THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Economic and Social Research Council

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

Jon Danielsson
Andreas Uthemann

SRC Discussion Paper No 125
November 2023

Systemic Risk Centre

Discussion Paper Series

ISSN 2054-538X

Abstract
As the financial authorities increase their use of artificial intelligence (AI), micro regulations, such as consumer protection and routine banking regulations, will benefit because of ample data, short time horizons, clear objectives, and repeated decisions, leaving plenty of data for AI to train on. It is different from macro, focused on the stability of the entire financial system, where AI can potentially undermine financial stability. Infrequent and mostly unique events frustrate AI learning and hence its use for macro regulations. Using distributed decision making, humans retain the advantage over AI for advising on and making decisions in times of extreme stress. 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, robustness and accuracy. We propose six criteria against which to judge the suitability of AI use by the private sector and financial regulation and crisis resolution.

JEL classification: G01,G28.

Keywords: Artificial Intelligence, Systemic Risk, Financial Regulations, Central Banks.

This paper is published as part of the Systemic Risk Centre's Discussion Paper Series. The support of the Economic and Social Research Council (ESRC) in funding the SRC is gratefully acknowledged [grant number ES/R009724/1].

Jón Danielsson, Department of Finance and Systemic Risk Centre, London School of Economics

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

Published by
Systemic Risk Centre
The London School of Economics and Political Science
Houghton Street
London WC2A 2AE

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

October 2023

Abstract

As the financial authorities increase their use of artificial intelligence (AI), micro regulations, such as consumer protection and routine banking regulations, will benefit because of ample data, short time horizons, clear objectives, and repeated decisions, leaving plenty of data for AI to train on. It is different from macro, focused on the stability of the entire financial system, where AI can potentially undermine financial stability. Infrequent and mostly unique events frustrate AI learning and hence its use for macro regulations. Using distributed decision making, humans retain the advantage over AI for advising on and making decisions in times of extreme stress. 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, robustness and accuracy. We propose six criteria against which to judge the suitability of AI use by the private sector and financial regulation and crisis resolution.

*Corresponding author Jon Danielsson, J.Danielsson@lse.ac.uk. We thank Charles Goodhart, 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.

1 Introduction

Artificial intelligence (AI) will transform how the financial system is governed, promising improved efficiency, robustness and impartiality at much lower costs than existing arrangements. AI also threatens financial stability because of how fixed objectives, unknown unknowns and strategic interactions affect it.

The usefulness of AI for any task, as noted by Russel (2019), depends on the structure of the task at hand, and it is helpful to think of its application to financial policy on a spectrum. On one end, we have a problem with ample data and fixed, immutable rules, where both play to the strength of AI and its learning algorithms. As the frequency of events drops and rules become mutable, the AI's advantage erodes until, on the other end, we reach once-in-a-working-lifetime events, such as having to resolve a systemic financial crisis.

We propose six questions to ask when evaluating the use of AI for regulations and crisis resolution purposes. Table 1 on page 19 shows how each relates to particular activities, ranging from routine risk management to systemic crises.

1. Does the AI engine have enough data?
2. Are the rules immutable?
3. Can AI be given clear objectives?
4. Does the authority the AI works for make decisions on its own?
5. Can we attribute responsibility for misbehaviour and mistakes?
6. Are the consequences of mistakes catastrophic?

There is no uniform notion of what AI is. Here, we adopt a common definition that AI is a computer algorithm that autonomously engages in goal-directed behaviour that humans would normally perform. It acts to achieve the best expected outcome given its objectives and its understanding of the problem domain — the rational agent approach according to the taxonomy of Norvig and Russell (2010) — which resonates with economic analysis. Here, "acts" can refer to advising human decision-makers on achieving an objective, i.e. "action A is preferable to action B" or making decisions independently, such as using self-driving cars or trading algorithms. The AI's objectives can be hard-coded by its human owners or learned from human feedback in the AI's training phase. As AI studies how human supervisors and crisis managers make decisions, it will develop an increasingly higher understanding of the objectives.

The authorities need to respond to AI, whether or not they want to, and most are formulating AI policy (see, e.g. Moufakkir, 2023). AI use is mainly concentrated today in areas such as AI-guided risk management and advice from large language models (LLMs) like ChatGPT, but it will probably grow rapidly. Along the way, AI will affect most functions performed by financial authorities, such as designing regulations, monitoring the system for the violation of rules, making supervisory decisions and advising on crisis resolution. Even if an authority does not want to use AI for decision making, that might happen regardless; AI provides critical analysis, so rejecting its conclusions might not be acceptable. Whereas some authorities might prefer a slow, deliberative and conservative approach, that will not be tenable for three reasons.

First, the private sector is significantly expanding its use of AI in applications such as risk management, internal control, credit allocation and fraud minimisation, and an authority that persists in regulating such activities with traditional methods may find itself outmanoeuvred. Because private sector AI will optimise differently and act and react faster than humans, it is likely to generate more data, and of a different kind, so analysing outcomes requires matching AI technologies. It seems inevitable that the authorities will have to keep up with the use of AI in the private sector to remain effective. That extends all the way to resolving the most severe financial crisis. The increased use of private sector AI amplifies the complexity of the financial system, so if the authorities are to understand the market situation, they will have to make extensive use of AI for analysis, advice and even decisions.

Second, a policy of not using AI for high-level decisions will probably be undermined by the stealthy adoption of AI. Even if AI is explicitly prevented from making important decisions, authority staff will probably adopt state-of-the-art techniques from academia and the private sector to solve their tasks. This will make AI highly influential. An authority could prevent such use of AI, but that would come at the cost of much inferior analysis.

Finally, there may be no feasible alternative to using AI for advising on essential tasks. Perhaps AI in a severe liquidity crisis is the only way to bring together all the disparate data sources and to identify all the various connections between the market participants necessary for providing the best advice to the leadership.

There may not be much difference between AI making decisions and AI providing crucial advice, particularly if AI's internal representation of the system is no longer intelligible to its human operators. What are the alternatives to accepting AI's advice if AI automatically does all the monitoring and analysis? We trust it to make small decisions, and as the system performs well, we expand its reach. Eventually, we come to rely on it. The risk is that we become overly dependent on a system we do not fully understand. This, of course, is not unique to the financial system; the demands for explainable AI are very strong, with significant resources being brought to bear. It could take a long time to reach satisfactory levels of AI explainability, especially when it is used for crisis resolution.

A financial authority entrusted with regulating the financial system has two objectives. The first is micro-regulation, which is concerned with day-to-day issues such as risk management, consumer protection and fraud. AI will be of considerable benefit to the micro-regulation authorities.

The second is macro-regulation, and here the picture is less clear. Macro-regulation focuses on broad-picture issues such as financial stability, as noted by Danielsson, Macrae and Uthemann (2022). The emphasis is decidedly long run, both to avoid systemic financial crises and large losses over years and decades into the future, and to resolve such crises if they occur. Macro-policies are much more difficult to execute and less accurate than micro-policies.

The usefulness of AI for the financial authorities tasked with micro- and macro-supervision and resolving stress events and crises is directly affected by four conceptual challenges: data, unknown unknowns, strategic interactions and the mutability of objectives.

The first, perhaps paradoxically, is data. After all, the financial system may appear to be the ideal use case for AI because it creates almost infinite amounts of data, leaving plenty for AI to train on. However, such data is often inconsistently and even inaccurately measured. Financial system data is collected by authority silos where data sharing, within an authority, between authorities in the same jurisdictions and across jurisdictions, is limited. Data for stress mitigation is rare because crises are infrequent. The Laeven and Valencia (2018) database finds that the typical OECD country only suffers a systemic financial crisis once in 43 years, suggesting that most senior decision-makers make important crisis interventions only once in their professional career. We usually don't know what data is relevant until after the stress event has occurred. This means that AI is at the risk of inferring an incorrect causal structure of the financial system.

The second challenge arises from the uniqueness of crises. Major financial crises share many common fundamental features: high initial leverage and an unexpected event that undermines confidence, followed by a liquidity dry-up as investors run for safety in increasingly opaque markets. However, every crisis is unique because the driving factors are specific to the institutional structure at the time and the particular regulations and political regimes in place. To learn from past crises, it is essential to map the unique details onto the common fundamentals and vulnerabilities. That requires a high level of understanding of areas such as politics, institutional structure and law. Whereas the authorities can search for the vulnerabilities that lead to financial crises, their job is frustrated by the effectively infinitely complex nature of the financial system, which allows them to only patrol a small part of it. Rarity and uniqueness in a sparsely monitored system imply that systemic crises can be seen as unknown unknowns or "uncertain" in Frank Knight's (1921) classification. Consequently, it can be risky to outsource analysis to AI because it does not fully understand all the important factors that affect both the likelihood of crises arising and their dynamics once they happen.

The third conceptual challenge facing AI relates to how the financial system responds to control, echoing Goodhart's law and the Lucas critique (Lucas, 1976). The complex feedback between regulations and responses to regulations frustrates the monitoring and control of risks in the system. Daníelsson and Shin (2002) classify financial risk as either exogenous or endogenous. Exogenous risk assumes that risk drivers arise outside the financial system and are not influenced by it. Most risk forecasting and micro-regulations assume that risk is exogenous, easy to analyse and usually not of serious concern. Endogenous risk recognises that the interaction of economic agents not only drives outcomes but also changes the structure of the financial system. Macro-risk is almost always endogenous, and any data-driven process can mislead AI because historical data might not be informative about ongoing endogenous risk dynamics.

The last conceptual problem stems from what objectives the AI optimises for. The rulebook is known in micro-regulations, and usually static on the timescale during which decisions are made. The objectives become increasingly dynamic as we get to longer timescales and less frequent and more severe events. Whereas 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.

That is a particular problem for macro-regulations, both for regulations and especially for crisis resolution. When faced with the most severe crises, society demands that we do what it takes to resolve them. If the extant rules and laws stand in the way of the preferred resolution process, they can be changed. There are repeated examples of parliament passing emergency legislation to change the existing law to facilitate resolution, such as in Switzerland during the Credit Suisse collapse.[1] Pistor

[1]https://www.bloomberg.com/news/articles/2023-03-20/credit-suisse-collapse-reveals-some-

(2013), in her legal study on 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.

This means it is impossible to prespecify the objectives of a crisis resolution process, except at the highest levels of abstraction. The design and enforcement of macro-rules may be subject to unpredictable political interference. When a severe crisis happens, the political leadership takes charge. That is 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 legitimacy to do that – AI is not.

The resolution of a crisis depends critically on information and interests that only emerge endogenously during the resolution process. Previous crisis resolution objectives will not apply because both the political environment and the crisis details will differ from previous crises. To effectively understand the objectives as they emerge, one needs to have an intuitive understanding of the ideas, knowledge and objectives of other stakeholders, which in a crisis can involve large private sector institutions, the judiciary, parliament and the political leadership. We cannot tell AI about preferences for things that have not happened so reinforcement learning will be impossible.

These four conceptual challenges have particular implications for the use of AI by the financial authorities. The effectiveness of AI depends on what we ask of it.

AI will broadly benefit the micro-authorities, where there are few concerns about its use. There is plenty of data to learn from. The rulebook is fairly static. Many decisions made by human supervisors feed into reinforcement learning, and the cost of mistakes is relatively low. These factors play to the strengths of AI. As we note in Table 1 on page 19, we expect the use of AI to grow rapidly to advise on regulatory design, monitor the system for compliance with rules and ultimately make supervisory decisions. Whereas the lack of AI explainability might hamper decision making today, explainability is very important across AI use. Once an AI supervisor can explain its decisions, there is no reason, in principle, why it could not make enforcement decisions and, along the way, provide justification that could be used in subsequent legal proceedings.

The picture is different for macro-authorities because in this sphere data is scarce, the rules are mutable and events are mostly unique and infrequent. Here, a data-driven learning approach based on pattern extraction and learning from past decisions can

ugly-truths-about-switzerland-for-investors be a problem. For AI to be effective and not make serious mistakes, it would have to gain a deep understanding of the stakeholders' reaction functions and objectives that change as part of the political process. Humans deal with this by intuitively mapping experiences from different domains, such as history, law and ethics. There is considerable risk that AI will provide the wrong advice or make poor decisions. Even if that is not the case, the potential that it could happen means the authorities need to be alert to risks arising from AI use in macro-regulations.

AI will probably be crucial for advising the decision-makers and will be much quicker in running scenario analysis than human experts, which will help to get a better outcome for the resolution process. However, it cannot be left to make all the decisions because its strength of constructing scenarios and inter- and extrapolating from statistical models becomes a weakness compared to the intuitive approach of a group of humans using distributed decision-making based on intuitive understanding. Ultimately, that means there may be little to distinguish between AI advising and decision making, especially in resolving crises, as an analysis of events can strongly direct decisions taken in response to them.

The public sector's use of AI in regulations raises tough questions. Who is accountable when the AI regulator makes decisions or provides crucial inputs for human decisions, and how can a regulated entity challenge decisions? The regulatory AI may not explain its reasoning or why it thinks it complies with laws and regulations. The supervisory AI will need to be overseen — regulated — differently to human supervisors.

As the private sector rapidly expands its use of AI, new risks emerge. AI systems used by the private sector are better at finding optimal solutions than their human counterparts, but at the risk of such solutions being socially undesirable. Financial markets are particularly susceptible to such outcomes because they have strong complementarities that can lead to undesirable phenomena such as liquidity hoarding, bank runs and fire sales. Worse, these complementarities will probably strengthen as AI accelerates the adoption of best-of-breed methodologies for measuring and managing risk. Whereas that is beneficial, it also means that market participants and the authorities will increasingly interpret shocks similarly, and regulations will induce market participants to react in the same way.

2 Conceptual challenges

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

The use of AI in financial regulations is frustrated by the complexity of the problem and limitations of AI. Financial market participants operate in highly uncertain social environments that are subject to frequent structural changes, while neither the rules nor the players' objectives might be known. Sometimes, the participants can even change the rules to their advantage in a way that others only partially observe. While not usually a problem for the micro authority, the uncertainty and mutability of the macro controllers' problem give rise to four conceptual challenges in AI macro use: data, unknown-unknowns, response to control and specification of objectives.

2.1 Usefulness of data

Data should play to AI's advantage as the financial system generates many petabytes 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 it to be easy for AI to study the financial system in detail and identify all the causal relationships. That is true for most micro problems, but not macro.

Start with the basic measurement. The standards for recording 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 that were 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 credit. The situation has improved somewhat since then due to mandatory trade reporting for many derivatives transactions and increasing availability of data on cash and repo transactions for bond markets. However, silos persists 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).

These three issues, data quality, silos and rarity of events, are usually not all that important for micro. Not so for macro where they amplify each other. That, in turn, impacts both the design of macro 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 macro 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 is, 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 and complexity 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 specifics that cause vulnerabilities, their job is frustrated by the almost infinite complexity of the financial system so that the supervisors can only patrol a small part of it. Even if 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 thresholds. The system's complexity and measurement noise mean 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 false positives, and be seen as socially unacceptable.

This uniqueness of crises creates particular problems for the designers of macro regulations because they really only know what data is useful after an event. That was, for example, the case in the crisis in 2008. It only became transparent afterwards that sub-prime mortgages being put into structured credit products — where 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 has an almost infinite amount of signals and an enormous amount 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 Strategic interactions

A key challenge for AI working for the macro authorities, but not generally for the micro 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, financial institutions in our case, 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, in response to a given crisis will change the responses of the private sector to a similar crisis in the future in potentially unexpected ways.

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 micro regulations, the frequency of events and lack of 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 final 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 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 micro. The rulebook is known, and static on the timescale, most decisions are made static. It does involve over time in response to events and the regulated response to regulations. Still, AI can quickly update its understanding of the objectives, both by adopting changes to the rulebook and by reinforcement learning from observing how the human supervisors act.

This is not the case in macro, as it operates on very long timescales, the time between events it is concerned with usually being in the decades. Furthermore, it can be very difficult to define the macro objective except at the highest levels of abstraction, such as preventing severe dysfunction in key financial markets and especially the failure of systemically important institutions. This applies to the design of macro regulations, macro 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 macroprudential regulations are susceptible to lobbying and political interference, which means that the objectives of regulations can change over time. Reinforcement learning using 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.[2]. 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. That is 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.

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 crisis, these silos break down. All relevant authorities, the affected private sector, the judiciary and especially the political leadership 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, philosophy, morality, history, technical background and objectives to the table. Such a process can be highly robust. All pertinent issues are discussed, including information that was until then confidential or implicit. Such analysis depends on implicit knowledge and intuitive understanding its participants have of each other and how that aligns with what they know and want. The intuitive understanding is so vital to crisis resolution and drives the objectives while being particularly difficult for AI to learn.

Crisis resolution is arguably the most important aspect of financial policy, especially for central banks — arguably one of their raisons d'être. Nevertheless, given the

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

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 in certain dimensions. 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.

3 AI usefulness for the financial authorities

One of the hardest problems for AI applied to decision making in complex social settings, like macro regulations, is the specification of its objectives. The algorithm needs a precise objective function that evaluates the cost and benefits of alternative courses of action given the current state of the environment and its future evolution. It needs to take into account how the system reacts to its actions. Misspecifying the problem leads to suboptimal 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 AI's, presence was not welcome.

If we ask ChatGPT whether it is okay for admirals to sink their own ships, it says no. However, this is a well-known example, likely in the 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 in many ways, most harder to detect than 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 make it more predictable, similar to how social media algorithms attempt to change user preferences. This can have severe consequences for both micro and macro regulations. In the former, perhaps leading to unethical or biased outcomes and, in the macro case, procyclicality, amplifying the financial cycle on its way up and down.

Even if the authority had an explicit policy of only using AI for advice and not for decisions, that might not be as big a distinction as it thinks. While the AI engine will have its own internal representation of the financial system, 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 directions in 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. The AI, when optimising, may even choose to present alternatives in that manner so as not to risk having the operator make inferior choices. In both cases, we have reached the stage where AI is effectively making decisions. This applies to all stages of AI use by financial authorities, whether macro or micro, in regulation design, enforcement or resolution. Eventually, we risk becoming dependent on a system for critical analysis and decisions we don't fully understand.

This particular problem is common in many AI applications, as there are many cases where it is crucially important to understand how it sees the world, analyses and makes decisions — explainable AI. Significant resources are being brought to bear on this particular problem and, if successful, will alleviate many of these concerns. However, the impact on financial authorities will not be uniform. Explainable AI will particularly benefit micro regulations because the problem is relatively simple with a rich sample of observations. It will be different in macro because then we ask AI to understand the entire financial system, which is almost infinitely complex. This is a far bigger challenge and will likely take much longer if it can be achieved at all.

While an authority might not wish to get to that point, the use of AI might end up there regardless. As we come to trust its analysis and decisions and see how well it performs in increasingly complex and important tasks, it may end up where the authorities do not want to be by stealth. By then, it may be impossible to turn the AI engine off because it is performing vital functions, and the risk of disastrous outcomes might be deemed unacceptably high.

Ultimately, the usefulness of AI for the financial authorities depends on what we want from it. Combining the conceptual challenges in the previous section with the AI concerns here, we get six criteria for using AI in financial policy that can be brought to bear on particular use cases. After applying this discussion to regulation and resolution in the next section, Table 1 on Page 19 applies this list to particular policy actions.

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?

4 Regulating AI

As the authorities increasingly use AI in the design of regulations, the enforcement of and the resolution of stress events, the usefulness criteria of AI discussed in Section 3 have particular implications for financial regulations.

4.1 Regulation of private AI

The financial markets have particular characteristics that challenge both micro and macro AI. The reason is that almost all outcomes in the system are determined by the interaction of well resourced and highly incentivised participants, maximising profits in normal times and survival during extreme stress. This homogeneity of objectives and ample resources make financial markets particularly vulnerable to coordinated behaviour. Many investment decisions have strong complementarities where the benefits of a given action increase in the number of market participants taking that action, leading to undesirable phenomena such as collusion, bank runs, fire sales and flights to safety.

While such outcomes have always been a feature of financial markets, long before the use of computer technology, they are strengthened by using computers and algorithms to analyse data and make decisions. The increased use of AI will further amplify that process. Given the homogeneity in private sector AIs' objectives, the data they are trained on, and their superior ability to find optima, we expect AI to exploit complementarities more efficiently than their human counterparts, potentially amplifying stress. Such outcomes can emerge innocently, where private and public sector AI push for best practices in measuring and managing risk, harmonising knowledge and action and amplifying the financial cycle. The result is increased systemic risk caused by procyclicality.

There is a further grey area where a private sector AI finds that in optimising, it can best meet its objectives by bypassing or manipulating both the regulatory system and the standards of conduct for the behaviour of private entities. It might use this to attack competing AI, collaborate with them, and opt to attack the authorities' AI. One example 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 both very complex and often illegal. AI is much better at handling complexity and is unaware of the legal nuances unless explicitly taught or instructed.

Some market participants might find it particularly beneficial to deploy AIs capable of identifying profitable loopholes in times of heightened stress since anybody forewarned of stress can profit, so creating and amplifying existing stress is profitable. We have seen many examples where human decision-makers seek to profit from creating stress. AI will make that even easier.

AI is also useful for those seeking to take advantage of the financial system for financial gain illegally and to cause damage, like criminals, terrorists and hostile nation-states. The reason is that the computational problem such entities face is much easier than that of the public authorities as they only have to find one area to exploit. In contrast, the authorities have to patrol the entire system. The increased private sector use of AI makes it increasingly harder for the authorities to monitor and control such criminal and terrorist behaviour.

The increased private sector use of AI furthermore creates legal problems for the financial authorities because it results in another 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. Consequently, the increased use of AI in the private sector facilitates the job of those economic agents seeking to utilise AI for nefarious purposes by providing them with yet another level of denial.

4.2 Regulation of public AI

Several issues arise when the financial authorities use AI for regulations. The first is that as we increase the use of AI for micro-regulations, the AI will tend to pick the same best-of-breed techniques for managing risk. While inherent in all regulatory designs, AI will likely amplify the use of similar risk measurement and management methods. That is a concern because the more market participants and regulators come to see the world in the same way, the more regulations will induce market participants to react to shocks in the same way. That amplifies the financial cycle, creating booms and busts — procyclicality. Even worse, this risk monoculture means the system becomes vulnerable to the same unknown-unknowns – further increasing systemic risk.

Even if the authority aims to mitigate such drivers of procyclicality, it may be inevitable because of how the public and private sector AI may communicate and how the private sector manages risk. Because of the high fixed costs in risk modelling and management, creating risk management systems is an increasing return to scale business, similar to cloud computing. That can drive risk management to a handful of AI vendors that amplify homogeneity in beliefs and actions, amplifying procyclicality. Such outsourcing to risk management as a service (RMaaS) is rapidly happening, not the least to BlackRock's Alladin.

The financial authorities face new issues when it comes to regulating financial institutions that extensively use RMaaS, where the authorities will likely also use the same vendors for data, models and computing resources. This blurs the regulator / regulated divide. The authorities will have to identify how to regulate such multiuse global public infrastructure. Furthermore, even if an authority would prefer such systems to operate from its own jurisdiction, that would be less important than the actual design of the system, which might be done elsewhere. Consequently, even if an authority wanted the RMaaS to be explainable, it might not be as achievable as if the RMaaS operator was solely in its jurisdiction.

An additional issue for public sector use of AI is accountability. If a human supervisor makes a mistake, we can hold someone accountable, give them more training, or dismiss them. It is not as easy with AI making the same decisions. Who is accountable? What does it mean to train them differently? We can't dismiss them because that might lead to an entire essential function in the supervisory apparatus left unaddressed. When the decisions of the AI are challenged, the regulatory AI may not be able to explain its reasoning or why it thinks it complies with laws and regulations.

Ultimately, this means that the internal supervision of AI use in the regulatory agencies and its interaction with the outside legal system will require different policies than those used for current human supervisors.

4.3 Evaluating issues in AI use in particular applications

The six step procedure for evaluating AI effectiveness shown in Page 15 and the conceptual challenges in Section 2 can be applied to particular applications as in the following table.

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
Micro 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	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
Management of global systemic crises	Very rare or not available	Full	Complex & conflicting	Multiple & international	Unclear even ex-post	Very severe

Source: Danielsson and Uthemann (2023)
"On the use of artificial intelligence in financial regulations and the impact on financial stability"

5 Conclusion

In this work, we have identified the main criteria for evaluating the pros and cons of AI use in the financial authorities and the conceptual problems that may arise. Many of the issues facing AI also affect human decision making. For some of these, AI will perform much better than human decision makers, such as in risk management and compliance, while in others perform worse than human decision makers.

The main areas where AI is disadvantaged relate to domains where data are limited, and the system is subject to frequent structural change. That frustrates learning since AI depends on data driven decision making processes, which can lead to potentially catastrophically bad decisions. Human decision makers are not as dependent on data driven analysis because they can bring in domain information from outside the problem being managed by AI and benefit from distributed decision making processes that help better solutions emerge. Human decision makers do not need the pre-specification of objectives on a level that AI requires. In some applications, that might not be very important, such as in many cases where AI only gives advice but does not make decisions. Furthermore, reinforcement learning can often help AI to learn from human analysts and decision makers. That is particularly relevant for micro.

AI excels and outperforms humans in risk modelling and management. It is making rapid inroads into detecting fraud, consumer protection and other misbehaviour, but its use is sometimes frustrated by data silos. However, there are technological solutions that may overcome such limitations. AI benefits such applications because there is plenty of data to train on. The decisions of human supervisors feed into the reinforcement learning algorithms AI makes such good use of. 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. AI can also amplify existing micro problems, such as those arising from algorithmic bias, perhaps with racial criteria for credit decisions, but better training should alleviate most of such concerns.

Several factors frustrate the use of AI for macro, and even worse, can cause it to mislead 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.

Furthermore, the increased use of AI, particularly in risk management, can potentially increase procyclicality and hence systemic risk. It will be better at finding state-of-the-art measurement and management techniques, which would then be similar across the system, harmonising beliefs and actions and inducing procyclicality. This will only be amplified by the increased outsourcing of risk management to a small set of institutions with superior technologies and data.

References

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.

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

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`.

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

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

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

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.

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

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.

Norvig, P. and S. Russell (2010). *Artificial Intelligence: A Modern Approach:*. Pearson.

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

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

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Economic and Social Research Council

Systemic Risk Centre

The London School of Economics
and Political Science
Houghton Street
London WC2A 2AE
United Kingdom

tel: +44 (0)20 7405 7686
systemicrisk.ac.uk
src@lse.ac.uk