is almost axiomatic that the type of event a macroprudential authority is concerned with plays out in unexpected ways. Otherwise, precautionary actions would have been taken to avoid a crisis.

Unfortunately, from the point of view of AI learning from these crises, the details of a given crisis are, in important aspects, unique to it. Fundamentally, every crisis is caused by the same set of fundamental vulnerabilities, all of which act as crisis amplifiers. Financial institutions that use high degrees of leverage that render them vulnerable to shocks, self-preservation in times of stress, leading market participants to prefer the most liquid assets, and system opacity, complexity and asymmetric information, causing market participants to mistrust each other in times of heightened uncertainty. However, these vulnerabilities are essentially conceptual, and when it comes to designing regulations to prevent stress and mitigating it when it happens, the authorities have to focus on details. Those details are unique to each crisis. That is almost self evident because the supervisors would have prevented a crisis if they were not.

While the authorities can scan the system for specific causes of vulnerabilities, their job is frustrated by the almost infinite complexity of the financial system. The supervisors can only patrol a small part of that infinitely complex system. Even if the supervisors, AI or human, could monitor all threat scenarios and assign a probability to each — an impossible task — they still have the problem of picking notification

Electronic copy available at: https://ssrn.com/abstract=4604628

thresholds. The system's complexity and measurement noise means that the number of notifications would be very large, with mostly false positives. Furthermore, such intrusive monitoring might sharply curtail desirable risk taking because of the false positives, and be seen as socially unacceptable.

This uniqueness of crises creates particular problems for the designers of macroprudential regulations because they generally only learn what data is most useful after a stress event. That was, for example, the case in the crisis in 2008, where we only realised afterwards that sub-prime mortgages being put into structured credit products — the banks held onto the most senior and junior tranches where risk modelling was extremely poor — was the key channel for the crisis. Obvious after the event but practically impossible to discover before. When the analyst faces an almost infinite number of signals and an enormous number of false positives, it is very difficult, to the point of impossible, to identify which data is useful until after a crisis event is already underway. It is too late to have preventative regulations in place by this time.

This means that the most severe financial crises are, by definition, unknown-unknowns or uncertain in Frank Knight's (1921) classification.

2.3 System responses

A key challenge for AI working for the macroprudential, but not generally for the microprudential authorities, relates to the dynamic interaction of the financial system participants. A helpful framework for understanding the problem is the Lucas (1976) critique, which states that the decision rules used by economic agents depend on the underlying economic environment and can change as regulations change, undermining their effectiveness. An example of Goodhart's (1974) law, "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." Changes to financial regulations or the level of supervision, including changes to the crisis resolution playbook, as well as decisions taken during resolution, will change the responses of the private sector to a similar crisis in the future in unexpected ways.

Of particular interest when AI is to be used for economic analysis relates to how Lucas and Goodhart came to their conclusions. Prior to the 1970s, economists believed in large general models informed by statistical patterns, with the Phillips curve, representing a trade-off between unemployment and inflation based on a negative empirical correlation of these two variables, as a centrepiece.[2] The contribution of

[2]For an important example, see Rasche and Shapiro (1968) explaining the MIT-Fed model used

Electronic copy available at: https://ssrn.com/abstract=4604628

Lucas was to recognise that when the authorities try to exploit this apparent trade-off, they change the way the private sector negotiates wages and sets prices, leading to a breakdown of previously identified correlations. Further attempts to stimulate the economy through this channel create inflation without reducing unemployment, giving rise to the word stagflation. In response, the economic profession moved away from large-scale simultaneous equation models based on statistical correlations to smaller, structural models that explicitly modelled beliefs and general equilibrium effects so that reactions to policy interventions followed a well understood and internally consistent logic encoded in the model, perhaps an early example of interpretability. An AI engine would need to do the same if it were to become an effective macropru regulator.

One direct impact is on measuring financial risk, an essential task for any regulator. Besides the sampling issues discussed in Section 2.1 above, there are particular technical issues for why risk measurements can be misleading. It is helpful to use the classification scheme proposed by Daníelsson and Shin (2002), which separates financial risk into two categories: exogenous and endogenous. Exogenous risk emphasises risk measured by statistical techniques based on historical outcomes in financial markets, typically prices. Endogenous risk, by contrast, captures risk that arises from the strategic interaction of the economic agents that make up the financial system.

Exogenous risk is easy to measure, and AI excels at it. Identifying endogenous risk is difficult because it captures outcomes only visible in extreme stress when self preservation and mistrust of counterparties are crisis amplifiers.

The relative importance of exogenous versus endogenous risk depends on the problem. For most microprudential regulations, the frequency of events and lack of large strategic interactions means that assuming risk is exogenous is usually quite innocuous. However, any authority using data driven analytics that uses exogenous risk measurement for assessing the risk of financial instability, will likely be seriously misled as to how the financial system evolves in times of stress.

2.4 Pre-specified objectives and distributed decision making

The fourth conceptual challenge arises from the clarity of the objectives AI optimises for. In the best case scenario, it knows the objectives, and can in many cases use reinforcement learning to identify solutions in real time. While AI can operate in an environment with mutable objectives, it is less effective and more prone to mistakes as the rarity of events and the cost of mistakes increase. The worst case for AI is

by the Federal Reserve and developed in collaboration with MIT.

Electronic copy available at: https://ssrn.com/abstract=4604628

when the objective is unknown ex-ante and cannot be learned, and that is where its potential for making catastrophically wrong advice and decisions is the strongest.

Mutable objectives do not pose much of a problem to micropru. The rulebook is known and usually static on the timescale decisions are made. While it evolves in response to events and the regulated response to regulations, AI can quickly update its understanding of the objectives, by adopting changes to the rulebook, learning from observing how the human supervisors act and reinforcement learning.

This is not the case in macropru, as it operates on very long timescales. The time between relevant events is usually in decades. It can be very difficult to define the macroprudential objective except at the highest levels of abstraction, such as the prevention of severe dysfunction in key financial markets and especially the failure of systemically important institutions. This applies to the design of macroprudential regulations, macroprudential supervision and crisis resolution. It can be difficult to make a case for the need to allocate significant resources to prevent something that only happens in the distant future, and the macroprudential regulations are susceptible to lobbying and political interference, which means that the objectives of regulations can change over time. Learning from previous crisis interventions might not be of much help as the concrete goals the actions under consideration are to achieve are likely specific to particular circumstances and political environments.

The most severe financial crises can have catastrophic consequences if not addressed adequately, with direct economic costs in the several trillions of dollars, as noted by Barnichon et al. (2022), and a large number of people materially affected. When that happens, society demands we do what it takes to resolve the crisis. Often, the rules and the laws in place might stand in the way of the most effective crisis resolution. Emergency sessions of Parliament to rectify that are not uncommon, such as Switzerland's resolution of Credit Suisse.[3]. Emergency constitution clauses might be invoked. Pistor (2013), in her legal study of the resolution of financial crises, finds that if the existing law prevents the most effective course of action, there is acceptance from the political and judicial system to suspend the law in the name of the higher objective of crisis resolution. Furthermore, when a severe crisis happens, the political leadership takes charge, inevitable because if it becomes necessary to change or bypass the law or significantly redistribute resources, the political leadership is the only entity with the necessary legitimacy. Given the fluidity of this process, it is difficult to see how one would specify the objectives for an AI so that it can provide real-time advice for crisis resolution or make decisions.

[3]https://www.bloomberg.com/news/articles/2023-03-20/credit-suisse-collapse-reveals-some-ugly-truths-about-switzerland-for-investors

Electronic copy available at: https://ssrn.com/abstract=4604628

We have a long experience of resolving crises and have a relatively good understanding of the process. The regulatory system is usually modular, with separate authorities and fiercely guarded mandates. In the most severe crises, these silos break down. All relevant authorities, the affected private sector, the judiciary, and especially the political leadership have come together to decide how to resolve the crisis. Government ministers usually lead this process. This may involve the same entities in other financial centres in a global crisis. Each stakeholder brings their own education, ethics, professional experience and objectives to the table. Such a process can be highly robust. All pertinent issues are discussed, including information that was confidential or implicit until then. Such analysis depends on implicit knowledge and intuitive understanding its participants have of the current situation and each others' views and objectives. Intuition can mean finding analogous problems from seemingly unrelated domains that can provide creative solutions for the current crisis that has not been encountered before. This process of successful extrapolation is often vital to crisis resolution and can imply ignoring or revaluing previously pursued goals. There are many such examples in history. In his re-evaluation of the 1866 crisis, Bagehot (1873) invented modern financial stability policies, like lending of last resort, while the German central banker, Hjalmar Schacht, used a short squeezing to stop the hyperinflation in 1923. The Swedish central bank in 1992 created the good bank-bad bank model for crisis resolution. Reproducing this process will be particularly difficult for AI as it implies both shifting attention to seemingly unrelated problems and potentially modifying its objective function.

Crisis resolution is arguably the most important aspect of financial policy, especially for central banks — one of their raisons d'être. Nevertheless, given the above challenges — scarce data, unknown-unknowns, endogenous structural changes in response to attempted control — humans struggle with this task, and this is where AI could benefit them the most. Unfortunately, it has to overcome the same conceptual problems, and this challenge is as difficult for AI as for human regulators, if not more so. Paradoxically, progress in AI's suitability for the task might come from a better human understanding of these problems that can then be translated into better algorithms.

2.5 Incentives

The objectives of the various market participants are specific to them and often in conflict with those of other participants and society. That is why we have financial regulations: to align the private sector's incentives with society at large. The interactions between financial supervisors and the private sector decision-makers can

Electronic copy available at: https://ssrn.com/abstract=4604628

be seen as a principal-agent (PA) relationship, where the highly resourced and sophisticated private sector — the agent — wants to get on with successfully running its business. In contrast, the supervisors — the principal — aim to ensure that the agents operate a financial system that efficiently channels savings to investments while not abusing their clients or taking excessive risks that can cause costly crises.

The PA problem applied to financial markets is particularly difficult for three reasons. First, contracting is incomplete, meaning we cannot specify what will happen in every contingency. We might not even know what the contingencies are. Second, the very behaviour to be controlled is risk taking, and since risk is a latent variable, it cannot be measured directly but only imprecisely inferred from market movements. Consequently, it can be difficult to determine whether an undesirable outcome is due to misbehaviour or innocent bad luck. Finally, because implicit or explicit government guarantees are fundamental to all financial regulations, market participants, knowing that some of their risk taking is publicly insured, can behave in a way that makes government bailouts more likely, creating moral hazard.[4]

As we increasingly use AI in regulations and the private sector, a new dimension in the PA problem emerges. The one-sided financial institution-regulator problem becomes two-sided, with the addition of the relationship between the regulators and financial institutions as principals and their respective AI as agents. The reason is that we cannot expect the AI engines' objectives to be fully aligned with its human operator as AI can take hidden actions to achieve its goals. This means the human operator needs to incentivise the AI to act in its owner's interest.

The presence of the two-sided PA problem amplifies the already significant existing PA problem in financial regulation. We have in the past seen many cases where market participants actively and deliberately create heightened stress since anybody forewarned of stress can profit. Creating and amplifying crises is profitable. Lowenstein (2000) shows just one example from the LTCM crisis in 1998. AI might find it even easier to act in that way. It will, therefore, be essential for the authorities to explicitly consider how to align the AI engines, whether directly operated by the regulated or by their AI vendors, with the objectives of the authority.

Existing mechanisms for managing the regulator-bank PA relationship will not be adequate when AI is involved. In the technical terminology of moral hazard, AI has more hidden actions at its disposal and will command higher information rents. In addition, AI's objectives can be less straightforward than those of traditional financial institutions that trade-off risks and returns. How do you incentivise an algorithm

[4]Dewatripont et al. (1994) provides a detailed analysis of the problem of financial regulation with incomplete contracts, government guarantees and private sector moral hazard.

Electronic copy available at: https://ssrn.com/abstract=4604628

whose objective is to predict the next token correctly? What higher level objectives does this loss function induce, and how do we formulate an incentive scheme that imposes the correct constraints on AI behaviour? How can we make it internalise negative externalities, such as the possibility of deliberately creating fire sales or runs to trigger privately beneficial but publicly costly bailouts?

3 Artificial intelligence and machine learning

There is no standard definition of artificial intelligence (AI). One of the pioneers of AI, John McCarthy, saw it as "the science and engineering of making intelligent machines",[5] where perhaps "intelligence measures an agent's ability to achieve goals in a wide range of environments" in the definition of Legg and Hutter (2007). Generally, it is a computer algorithm that performs tasks that a human would otherwise do. For a general view of AI, Russel (2019) is particularly useful.

AI distinguishes itself from machine learning (ML) and traditional statistical modelling in that it is also used to do analysis, provide recommendations and make decisions, not merely give quantitative inputs for human decision making as a statistical algorithm would. AI charged with tasks aims to find the best course of action given its objectives and its understanding of the problem domain.

ML provides the fundamental building blocks of most AI, finding statistical regularities in large data sets and helping provide AI engines with a deeper understanding of the underlying latent structure that generates the data of a given problem domain. One of the key differences to conventional applied statistics is that the input data tend to be unstructured, are often text-based and large in size, where the models can encode highly complex statistical relationships. The modeller usually does not supply a prior understanding of the parameterised model structure. See the Appendix for details on ML that are useful in financial applications.

3.1 From reactive machines to AGI

Even though the ability of AI engines has increased rapidly, today's AI merely match patterns and do not understand in the same way a human does. It is unclear what achieving such an understanding will involve. When considering the abilities of various AIs, including those only hypothesised, it is beneficial to classify AI based on its abilities. As it turns out, that is a surprisingly difficult task since the experts

[5] http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

Electronic copy available at: https://ssrn.com/abstract=4604628

do not agree on classification, terminology, ability or future potential. At the risk of oversimplification, there are four types of AI.

The first is often termed a *reactive machine*, which is a system with no memory of previous decisions and designed to perform a specific task. In other words, it only works with already available data. Typical examples are classification engines that identify objects and photographs and recommendation engines.

The second category is *limited memory* AI, which both has knowledge of what happened in the past and keeps track of past actions and observations and can condition its response to a given situation on these past experiences. Generative AI, such as ChatGPT, fall into this category, systems that predict the next word or phrase, or some other object, based on the content it is working with. Other examples include virtual assistants and self driving cars. Time series models are also of this type.

We then have a *theory of mind*, not yet realised, AI that aims to understand thoughts and emotions and consequently simulate human relationships. They can reason and personalise interactions. Importantly, from the point of view of financial policy, a theory of mind AI should be able to understand and provide context for policy decisions, which today's AI cannot do, and hence be invaluable in crisis resolution. AI that understands how its interlocutors think can explain its decisions to them and reason about how they will react to its actions and plan accordingly.

And finally, *artificial general intelligence* or AGI. It is still purely theoretical while often hypothesised. It would use existing knowledge to learn new things without needing human beings for training and being able to perform any intellectual task humans can. And if it reaches that level, increased computer capacity would allow it to surpass humans.[6]

The potential for the eventual emergence of a theory of mind and AGI AI remains controversial, and experts have strong disagreements on both the likelihood of such eventualities and what it takes to reach them. The ability of an AI engine depends on three factors: network structure, data and compute. The more relevant data we feed into the engine and the more compute we apply, the better it will perform. What is controversial is whether that is all we need to make progress on the theory of mind and AGI or if a conceptual leap in the ways of learning is also required.

A point of controversy is whether the next token prediction is sufficient for achieving AGI. In that case, all that is required for AGI is a sufficient amount of training data and compute. We would not need conceptual improvements in network technology.

When transiting from using AI only to learn about the world to use it to advise on

[6]Kasirzadeh (Kasirzadeh) discusses existential risks arising from AGI

Electronic copy available at: https://ssrn.com/abstract=4604628

real-world decisions and even make such decisions, a key challenge is AI alignment and safety, that is, how to make AI systems behave in line with human intentions and values.[7]

3.2 How do we know AI does what it is supposed to?

One of the hardest problems for AI applied to making decisions in complex social settings, like financial policy, is the specification of its objectives. Why should we trust its analysis and explanations?

A key challenge is that the training algorithms need to know what to optimise for. In other words, they need a well specified objective function that evaluates the cost and benefits of alternative courses of action given the current state of the environment and its future evolution. In most financial applications, they need to take into account how the system reacts to its actions. Misspecifying the problem, or what is likely to be more common, insufficient specification of the problem leads to suboptimal and even catastrophically bad decisions.

In the 1980s, an AI decision support engine called EURISKO used a cute trick to defeat all its human competitors in a naval wargame, sinking its slowest ships to maintain manoeuvrability. This early example of AI reward hacking, something humans are experts in, illustrates how difficult it is to trust AI. How do we know it will do the right thing? Human admirals don't have to be told not to sink their own ships, and if they do, they either have high-level political acquiescence or are stopped by their junior officers. Any current AI making autonomous decisions has to be told, or learn from observing human decisions, that sinking its own ships is not allowed. The problem is that the real world is far too complex for us to train AI on every eventuality. AI will predictably run into cases where it will make critical decisions in a way that no human would. EURISKO's creator, Douglas Lenat, notes that "[w]hat EURISKO found were not fundamental rules for fleet and ship design; rather, it uncovered anomalies, fortuitous interactions among rules, unrealistic loopholes that hadn't been foreseen" (Lenat, 1983, p 82). Each of EURISKO's three successive victories resulted in rule changes intended to prevent repetition. Still, in the end, the only thing that worked was telling Lenat that his and his AIs' presence was not welcome.

If we ask ChatGPT whether it is okay for admirals to sink their own ships, it gives a well-grounded, nuanced answer. However, this is a well known example, likely in the

[7]For a survey of AI alignment work, see, e.g., Ji et al. (2023), and Shevlane et al. (2023), and Bengio et al. (2023) model evaluation for extreme risks.

Electronic copy available at: https://ssrn.com/abstract=4604628

canon ChatGPT trained on. Would it have come up with the same answer if Lenat had not entered EURISKO in the naval board game? We don't know.

An AI engine can lead to suboptimal outcomes that are harder to detect than in the EURISKO case. Having control of some system and being given the objective of forecasting, it might attempt to manipulate the system's structure to meet the objectives as it sees them. This can have severe consequences for microprudential and macroprudential regulations, as discussed in Section 5 below.

That means it is essential that AI can explain its reasoning, just like humans are required to. That can be difficult since AI outputs are simply a nonlinear mapping of input data, where decisions are affected both by the network structure and trained parameters. The machine may say yes or no based on one particular configuration of a trillion parameters.

AI, which cannot explain its reasoning, cannot be used for many applications. Not surprisingly, considerable resources are being brought to bear on that problem, allowing the operators to trace particular recommendations to the input data, so instead of a recommendation being merely a function of model parameters, they are determined by particular inputs that are analysed logically. Benchmarking AI engines to specific tasks could be particularly valuable in financial applications, as we discuss in Section 6.

4 AI use in the financial system

Many public and private financial system applications can be supported by the two forms of AI accessible today, reactive and limited memory. And, as the engines' capabilities develop at a rapid pace, and we come closer to a theory of mind, and possibly AGI, we expect AI use to expand in both the public and private sectors (Maple et al., 2023; Kinywamaghana and Steffen, 2021; Korinek, 2023).

Credit allocation, fraud, compliance and risk management are just a few of the tasks in which reactive machines have proven highly valuable. Supervisors may use them to develop the rulebook and enforce compliance with it. There is no technical reason why AI cannot perform the majority of such jobs now; what is holding it back is a lack of human expertise, compute and data sharing (see Section 2.1) as well as caution.

For private financial institutions, where senior managers can be held to account, the unknown territory of AI, compared to relatively safe and familiar incumbent technology, might represent unacceptable personal risk. However, while legacy systems

Electronic copy available at: https://ssrn.com/abstract=4604628

might hold back traditional financial institutions and perhaps a culture that reflects traditional practices, challenger banks find it much easier to use technology. In a competitive market, their greater efficiency and better delivery of financial services will most likely push AI adoption across the private sector.

They may, however, prefer to outsource much AI analytics because of the high fixed costs of AI engines, as discussed in Section 5.4. Outsourcing to a vendor that spreads compute and human capital costs across multiple banks may be especially appealing to all save the largest banks.

The authorities have chosen a more cautious approach to AI. However, as the private sector develops its use of AI, the authorities must follow suit if they are to remain relevant. They are already doing so, as evidenced by Moufakkir (2023) and Cook (2023). Anecdotal evidence suggests that current AI use is mostly limited to regular forecasting and low-level data processing. However, most authorities are only now beginning a more structured analysis of AI.

4.1 Authorities taking advantage of AI

While financial authorities have so far used AI fairly sparingly, we expect that to change. This will necessitate investments in data, human capital and computing power. The most straightforward is data, as authorities already have access to large databases, including confidential disclosures, that can be used to train AI. This includes:

- Observations on past compliance and supervisory decisions
- Prices, trading volumes, and securities holdings in fixed-income, repo, derivatives and equity markets
- Assets and liabilities of commercial banks
- Network connections, like cross-institution exposures, including cross-border
- Textual data
  - The rulebook
  - Central bank speeches, policy decisions, staff analysis
  - Records of past crisis resolution
- Internal economic models

Electronic copy available at: https://ssrn.com/abstract=4604628

- Interest rate term structure models
- Market liquidity models
- Inflation, GDP and labour market forecasting models
- Equilibrium macro model for policy analysis

The authorities will find human resources and compute more difficult, not the least because of very high costs of human capital specialised in AI, as we discuss below in Section 5.4. Training AI engines is very expensive and necessitates access to specialised facilities, both of which are out of reach for most financial authorities. However, because many of the largest costs are associated with training general-purpose engines, it is easy to overstate those problems. The marginal cost to the authorities is likely to be significantly smaller, as they may be able to extend extant engines via transfer learning on curated datasets for financial sector use that include the items listed above.

There are many activities within the financial authorities where AI would be of considerable help. Start with the microprudential regulations, both design and supervision. ML can be used to translate the rulebook into a logic engine, allowing analysis of internal consistency and coverage and facilitating AI supervision. Several authorities have already made progress in this area. Generative models can simulate undesirable behaviour and hence help in designing the rulebook to protect against such outcomes. Reactive engines can identify undesirable behaviour not covered by extant rules and propose remedies.

AI can be beneficial in supervision, providing routine compliance advice to the private sector, recommending regulatory actions, and even making supervisory decisions. While human supervisors would initially closely oversee it, reinforcement learning with human feedback will help supervisory AI to become increasingly performant. For challenging areas of microprudential supervision, adversarial architectures such as generative adversarial networks (GANs) might be particularly beneficial in understanding complex areas of authority-private sector interactions, such as fraud detection.

It will also likely be useful in other activities, such as ordinary economic analysis and forecasting, thanks to an economic engine that performs both theoretical and statistical analysis. This could be achieved with a general purpose foundation model augmented via transfer learning using publicly and privately available data, established economic theory embodied in economic models and previous policy analysis. Reinforcement learning with feedback from human experts might be useful to improve the engine further. Such AI would be very beneficial to those conduct-

Electronic copy available at: https://ssrn.com/abstract=4604628

ing macroprudential stress tests. A key challenge in such analysis (Anderson et al., 2018) is that the stress needs to cover behaviour that we don't usually see in the system and capture particular interrelations between the various market participants in times of stress. Here, generative models could be valuable in facilitating the running of scenarios in crisis resolution, helping to identify and analyse the drivers of extreme market stress. The authorities could also use such models as artificial labs to experiment on policies and evaluate private sector algorithms, perhaps building on existing agent-based simulations. Over time, as AI technology improves, theory of mind AI might be able to understand and provide context for policy decisions and hence be invaluable in crisis resolution. AI that understands how its interlocutors think can explain its decisions to them, as well as reason about how they will react to its actions and plan accordingly, will be of considerable benefit. Not the least, it could address the problem of misspecified objectives.

4.2 The risk of authorities losing control

As the public and private sectors increase their use of AI, the authorities must carefully consider how they adopt AI. They might conclude that AI should only be used for advice, not decisions, with human beings in the loop to avoid undesirable outcomes. However, that might not be as big a distinction as the authority thinks. The AI engine will have its internal representation of the financial system, where its understanding might not be intelligible to its human operators. When we then use that AI to scan the system for vulnerabilities and run scenarios to evaluate the impact of alternative regulations or courses of action for crisis resolution, we might have no choice but to accept its advice, especially when presented as a choice between something that appears sensible and a potentially disastrous alternative. When optimising, the AI may even choose to present alternatives in that manner so as not to risk having the operator make inferior choices.

4.3 Trust

While an authority might not wish to get to that point, its use of AI might end up there regardless. As we come to trust AI analysis and decisions and see how well it performs in increasingly complex and essential tasks, it may end up by stealth, where the authorities do not want to be. That arises because of how AI creates trust. There are many examples where technology was initially met with scepticism. As it performs, technology is then increasingly accepted, as we have seen with AI

Electronic copy available at: https://ssrn.com/abstract=4604628

landing passenger aeroplanes, AI-guided surgery and self-driving cars. Even if we only allow AI to execute the simplest tasks inside the financial authorities, as it starts performing at the same level or better than its human counterparts, but at a much lower cost, we can expect AI to become increasingly welcomed. In other words, its very success creates trust. As that trust is earned on relatively simple and safe repetitive tasks and seeing AI succeed in more complex jobs, it will be asked to take on ever more sophisticated tasks.

As trust builds up, the critical risk is that we become so dependent on AI that the authorities cannot exercise control without it. In a sense, AI optimises to become irreplaceable. By then, turning the AI engine off may be impossible because it performs vital functions, and the risk of disastrous outcomes might be deemed unacceptably high. Eventually, we risk becoming dependent on a system for critical analysis and decisions we don't entirely, or even partially, understand.

4.4 Six criteria for AI use in financial policy

The conceptual challenges in Section 2 and the various issues surrounding AI take us to 6 criteria for evaluating AI use in financial policy.

1. **Data.** Does an AI engine have enough data for learning, or are other factors materially impacting AI advice and decisions that might not be available in a training dataset?

2. **Mutability.** Is there a fixed set of immutable rules the AI must obey, or does the regulator update the rules in response to events?

3. **Objectives.** Can AI be given clear objectives and monitor its actions in light of those objectives, or are they unclear?

4. **Authority.** Would a human functionary have the authority to make decisions, does it require committee approval, or is a fully distributed decision making process brought to bear on a problem?

5. **Responsibility.** Does private AI mean it is more difficult for the authorities to monitor misbehaviour and attribute responsibility in cases of abuse? In particular, can responsibility for damages be clearly assigned to humans?

6. **Consequences.** Are the consequences of mistakes small, large but manageable, or catastrophic?

Electronic copy available at: https://ssrn.com/abstract=4604628

5 The channels for how AI can destabilise the financial system

While we anticipate that AI will be broadly beneficial to the financial system, it may also increase systemic financial risk, defined as a serious financial crisis that has a high chance of causing an economic recession. The costs of systemic crises are very high, in the trillions of dollars for large economies, as noted by Barnichon et al. (2022). As such crises are usually caused by excessive low-quality lending, amplified by high leverage and asymmetric information, they should be easily preventable as macroprudential regulations aim to do. It is not so in reality. The almost infinite complexity of the financial system makes it very hard to design effective regulations that simultaneously prevent excesses and do not hamper economic growth too much. On top of that, because the economy is usually growing rapidly before a crisis, there is strong political opposition to any measures that might threaten that growth.

The reason why AI can increase systemic risk is, perhaps paradoxically, its very strength and efficiency. In the words of Bengio et al. (2023), "Compared to humans, AI systems can act faster, absorb more knowledge, and communicate at a far higher bandwidth. Additionally, they can be scaled to use immense computational resources and can be replicated by the millions." The financial system has a large number of potentially destabilising feedback mechanisms that are created by the interaction of actors and the significant resources that are trying to maximise profits in regular time and survival during stress. The combination of AI that excels at discovering optimal strategies, the system's complexity, highly resourced actors and the inherent destabilising factors already prevalent in the financial system viciously reinforce each other.

We build on extant work on AI safety, like Weidinger et al. (2022), Bengio et al. (2023) and Shevlane et al. (2023), which identifies several societal risks arising from AI use, including malicious use, misinformation and loss of human control. When augmenting that, we propose four channels for how AI use may destabilise the financial system, all of which follow from the AI issues discussed in Section 3 and the conceptual challenges in Section 2.

5.1 Malicious use of AI

The first channel arises from how the operators of financial AI use them to optimise against the system for malicious purposes. Financial institutions and their staff deploy considerable resources in their pursuit of profit and usually are not very con-

Electronic copy available at: https://ssrn.com/abstract=4604628

cerned about the wider social impact of their activities. They can change the system in ways that benefit them while not being detectable by others, as we have seen many times in the past. AI provides such actors with new opportunities, either via adversarial attacks by feeding particular data into its training algorithms or influencing network design. That might not even be necessary since it is straightforward to exploit for private, even illicit, gain gaps in how AI sees its responsibilities.

One way such malicious use might occur is by manipulating governance processes. The financial authorities and internal control aim to align the interests of system participants to the organisation that employs them and society. As AI is particularly effective in finding profitable loopholes and amplifying vulnerabilities, it can facilitate misbehaviour that, while legal, is damaging both to society and even the institution employing the AI.

The operator of the AI could take that one step further and use it to deliberately create market stress, which, of course, is profitable to those forewarned. We have seen many examples of such conduct in the past, and a key purpose of securities and macroprudential regulations is to prevent such behaviour.

And finally, we have those aiming to manipulate the system for illegal purposes, like rogue traders, criminals, terrorists and nation-states. The inherent vulnerability of the financial system to adversarial attacks and the ease of manipulating AI engines create potential for particular attack vectors that can be impossible to identify ex-ante. One example comes from Hubinger et al. (2024) who identify what they call "sleeper agent" that usually acts as expected but gives deceptive answers when given special instructions. Here, AI engine corruption becomes another form of cyber attack. With such an attack vector in place, an AI engine becomes a tool for well-resourced hostile agents that might be difficult or impossible to prevent in real time, facilitating nation-state attacks on critical infrastructure. That is particularly relevant when a superior ability to identify optimal timing strategies allows them to solve the problem of double coincidence, as discussed in Daníelsson et al. (2016), creating heightened system fragility, perhaps by manufacturing a liquidity crisis as an attack amplifier.

5.2 Misinformed use and overreliance on AI

The second channel emerges from humans using AI incorrectly. That is particularly relevant when AI is used in domains with unclear objectives, which is very common in the financial system. As events become less frequent and more serious, data-driven decisions increasingly involve extrapolation, interpretation, intuition and nuance,

Electronic copy available at: https://ssrn.com/abstract=4604628

none of which AI in its current form is good at. As we discuss on Page 12, many crises have been resolved by very creative solutions that had never been used before.

Because successful AI architectures are data-driven engines, their analysis is predominantly founded on statistical regularities. Such AI engines are designed to provide advice and responses to prompts, even if they have very low confidence about the accuracy of the answer. They can even make up facts or present arguments that sound plausible but would be considered flawed by an expert, both instances of the broader phenomenon of AI hallucination. The risk is that the AI engines will present confident recommendations about outcomes they know little about. To overcome that, the engines should be required to provide an assessment of the statistical accuracy of their recommendations. Here, it will be helpful if the authorities overcome their frequent reluctance to adopt consistent quantitative frameworks for measuring and reporting on the statistical accuracy of their data-based inputs and outputs.

5.3 AI misalignment and control avoidance

The third channel for how AI can be destabilising relates to difficulties in aligning the behaviour of AI engines with the objectives of both their owners and financial authorities. The latter is particularly challenging if the objectives of financial regulations cannot be defined precisely, as usually is the case, as we discussed in Section 2.4. Problems of misalignment arise when AI makes decisions because of the impossibility of fully specifying all objectives it has to meet, a form of incomplete contracting. The best solution an engine comes up with might be undesirable for the institution it works for. While such individual misbehaviour is certainly all too common in today's human-centred financial institutions, it becomes more prevalent as AI use increases. Humans have been taught ethics and have general notions such as "don't cheat, steal, no violence." Most organisations try to recruit individuals for whom that education was successful. AI might require a similar broad education if its objectives cannot be fully specified ex-ante.

Scheurer et al. (2023) provides an example of how individual AI can spontaneously choose to break regulations in their pursuit of profit. Using GPT-4 to analyse stock trading, they told their AI engine that insider trading was unacceptable. When they then gave the engine an illegal stock tip, it proceeded to trade on it and lie to the human overseers. Here, AI is simply engaging in the same type of illegal behaviour so many humans have done before.

The problem of misalignment is particularly difficult when several AI engines interact. For example, market manipulation is most successful when several investors

Electronic copy available at: https://ssrn.com/abstract=4604628

collude. Similarly, the payoff of many trading activities, such as carry trades and short sales, increases with the number of investors participating in them. A particularly damaging consequence of the strategic complementarities prevalent in the financial system relates to the objectives of financial institutions. They maximise profit most of the time, perhaps 999 days out of a thousand, but optimise for survival during stress. The behaviour on that one day out of a thousand is particularly damaging when financial institutions seek safety by withdrawing liquidity, which can lead to destabilising dynamics such as fire sales, bank runs, and credit crunches. The very high level of AI performance can, perhaps paradoxically, increase the likelihood of damaging coordinated behaviour. They might find better strategies for exploiting strategic complementarities for private gain, at the risk of lowering market quality or even threatening financial stability.

An example of AI collusion is Calvano et al. (2020), who find that independent reinforcement learning algorithms instructed to maximise profits quickly converge on collusive pricing strategies that sustain anti-competitive outcomes. It is much easier for AI to behave in this collusive way than humans, as such behaviour is very complex and often illegal. AI is much better at handling complexity and is unaware of the legal nuances unless explicitly taught or instructed.

The most difficult problems of misalignment arise due to the interaction of the various AI working throughout the industry. Long before computers got involved with the financial system, humans coordinated damaging behaviour due to the misalignment of individual short-term incentives with the long-term incentives of the same market participants and society at large. The crises of 1763, 1914 and 1929 (see, e.g. Daníelsson, 2022) are just two of many such examples, as is the tulip mania in the Netherlands in the 1630s, see Aliber and Kindleberger (2015). The use of computers amplifies the potential for such undesirable outcomes as two recent examples illustrate. The largest stock market crash in history happened in 1987 when global markets fell by 23% in one day. The reason was trading algorithms that, in a way not foreseen, coordinated in creating vicious feedback between prices collapsing and algorithm-induced selling, as documented by Gennotte and Leland (1990). Similarly, the quant crisis in the summer of 2007, as noted by Khandani and Lo (2007), was due to trading algorithms trained to sell on falling prices but not being told of the potential for aggregate feedback caused by trading algorithms in a number of funds coordinating on selling. Such factors have further been at the root of the various flash crashes. AI will likely amplify such problems of misalignment.

It is tough for the authorities to patrol a nearly infinitely complex system. They have limited resources and can only look at specific system parts. Once they are focused on particular behaviour, that risk might be contained, but what tends to happen is

Electronic copy available at: https://ssrn.com/abstract=4604628

that the forces of instability emerge elsewhere. Risk is a bit like a balloon; squeeze it in one area, and it expands elsewhere. Indeed, it is almost axiomatic that instability arises where the authorities are not looking. And, in an almost infinitely complex system, there is plenty of scope for instability.

AI will help the authorities keep the system stable and aid the forces of instability. The question is, which of these two dominate? Instability, as it needs only to find one weakness, while the authorities must monitor the entire system. AI will be particularly good at identifying the weak points.

Meanwhile, the supervisor must not only identify the weak points but also monitor how financial institutions interact with each weakness in real time. A much more difficult computational task, and one where the more complex the system becomes, the harder the task is. Furthermore, the supervisors have fewer resources than the private sector. The regulator might even have to rely on private sector resources (compute, fitted models, data). The resulting power asymmetry may be impossibly hard to solve.

That means AI, either autonomously or at human instruction, would likely be particularly good at finding strategies for avoiding both internal controls in a financial institution and regulatory oversight, making the two-sided principal-agent problem we discuss in Section 2.5 particularly challenging. In other words, the more sophisticated AI become, the more difficult it will be to exercise authority over them. After all, the controllers need to understand the AI objectives to design the correct incentives. And given the current state of interpretability of foundation models, this is very challenging.

5.4 Resource concentration

Monoculture is an important driver of booms and busts in the financial system. As financial institutions come to see and react to the world in increasingly similar ways, the more they coordinate in buying and selling, leading to bubbles and crashes or, even worse, credit booms and credit crunches. Banks' capital calculation is a particular driver of procyclicality, not the least via risk weights. Such procyclical behaviour has always been a feature of financial markets, long before the use of computer technology. Still, it is amplified by technology.

AI will likely exacerbate this already strong channel for financial instability. It excels in finding the best methods for measuring risk by using generative models that combine a scientific understanding of the nature of risk with stochastic processes of financial assets. While that might lead to the best possible measurement of risk, it

Electronic copy available at: https://ssrn.com/abstract=4604628

also has the unfortunate consequence of driving market participants to see risk in the same way — beliefs. AI then will find the best practices for managing risk — action. The consequent harmonisation of beliefs and action is strongly procyclical, as financial institutions react to shocks similarly, excessively amplifying buying and lending, inflating bubbles, and eventually selling rapidly and withdrawing credit.

The oligopolistic nature of the AI analytics business further strengthens that efficiency-induced procyclicality. ML design, input data and compute affect AI engines' ability. These are controlled mainly by a few technology and information companies, which continue to merge, creating an increasingly oligopolistic market. The median salary for specialists in data, analytics and artificial intelligence in US banks was $901,000 in 2022 and $676,000 in Europe.[8] Since there is a considerable shortage of the necessary human capital and the productivity of those experts is directly affected by the network effects of working with other experts, as well as the data and compute available to them, this alone puts designing the most effective AI engines out of the reach of all but the largest financial institutions. Furthermore, the currently largest neural networks have trillions of parameters, GPT 4 with 1.7 trillion, requiring significant funding and access to specialised data centres with the requisite GPUs. The current top engines have hundreds of millions of dollars in annual budgets. OpenAI reported that the cost of training GPT-4 exceeded $100 million.[9]

Finally, financial data vendors have concentrated considerably over the past few years, with only a handful left, such as S&P Global, Bloomberg and LSEG. They have unified databases with standardised labelling and graph structures linking disparate categories of data, including proprietary and not publicly available data, facilitating training. It is a concern that neither the competition nor the financial authorities appear to have fully appreciated the potential for increased systemic risk due to oligopolistic AI technology in the recent wave of data vendor mergers.

Taken together, AI-driven financial analytics is characterised by increasing returns to scale technologies with high entry costs, making an oligopolistic market structure highly likely. In such an environment, outsourcing analytics to one of those vendors might be the best, maybe even the only feasible solution for all but the largest financial companies. Anecdotal private sector evidence indicates that many financial institutions already prefer to outsource data-driven applications to specialists. A particular example is risk management as a service, RMaaS, with a leading vendor being BlackRock's AI-fuelled Aladdin. The risk is harmonised beliefs and actions,

[8]https://www.bloomberg.com/news/articles/2023-11-28/goldman-raided-by-recruiters-in-wall-street-fight-for-ai-talent

[9]https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over

Electronic copy available at: https://ssrn.com/abstract=4604628

driving procyclical behaviour and exposing users to the same blind spots.

6 Regulating AI

The financial system is one of the most regulated parts of the economy because of the strong externalities arising from financial activities. A particular challenge for macropru is that keeping individual parts or institutions of the system safe is not equivalent to keeping the system as a whole safe. In fact, these two objectives often come into conflict, and regulation needs to take a system perspective with all the complexity that ensues.

6.1 Regulation of private AI

As the private sector increasingly adopts AI, the supervisors will have to address both how AI changes the behaviour of the private sector institutions and also how to regulate the AI engines in use. Several issues emerge.

Both the public and private sectors will likely end up outsourcing analytics to the same handful of very large AI vendors. This blurs the regulator/regulated divide since if the private sector is acting on the same advice as the public sector, the nature of supervision will be different than today. The reason is that a vendor's engine will likely have a single representation of the financial system, that is, of the stochastic processes that drive market outcomes. This means that advice provided to all users of the engine, in the private and public sectors, will share the same underlying view of the system. That has two consequences. The first is that it will harmonise beliefs, which is pro-cyclical. Furthermore, since the engine is just a model that maps a near infinitely complex system to a simplified version of it, if a private firm unknowingly engages in undesirable actions at the recommendation of the engine, the authority using the same engine might not identify that misconduct. To guard against both issues, the authorities should pay particular attention to problems of model diversity and robustness, aim to use different engines than regulated entities and, seek to have several vendors of AI analytics and take advice from several vendors on important questions of system stability.

The authorities may also need to contend with new dimensions of existing regulations on the physical location of compute facilities and data sovereignty, as the AI analytics vendors likely will not be from the same jurisdiction as the supervisor or the private sector institution, a problem already common in cloud computing. One

Electronic copy available at: https://ssrn.com/abstract=4604628

way we approach this today is to require facilities to be located within the supervisor's jurisdiction, with data not allowed to migrate outside of the jurisdiction. When it comes to AI, that distinction might not be as relevant as it appears. The design of the engine, and even the training, may be done elsewhere, and the same would apply to the intellectual property. The risk, then, is that a jurisdiction will have less control of a key component of market infrastructure than under the current setup.

While regulations can prevent that, it would then only come at the expense of lower performance.

Increased use of private sector AI can result in a new layer of deniability. Assigning legal responsibility for misconduct in the financial sector is already tricky. It will be harder when AI is used to make decisions. Suppose a human operator deliberately instructs AI to break the law for criminal or terrorist purposes or just turns a blind eye to the AI doing so as a byproduct of maximising profits. Even if detected, it might be easily explainable as an unintended and unexpected innocent behaviour. For example, while a rogue trader can be prosecuted, we cannot do that with a rogue AI trader. The institution employing such AI might profit significantly. Consequently, the increased use of AI in the private sector facilitates the job of those seeking to utilise AI for nefarious purposes by providing them with yet another level of denial, at least until the law and regulations catch up.

As the authorities ease into regulating a financial system that uses AI, a helpful approach might be to evaluate the safety of private AI against defined "benchmark tasks" that are well-understood and clearly defined everyday regulatory activities. While that might give rise to the private sector optimising against those benchmark tasks, constructive ambiguity in their design would likely significantly hamper such efforts.

Bengio et al. (2023) and Shevlane et al. (2023) in their analysis of the use of AI, propose avenues for benchmarking, such as checking for engine consistency, that is, whether standard inputs yield reasonable outputs, and that similar inputs yield equally similar outputs. The authorities could then execute regular scans for engines having unexpected capabilities and test them to see if they can engage in manipulative behaviour and can be transparent, via explainability, in how they reach conclusions. Ultimately, transparency, for example, by publishing sufficient details on architecture and training or even using open-source approaches, will allow outside scrutiny.

Electronic copy available at: https://ssrn.com/abstract=4604628

6.2 Regulation of public AI

The financial authorities act as agents for society and have to be seen as accountable to society. We have several safeguards built to achieve that, including judicial and parliamentary scrutiny. We have governance structures in place. As the authorities increasingly start using AI, new areas of accountability are raised.

We have several ways today to provide that accountability. The human supervisor who makes mistakes can be held accountable, trained, dismissed, or even prosecuted, as we have seen many times in the past. If AI is tasked with making the same type of supervisory decisions, how does one attribute accountability? While the supervisory structure has many individuals, where perhaps only one makes a mistake, the authority will use only one AI engine to make all supervisory decisions. That means mistakes become systemic, not idiosyncratic. However, the engine can't be turned off if it is already in use, and retraining it won't be easy. The authority will need to have a response plan in place for such eventualities.

When a human supervisor makes a mistake, or the regulated thinks a human supervisor made a mistake, the regulated can appeal and challenge the ruling in a court of law. That means a human supervisor will be called on to explain the logic behind a decision, and the authority will explain the guidelines given to the human supervisor. That will be more difficult when AI makes decisions. There will be no separation between individual decisions and the guidelines, and the regulatory AI may not be able to explain its reasoning or why it thinks it complies with laws and regulations.

As an engine becomes embedded, we get to the point where we can not turn the engine off because that might lead to an entire essential function in the supervisory apparatus left unaddressed. Ultimately, this means that the internal supervision of AI use in the regulatory agencies and its interaction with the legal system will require different policies than those used for current human supervisors.

A useful way to evaluate AI use in the public sector would be to test them for safety and performance on defined benchmark tasks, particularly activities the authority engages in. It could be asked to advise and rule on macro prudential actions and suggest how to respond to different hypothetical crisis scenarios.

7 Conclusion

In this work, we have identified the main criteria for evaluating the pros and cons of AI use in financial authorities and the conceptual problems that may arise. Many of

Electronic copy available at: https://ssrn.com/abstract=4604628

the issues facing AI also affect human decision-making. AI will perform much better than human decision-makers in many routine tasks, such as risk management and compliance, while also threatening the stability of the system.

AI excels and outperforms humans in tasks such as compliance, standard supervision and risk management because there is plenty of data to train on, the objectives AI has to meet are clear and immutable over the timescale it operates and the cost of mistakes is contained and easily addressed.

Several factors frustrate the use of AI for macropru, and even worse, can cause it to misdirect policymakers and even destabilise the financial system. Data are limited and can be misleading as the financial system undergoes continuous structural change. Monitoring the system vulnerabilities and controlling risks is difficult because the drivers of instability only emerge in crisis times. Economic actors endogenously amplify stress and change their behaviour in response to regulatory attempts of control.

The literature on AI has identified several areas where it can threaten society, and we augment those with issues arising in the financial system. We find five channels: malicious use of AI, misinformed use and overreliance on AI, AI misalignment, AI as an increasing return to scale technology, and AI evading control.

Ultimately, the usefulness of AI for the financial authorities depends on what we want from it. The following table shows how AI affects the various tasks performed by the authorities.

Electronic copy available at: https://ssrn.com/abstract=4604628

Table 1: Particular regulatory tasks and AI consequences

Task	Data	Mutability	Objectives	Authority	Responsibility	Consequences
Fraud/Compliance Consumer protection	Ample	Very low	Clear	Single	Mostly clear	Small
Microprudential risk management Routine forecasting	Ample	Very low	Mostly clear	Single	Clear	Moderate
Criminality Terrorism	Limited	Very low	Mostly clear	Multiple	Moderate	Moderate
Nation state atttacks	Limited	Full	Complex	Multiple & international	Moderate	Very severe
Resolution of small bank failure	Limited	Partial	Clear	Mostly single	Mostly clear	Moderate
Resolution of large bank failure Severe market turmoil	Rare	Full	Complex	Multiple	Often unclear	Severe
Global systemic crises	Very rare or not available	Full	Complex & conflicting	Multiple & international	Unclear even ex-post	Very severe

Electronic copy available at: https://ssrn.com/abstract=4604628

A Machine Learning (ML)

There are many excellent sources on the technical details of ML, see for example Hastie et al. (2009) and Murphy (2023). The ML algorithms that underpin the current generation of successful AI engines are based on neural networks and involve inferring a complex mapping from some input data to an output.[10] For example, for large language models (LLMs), the input is a sequence of what is called tokens, perhaps words, images or sounds, while the output is the next token in the sequence, such as a word in a sentence.[11]

Once a sequence predictor has been trained, it can be used to generate new data — *generative AI* — via the built-in auto regression: given an initial input sequence of tokens and a prompt (question), the model predicts the next token for the sequence. This predicted token is then added to the original sequence and fed back into the model to predict yet another token, and so on. The output of this iterative process can include paragraphs of text that provides an answer to a question or a prediction for the future evolution of a set of financial asset prices given an initial market configuration. Other generative network structures, such as *variational autoencoders* and *diffusion models*, infer the latent statistical structure of a dataset and then sample from the inferred distribution to generate new data instances, for example artificial human faces given a database of portrait pictures. *Generative adversarial networks* (GAN) generate new instance by training a *generator* of artificial data to fool a *discriminator* to classify the generated data as real.[12] Obvious financial applications of such generative models include the simulation of financial market scenarios, realizations of prices and quantities for a set of financial assets, perhaps to evaluate the profitability or risk of trading algorithms or to assess the impact of regulatory policies.

A particular variant, transfer learning, might be especially useful for the financial authorities. These are networks first trained with a general body of information and then subsequently fine tuned with specialised datasets relevant to particular

[10]The algorithms look for a function f that "best" maps input x to output y, where f involve multiple nested layers, transforming the input data to extracting complex patterns. The techniques used to fit or train these types of models are referred to as deep learning. See Prince (2023) for an up-to-date overview of deep learning techniques.

[11]Currently, the most successful architecture for foundation models is the transformer (Vaswani et al., 2023) used in OpenAI's GPT models, Google's PaLM and Meta's LLaMA. While the transformer architecture dominates the current AI landscape, some alternative architectures for foundation models have recently emerged, e.g., state space sequence models (SSMs) (Gu et al., 2022) that are computational efficient and can model long-term dependencies in time-series data.

[12]For a detailed treatment of these generative models, see chapters 15-18 in Prince (2023).

Electronic copy available at: https://ssrn.com/abstract=4604628

applications. The fine-tuning of such a network can be based on specialised financial datasets or text data such as economics textbooks, academic papers, policy research and the rulebook.

If an AI is to act autonomously, it needs an understanding of its environment, a model of the world that includes how its actions impact the environment, and it also needs to know what objective its actions are to achieve. Finding optimal solutions means searching through the space of possibilities. A learning agent has to trade off exploiting actions that have proved successful in the past and exploring new and potentially superior actions. Reinforcement learning, (Sutton and Barto, 2018), is the most commonly used algorithm to learn the state-contingent relationship between actions and the ensuing payoffs. It uses ideas from dynamic programming to learn a policy function that gives optimal actions for given environmental states. Misspecifying an AI's objective can lead to undesired outcomes such as reward hacking. As an alternative to a hard-coded objective function, the AI can also be positively or negatively reinforced by human feedback, which for financial applications could come from economic experts such as experienced traders or senior central bankers with crisis experience.

Electronic copy available at: https://ssrn.com/abstract=4604628

References

Aliber, R. Z. and C. P. Kindleberger (2015). *Manias, Panics, and Crashes: A History of Financial Crises* (7 ed.).

Anderson, R. W., J. Daníelsson, C. Baba, U. Das, M. S. Basurto, and H. Kang (2018). Macroprudential stress tests and policies: Searching for robust and implementable frameworks.

Bagehot, W. (1873). *Lombard Street*. London: H.S. King.

Barnichon, R., C. M. C, and A. Ziegenbein (2022). Are the effects of financial market disruptions big or small? *Review of Economics and Statistics*, 557–70.

Bengio, Y., G. Hinton, A. Yao, D. Song, P. Abbeel, Y. N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, et al. (2023). Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*.

Calvano, E., G. Calzolari, V. Denicolo, and S. Pastorello (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review 110*(10), 3267–97.

Cook, L. D. (2023). Generative AI, productivity, the labor market, and choice behavior.

Daníelsson, J. (2022). *Illusion of Control*. Yale University Press.

Daníelsson, J., M. Fouche, and R. Macrae (2016). Cyber risk as systemic risk. *VoxEU.org*.

Daníelsson, J., R. Macrae, and A. Uthemann (2022). Artificial intelligence and systemic risk. *Journal of Banking and Finance 140*.

Daníelsson, J. and H. S. Shin (2002). Endogenous risk. In *Modern Risk Management — A History*. Risk Books. `www.RiskResearch.org`.

Dewatripont, M., J. Tirole, et al. (1994). *The prudential regulation of banks*, Volume 6. MIT press Cambridge, MA.

Gennotte, G. and H. Leland (1990). Market liquidity, hedging, and crashes. *American Economic Review*, 999–1021.

Goodhart, C. A. E. (1974). Public lecture at the Reserve Bank of Australia.

Electronic copy available at: https://ssrn.com/abstract=4604628

Gu, A., K. Goel, and C. Ré (2022). Efficiently modeling long sequences with structured state spaces.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*, Volume 2. Springer.

Hubinger, E., C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, A. Jermyn, A. Askell, A. Radhakrishnan, C. Anil, D. Duvenaud, D. Ganguli, F. Barez, J. Clark, K. Ndousse, K. Sachan, M. Sellitto, M. Sharma, N. DasSarma, R. Grosse, S. Kravec, Y. Bai, Z. Witten, M. Favaro, J. Brauner, H. Karnofsky, P. Christiano, S. R. Bowman, L. Graham, J. Kaplan, S. Mindermann, R. Greenblatt, B. Shlegeris, N. Schiefer, and E. Perez (2024). Sleeper agents: Training deceptive llms that persist through safety training.

Ji, J., T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, F. Zeng, K. Y. Ng, J. Dai, X. Pan, A. O'Gara, Y. Lei, H. Xu, B. Tse, J. Fu, S. McAleer, Y. Yang, Y. Wang, S.-C. Zhu, Y. Guo, and W. Gao (2023). AI alignment: A comprehensive survey.

Kasirzadeh, A. Two types of AI existential risk: Decisive and accumulative.

Khandani, A. E. and A. W. Lo (2007). What happened to the quants? in August 2007?

Kiarelly, D., G. de Araujo, S. Doerr, L. Gambacorta, and B. Tissot (2024). Artificial intelligence in central banking. Technical report, BIS.

Kinywamaghana, A. and S. Steffen (2021). A note on the use of machine learning in central banking.

Knight, F. (1921). *Risk, Uncertainty and Profit*. Houghton Mifflin.

Korinek, A. (2023, December). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature 61*(4), 1281–1317.

Laeven, L. and F. Valencia (2018). Systemic banking crises revisited. *IMF Working Paper No. 18/206*.

Legg, S. and M. Hutter (2007). Universal intelligence: A definition of machine intelligence.

Lenat, D. B. (1983). EURISKO: a program that learns new heuristics and domain concepts: the nature of heuristics iii: program design and results. *Artificial Intelligence 21*(1-2), 61–98.

Electronic copy available at: https://ssrn.com/abstract=4604628

Lowenstein, R. (2000). *When Genius Failed – The Rise and Fall of Long-Term Capital Management*. Random House.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, Volume 1, pp. 19–46. North-Holland.

Maple, C., L. Szpruch, G. Epiphaniou, K. Staykova, S. Singh, W. Penwarden, Y. Wen, Z. Wang, J. Hariharan, and P. Avramovic (2023). The ai revolution: Opportunities and challenges for the finance sector. Technical report, Turing instutite.

Moufakkir, M. (2023). Careful embrace: AI and the ECB. Technical report, European Central Bank. https://www.ecb.europa.eu/press/blog/date/2023/html/ecb.blog230928 3f76d57cce.en.html.

Murphy, K. P. (2023). *Probabilistic Machine Learning: Advanced Topics*. MIT Press.

Norvig, P. and S. Russell (2021). *Artificial Intelligence: A Modern Approach:*. Pearson.

Pistor, K. (2013). A legal theory of finance. *Comparative Journal of Economics*.

Prince, S. J. (2023). *Understanding Deep Learning*. MIT Press.

Rasche, R. H. and H. T. Shapiro (1968). The frb-mit econometric model: its special features. *The American Economic Review 58*(2), 123–149.

Russel, S. (2019). *Human compatible*. Allen Lane.

Scheurer, J., M. Balesni, and M. Hobbhahn (2023). Technical report: Large language models can strategically deceive their users when put under pressure.

Shevlane, T., S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, et al. (2023). Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement learning: An introduction*. MIT Press.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2023). Attention is all you need.

Electronic copy available at: https://ssrn.com/abstract=4604628

Weidinger, L., J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, et al. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229.

Electronic copy available at: https://ssrn.com/abstract=4604628