Journal of Banking and Finance 140 (2022) 106290

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

Artificial intelligence and systemic risk☆

Jón Daníelsson[a,*], Robert Macrae[a], Andreas Uthemann[a,b]

[a] Systemic Risk Centre, London School of Economics, United Kingdom
[b] Bank of Canada, 234 Wellington Street, Ottawa ON K1A 0G9, Canada

ARTICLE INFO

Article history:
Received 31 October 2020
Accepted 13 August 2021
Available online 28 August 2021

ABSTRACT

Artificial intelligence (AI) is rapidly changing how the financial system is operated, taking over core functions for both cost savings and operational efficiency reasons. AI will assist both risk managers and the financial authorities. However, it can destabilize the financial system, creating new tail risks and amplifying existing ones due to procyclicality, unknown-unknowns, the need for trust, and optimization against the system.

1. Introduction

Artificial intelligence (AI) is rapidly changing how financial institutions are operated and regulated. While AI will bring considerable economic benefits, it also poses specific threats to the stability of the financial system — increasing systemic risk — both because of conceptual problems but also how its use will impact and alter the financial system.

The task of managing and interacting with the financial system, whether from the point of view of the regulatory authorities or the private sector, has two distinct, and in practice, separate, dimensions. The micro problem encompasses both microprudential regulations and internal risk management of financial institutions, focused on day-to-day risk, such as large daily losses on individual positions, fraud and regulatory compliance. While immensely detailed, the emphasis here is on the short and medium run and the control of many repeated similar events. The mapping from individual actions, whether private or regulatory, to the state of the system is clear, as is the ability to judge a state in light of the objectives. It is these characteristics that facilitate the work of the micro AI. The picture is different with macroprudential policy and related private sector objectives such as long-term tail risk — the macro objectives to be executed by the macro AI. Here the emphasis is decidedly long run, avoiding systemic crises and large losses decades hence, where the mapping between objectives, actions, and outcomes is highly uncertain and the events being controlled are very few and mostly unique.

In this work, we contend that AI is well suited for the micro problems while facing serious conceptual and practical challenges when used for public or private macro objectives. In order to identify those challenges, we trace out the systemic consequences of using AI for macro control. A significant conceptual challenge is data availability. By their very definition, there are very few observations on extreme stress in financial markets, where those observations are unique in important aspects. Furthermore, such data is generated within a specific policy framework, and as both market participants and authorities learn from past stress, they change the environment, a direct application of the Lucas critique (Lucas, 1976). And finally, any AI has to be given precise objectives, difficult when the macro problem does not provide a clear and actionable formulation of its objectives, and any fixed objective is inherently vulnerable to unknown-unknowns. In turn, these conceptual problems give rise to many practical challenges facing those who want to use AI for macro control, including optimisation against the macro AI, excessive trust in the system, and increased procyclicality.

While there is no uniform notion of what AI is, we adopt a common definition that maintains that AI is a computer algorithm that makes decisions that otherwise would be taken by human beings. It is given objectives and instructed to guide some process towards meeting those objectives — the rational agent approach according to the taxonomy developed in Norvig and Rus-

☆ Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the Bank of Canada. We thank Will Cong, Jean-Philippe Dion, Travis D. Nesmith, Jun Yuan, and conference and seminar participants at "The impact of machine learning and AI on the UK economy" conference at the Bank of England, the "Data Driven Financial Stability: Opportunities and Challenges in Big Data" conference at the Bank of Finland, and the European Central Bank for comments and suggestions. We thank the Economic and Social Research Council (UK) [grant number ES/K002309/1] and the Engineering and Physical Sciences Research Council (UK) [grant number EP/P031730/1] for their support. First version: November 2017.

* Corresponding author.
E-mail address: j.danielsson@lse.ac.uk (J. Daníelsson).

https://doi.org/10.1016/j.jbankfin.2021.106290

sell (2010). The AI could be guided purely by some rules of play, like in games, but more often obtains information by machine learning (ML), whereby a computer algorithm takes in data and infers the reduced form statistical model governing the data. The usefulness of AI for any task, as noted by Russel (2019), depends fundamentally on the structure of the task at hand. Its best use case is a single agent decision problem with known immutable objectives, rules and a predetermined bounded action space. The more we deviate from that ideal scenario, the more poorly AI performs.

Three conceptual reasons frustrate the macro AI working at the behest of the financial authorities and the private sector. The first follows from the Lucas critique (Lucas, 1976) — economic agents' optimal decision rules depend on the structure of the underlying environment. Any changes to this environment, including those brought about by a macro AI, will lead agents to adapt their decision rules. Behavioural responses that the AI engine infers from historical data are contingent on the observed environment and can break down if the engine attempts to exploit them for control purposes. To regulate the financial system, the macro AI will not be able to solely rely on conventional ML techniques that infer patterns in existing data. It will have to complement this with an understanding of the system's causal structure, including economic agents' reaction functions and the underlying political system.

The second challenge facing any AI put in charge of the macro objective is paradoxically data. While the financial system might seem to be the ideal use case for AI as it generates seemingly infinite amounts of data, measurement problems, silos, and hidden interconnections limit the information that can be gleaned. While not much of a hindrance in micro applications, it likely misinforms the macro AI.

Finally, it is a general property of crises that they catch everyone by surprise. Systemic crises are typically unknown-unknowns, where every crisis has statistical patterns that make it unique. This makes learning from existing data, even if these data include numerous previous crises, challenging for any AI. It also means that the regulators only know what to exactly guard against ex-post. All they can do ex-ante is to specify general objectives. But then, how do we define a concrete objective such as "keep the financial system stable"? This is currently achieved by modular organisational structures having formal and informal communication channels, with personnel selection based on education, experience, and performance. It is not known how to replicate such decentralised objective formulating mechanisms when designing AI. If a regulatory AI is to act autonomously, humans will have to first fix its objectives. But a machine with fixed objectives, let loose on a highly complex environment, will have unexpected behaviour (Russel, 2019).

The three conceptual challenges facing the financial AI: How economic agents' responses to AI affect the system, data, and unknown-unknowns, in turn, cause the AI to impact the financial system in undesirable ways.

The first follows from the need to give the macro AI a clearly defined fixed set of objectives that make it more transparent and predictable than human regulators who can use strategic ambiguity in their communications. The precise objectives and transparency facilitate the inadvertent or deliberate attempts of economic agents to escape control and exploit loopholes. Some agents inadvertently use individually innocent strategies that, in aggregate, are damaging, while others deliberately act in a destabilising way for profit, legally and otherwise. The biggest challenge may be those agents intent on damage, terrorists and nation states, as a lack of a profit motive can make it harder to identify them early. All of that ultimately threatens the stability of the financial system.

The second consequence of AI's use for the financial system arises from how AI will gain trust. The problem is that trust creeps upon us. When we see AI performing well in low-level functions, it gives the green light to higher-level functions. Cost savings on expensive human domain knowledge will provide additional incentives to adopt AI for decision making. While, of course, present in the current setup, there are crucial differences between human decision-makers and AI that make the problem of trust particularly pernicious. It is harder to ascertain how AI reasons than a human decision-maker, nor can we hold AI to account. And because we do not know how it would react to unknown-unknowns, the question of trust becomes increasingly pertinent as AI encroaches on macro like problems. The longer we leave a macro AI in charge, the harder it will be to switch it off. Its knowledge of the financial system and internal representation of data will become unintelligible to humans. Turning it off risks disrupting the system in unforeseen ways.

The final practical consequence arises from how the financial system is affected by policy changes. In particular, AI is likely to amplify the inherent procyclicality of the financial system above and beyond what the current decision-making process does. There are two reasons for this. The first is that the AI will much more robustly find best-of-breed processes, so the AI will settle on a small homogenous set of risk management techniques performing well most of the time but also vulnerable to the same unknown-unknowns. Furthermore, their superior performance in good times will increase trust in the AI and induce additional risk-taking. Both amplify the financial cycle — AI is procyclical. Ultimately, it will have to contend with Minsky's dictum "stability is destabilising".

Most research on AI and ML in finance and economics focuses on finding better solutions for applied problems, such as improved portfolio selection (see e.g. Ding et al., 2018; Cong et al., 2020) or better prediction of asset returns and measurement of risk premia (see e.g. Gu et al., 2020; Bianchi et al., 2020). Some work points out the conceptual problems that arise when ML techniques are used for decision making. Beyond the well known problem of algorithmic bias (Cowgill and Tucker, 2019), AI driven approaches for policy have raised the issue of how an AI can infer causal links from statistical correlations (see e.g. Athey, 2017; Athey et al., 2019) and how it deals with measurement error when decision making is based on ML predictions (see e.g. Mullainathan and Obermeyer, 2017; Kleinberg et al., 2018). Here, we identify new types of algorithmic biases, broadly understood as erroneous internal representations that biases decision making away from desired outcomes, that arise when an autonomous AI is used for macro control and trace out their consequences for financial stability. Some of the problems we identify in this work are closely connected to the debate on the value of reduced-form macroeconometric evidence for aggregate control dating back to Goodhart (1974) and Lucas (1976).

The organisation of the remainder of the paper is as follows. We start by discussing AI in Section 2, focused on definitions and key issues, and then use that in Section 3 where we formally define macro and micro control, how AI interacts with them and especially its relationship to endogenous risk. That takes us to the conceptual challenges in Section 4, how AI changes the system, data, and specifications of its objectives. We then move on to the practical issues in Section 5 like optimisation against the system, trust and procyclicality. Section 6 concludes.

2. AI and how financial complexity affects it

Artificial intelligence (AI) is a broad field of research that does not admit a simple definition.[1] Here, we focus on the rational agent approach to AI, computer programs that act to achieve the

[1] See Norvig and Russell (2010) for a detailed discussion of its history and different approaches to defining AI.

best expected outcome given pre-specified objectives — a familiar perspective for economists. The tasks AI is to perform require a structured representation of the environment, knowledge of the rules that have to be followed and a formal specification of the objectives to be achieved.

While the concept of AI relates to the behaviour of an algorithm, machine learning (ML) is a fundamental component in most AI applications. ML is the process of acquiring knowledge about the environment the AI engine is tasked with controlling. The available methodological approaches and the accuracy of what the engine learns depends directly on the quantity and quality of data. Learning can be supervised or unsupervised. Unsupervised learning involves discovering patterns in the data without requiring any a priori knowledge of a problem's structure. These patterns are then mapped into mathematical logic, generating information on statistical relationships between observations. Typical use cases are the de-noising of large datasets (Ng, 2017) or asset clustering (Bryzgalova et al., 2020). In supervised learning tasks, an algorithm infers a mapping from inputs to output based on example pairs. The researcher feeds the ML existing domain knowledge, like information on the data generating process or causal links, thereby reducing the dimensionality of the learning process. Typical applications of supervised learning in finance are prediction tasks, such as forecasting asset returns and inferring risk premia (Gu et al., 2020; Bianchi et al., 2020).

If the AI is to act autonomously, it has to learn how to make optimal decisions. This requires information about the structure of the environment, either provided to it by a human expert or inferred from data using ML. But the AI also needs to be given clearly defined objectives that its actions are to achieve. Then, given a representation of its task environment and objectives, the AI can act. To teach the AI to make optimal decisions, most applications use reinforcement learning (Sutton and Barto, 2018), which employs dynamic programming techniques and a combination of exploration and exploitation to allow the AI to learn the state-contingent relationship between its actions and the ensuing payoffs.

Reinforcement learning algorithms were initially developed for single agent Markov decision problems but have lately seen successful applications to strategic interactions in which the AI engine interacts with humans or other algorithms. Some interactive tasks are naturally suited for this, such as the strategic recreational games Go and chess. When DeepMind's AlphaZero AI was shown the rules of Go, it figured out how to play the game better than any human in less than 48 hours, simply by playing against itself (Silver et al., 2017). Games like chess and Go belong to a particular category of problems, games of complete information. The players of such games have complete information on the strategic situation and are fully informed about all feasible moves. They know their objective and, importantly, also their opponents' objectives. The current state of play gives the strategic situation, say a board position, and is fed as an input into a flexibly parameterized function, typically a deep neural network, that outputs both suggestions for next moves and an evaluation of the current situation in terms of the probability of winning.

For most strategic settings, the AI engine needs to be endowed with a more sophisticated *theory of mind*, meaning an internal representation of opponents' objectives and beliefs about the environment that goes beyond merely thinking of them as clones of itself. Recent AI advances help it in such cases. Brown and Sandholm (2019), for example, provide a successful application of self-play to Poker, a multi-player game with incomplete information. Particularly challenging are tasks that are not purely adversarial zero-sum games but require some cooperation among players, like the game Diplomacy and many real-world problems like driving a car. Here, the benefits of coordination among players can lead to multiple local optima, which creates additional problems for a learning algorithm (Bard et al., 2020).

While such strategic games can be seen as an ideal case for AI applications, they do not reflect the reality of interactions in financial markets. Unlike in strategic games, the strategically relevant state variables in market interactions are rarely obvious a priori. They either have to be provided to the AI based on existing economic theories or other human domain knowledge, or the AI has to learn them. When playing games, AI benefits from knowing that its opponents have simple objectives — all they want to do is win — which allows it to generate training data via self-play. This assumption becomes problematic in games of incomplete information, like all finance applications, where AI is uncertain about the types of opponents it faces. Historical data could, of course, be used to simulate the behaviour of market participants. However, the financial system continually undergoes structural changes; new types of market participants enter the game all the time; others drop out, and financial innovation opens up new moves. This reduces the value of historical data for simulations, especially when it comes to extreme events, those that are destabilizing and a concern for the authorities.

Finally, when algorithms make decisions, they will also make mistakes. Systematic mistakes by the AI lead to algorithmic bias: erroneous internal representations bias decision making away from the desired outcome. There are several reasons for bad AI decisions. It might have been provided with erroneous data, leading the AI to perpetuate human error or bias in a supervised learning task, for example, racial bias in credit scoring of loan applications (Klein, 2020). The algorithm might also make decisions based on statistical patterns in data that are either spurious (Mullainathan and Obermeyer, 2017) or that change once the AI engine attempts to exploit them (Athey, 2017). One of the hardest problem for AI applied to decision making in complex social settings relates to the specification of its objectives. The algorithm needs to be given 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 given chosen actions. Misspecification of the structure of the problem will lead to suboptimal decisions.

3. The macro and micro problem

Finance is essential. It provides financial intermediation — channelling funds from one person to another across time and space. Finance reallocates resources, diversifies risk, allows us to build up pensions for old age and enables companies to make multi-decade investments. Finance is also dangerous and exploitative. Banks fail, financial crises happen and financial institutions exploit their clients. The response of society is to enjoy the benefits of the financial system while also demanding it be regulated and controlled heavily.

The regulation and control of financial activity can be classified into two main categories, micro and macro. Micro control, to be executed by the micro AI, encompasses microprudential regulations and most internal risk management in financial institutions. It is inherently concerned with day-to-day activities of financial institutions, is hands-on and prescriptive, designed to prevent large losses or fraudulent behaviour, mandating and restricting how institutions should operate, what they can and cannot do, codified in the *rulebook*. While the rulebook was once in paper form, it is now increasingly expressed as digital logic, allowing programmatic access. Most, but not all, of the objectives a micro AI has to meet exist in the rulebook, and it generally has an ample number of repeated similar events to train on. All of this facilitates the application of AI to micro financial problems.

Longer term objectives, such as the solvency of key institutions, financial stability and tail risk, risks that threaten the functioning of the financial system – systemic risk – are macro problems. Inside the regulatory space, that encompasses macro prudential regulations, and in the private sector, the management of solvency and liquidity risks for large financial market participants such as banks, insurance companies or mutual funds. The macro task is much harder. Macro risk is created by the strategic interactions of many players and involves aggregate phenomena such as bank runs or fire sales (Benoit et al., 2017). It is inherently global, but the devices of control are predominantly local.

A multitude of national regulators aim to control macro risk. The ability to coordinate on regulatory responses is severely limited by institutional factors such as localised control of data, national law, and domestic political objectives. By contrast, micro is predominantly local, facilitating the job of individual authorities. Furthermore, macro risk is concerned with infrequent and severe outcomes, while micro focuses on many similar events of smaller magnitude. Crises are rare and unique, the outcomes of decisions made years and decades earlier, typically in times when all outward signs point to stability, so taking on more risk was not seen as problematic. This is why being safe can lead to excessive risk taking, as noted by Ip (2015). The challenge for the financial authorities is crystallised in the words of Minsky (1986) "Stability is destabilising", economic agents, when they perceive the world as safe, want to take more risk. Daníelsson et al. (2018) provide empirical evidence for this mechanism.

3.1. Risk, exogenous and endogenous

One of the hardest problems for anyone tasked with controlling some aspect of finance, whether macro or micro, is the measurement of risk. After all, an essential part of meeting a macro objective is controlling risk of large shocks tomorrow, as well as years and decades hence. The concepts of exogenous and endogenous risk, as proposed by Daníelsson and Shin (2002), are helpful in conceptualising the challenges of risk measurement. Exogenous risk is readily measured by statistical techniques, whether traditional risk models or machine learning. The measurement process takes in historical observations on prices and other pertinent variables, inferring the distribution of future outcomes. A fundamental assumption to risk being exogenous is that the economic agents who interact with the financial system do not change the system.

Endogenous risk maintains that everybody who interacts with the system changes it. Endogenous risk arises from the interaction of economic agents and is typically most severe when they stop behaving independently and start coordinating. This happens when stress constrains their behaviour, such as increased capital and margin requirements or the need to liquidate investments to meet redemptions. The consequence can be a vicious feedback loop between market stress, binding constraints, and harmonised behaviour, ultimately culminating in a significant stress event or a crisis (Brunnermeier and Pedersen, 2008).

AI is well suited for measuring and managing exogenous risk because it can use large data samples, well-established statistical techniques, and many repeated events to train on while the objectives are straightforward. All of these facilitate ML, allowing for classification, simple extrapolation, and, eventually, better portfolio selection (Ding et al., 2018; Cong et al., 2020). Consequently, AI will likely make significant inroads into micro regulations and internal risk management in banks. The technology is mostly here already, and the cost and efficiency gains considerable. BlackRock's Aladdin and MCSI's RiskMetrics are widely used risk control platforms that make extensive use of micro AI for decision making (BlackRock, 2019).

To measure endogenous risk, it is necessary to identify the build-up of threats today that may culminate in a crisis many years in the future. Meanwhile, the nature of these rare crises varies a great deal making it hard to extract general patterns, frustrating the use of ML. Moreover, there is no obvious way of measuring such endogenous risk. The underlying drivers of bad outcomes are hidden until they manifest themselves at the time of crisis. All large shocks and crises are fundamentally endogenous in nature, which means that measurement processes based on exogenous risk such as SES, SRISK and ΔCoVaR are unable to capture the most severe risks, as argued by Daníelsson et al. (2017).

For the most part, the micro authorities are concerned with exogenous risk, the reason why even current AI is useful to them. It is not so for the macro authorities because the risk they care about is endogenous risk. While beneficial for basic data handling and modelling tasks, for AI to be of more fundamental help, it needs to understand endogenous risk and reason and act strategically, taking into account how market participants will react to hitherto unseen events.

4. Conceptual challenges

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

The most celebrated successes of AI pertain to well-defined strategic games such as chess or Go. That does not mean AI will do well with real world problems that are more complex and unstructured. Financial market participants operate in highly uncertain social environments in which even the game being played continually changes. The rules or the objective of each player are not generally known and the players can change the rules to their advantage in a way that the other players only partially observe.

The uncertainty and mutability of the macro controllers' problem give rise to three serious conceptual challenges; how economic agents' responses to AI affect the system, data for macro problems and how AI reacts to unknown-unknowns.

4.1. System response to AI

The first conceptual challenge for the AI designed to understand and control the financial system is the Lucas critique (Lucas, 1976). Economic agents' optimal decision rules depend on the structure of the underlying environment. Any changes to this environment, including those brought about by macro regulators, will lead agents to adapt their decision rules. Behavioural responses that an AI engine infers from historical data are contingent on the observed environment and can break down if the AI engine attempts to exploit them for control purposes.

A crucial element for successful control is understanding market participants' beliefs about the environment, including their understanding of the macro controllers' strategies. Unfortunately, these beliefs are generally latent. We only learn indirectly about them by observing agents' actions or aggregate outcomes such as prices. Consequently, the ML working for the AI will only capture a reduced form model of economic reality. This includes information on how economic agents have reacted to particular instances of control in the past but misses out on how agents' beliefs and, consequently, actions will change in reaction to new policies generated by the AI.

Recent work has considered how ML can be used to identify causal channels (Athey et al., 2019). There is also a large body of work, both in microeconomics and macroeconomics, that has developed econometric methods to identify the causal effects of

policy interventions (Angrist and Pischke, 2008; Nakamura and Steinsson, 2018). However, these techniques cannot provide insights into policy regimes that have not been implemented before. If data on the relationship between an action a and an outcome y in a given state of the world x do not exist, AI in its current form will not be able to make statements about the causal impact of a on y given x — the inherent problem of reduced form models. Conventional econometric identification schemes mostly rely on human domain expertise, such as narrative approaches or a priori sign restrictions on effects, to establish causality.

Sudden changes in the global environment that expose hitherto unknown vulnerabilities are among the challenges that all regulators have to address. While the rule of law attempts to codify and regularize large areas of social activity so that plans can be made and bargains struck, these rules and laws result from a political process. They evolve, as witnessed, for example, by the vast number of new financial regulations created over the last decade in response to the crisis in 2008. To overcome the lack of data, the AI engine would have to work with structural models that encode causal links between control actions and system responses. This requires knowledge of economic theory and the ability to handle abstract concepts, not a trivial undertaking. Such knowledge currently exists in distributed human form located in regulatory agencies, academia and businesses with communication channels across institutions to establish a common understanding of the environment.

And finally, to control system behaviour, AI needs to be understood. Market participants need to know what to expect from the AI to make plans. In micro decisions, that is important to avoid legal challenges, and in macro, so that aggregate quantities, such as inflation, can be taken into account when contracting. Clear and credible communication with the public is a challenge for human regulators. For AI, it is likely an even harder problem. We can always ask a human regulator how they arrived at a decision, but AI can use impenetrable logical structures to execute its task. Significant advances have been made recently in the interpretability of AI that facilitate this task (Hase and Bansal, 2020), but so far, we are far away from being able to ask any financial AI to explain its decision-making process.

4.2. The usefulness of data

The second conceptual challenge for AI's application to macro control is data. On the face of it, data should play to AI's advantage. The financial system generates a vast amount of it. Every transaction is recorded, decisions are documented, decision makers are monitored and recorded, and we can track processes over their lifetime. Financial institutions are required to report some of this data to the financial authorities, which should be able to directly ascertain whether the financial institutions are behaving according to the objectives set by the authority.

There are, however, four reasons why this sea of data may not facilitate the work of AI. The first relates to measurement. Standards are inconsistent so that the same transaction could be recorded differently, with different codes when reported by the various counterparts. Furthermore, many internal systems were not set up with data collection and sharing in mind. It is both expensive to capture data, and data can be inconsistent given the absence of standards and system heterogeneity. Fortunately, while real today, these problems will likely be overcome slowly over time.

A bigger problem is the lack of data sharing and silos. Most data stays within a financial institution and is not shared with anyone. Some are shared with the relevant financial authority, but even then, the financial institution can retain copyright and access control, only allowing the authority data access for compliance purposes. The authority may be legally prevented from using that data for risk control purposes. Individual financial authorities are reluctant to share data with other authorities in the same country, not to mention those abroad, explaining the lack of a global risk database. That problem was made clear in the 2008 crisis when the vulnerability from structured credit products was only visible when considering the aggregate market, but there was no way to do such measurements ex-ante.

Furthermore, the type of events of concern to the macro controllers are, by definition, rare. The typical OECD country only suffers a systemic crisis one year out of 43 according to the IMF crisis database maintained by Laeven and Valencia (2018), and the type of tail losses that undermine the solvency of other major market participants, such as pension funds or insurance companies, are almost as rare. Even more problematically, each of these crises, significant stress events and tail losses are to a considerable extent unique.

Politics is a strong driver of both short term and long term economic and financial system outcomes. Political uncertainty directly maps onto the financial markets as shown by Kelly et al. (2016) and the fortunes of the political leadership is directly affected by the financial system, as Liu and Shaliastovich (2021) demonstrate. However, it is hard to quantify the quality of political decisions and how they impact on the system. Therefore, any AI solely making use of financial market data for learning about and controlling the financial system in the short and long run, will miss out on the crucial political dimension.

Major stress events arise from interconnections between seemingly disparate parts of the system, fueled by political linkages, connections that only manifest themselves once stress is underway. In times of stress, a particular combination of observations may induce financial institutions to behave in a particular way, for self-preservation purposes, like hoarding liquidity. That response is particularly damaging for stability, but since we do not know how an institution will respond, it is unclear whether a particular set of observations is damaging or not. In other words, we do not know ex-ante what data to feed to the ML servicing the macro AI. We only know ex-post. The sources of fragility, fire sales, runs and other negative feedback loops are well understood theoretically. Still, the concrete form they take is context specific and depends, most importantly, on the current financial market structure and political environment. That is why even if the AI trains on an exhaustive dataset containing detailed observations on previous crises, it will not find all the vulnerabilities. It will likely miss the most important ones – those nobody, neither the regulators nor the market participants, have been aware of ex-ante.

4.3. Unknown-unknowns and fixed objectives

The former US Secretary of Defense, Donald Rumsfeld, classified events into three categories: known-knowns or certainties; known-unknowns, events we think might happen; and the unknown-unknowns that are a complete surprise. In the language of Knight (1921), known-unknowns are risk and unknown-unknowns are uncertainty.

Known-unknowns do not tend to cause crises as we can anticipate and prepare for them. If the US stock market were to go down by $200 billion today, it would have a minimal systemic impact because it is a known-unknown. Even the largest stock market crash in history, on October 19, 1987, with a one day downward move of about 23%, implying losses in the US of about $600 billion, or $1.4 trillion in today's dollars and global losses exceeding $3 trillion in today's dollars, only had limited impact on financial markets and practically no impact on the real economy. Losses in the financial crisis of 2008 were surprisingly small. The overall subprime market was less than $1.2 trillion, and if half of the mort-

gage holders had defaulted with assumed recovery rates of 50%, the ultimate losses would have amounted to less than $300 billion. And that is an extreme scenario as actual losses were much smaller (Ospina and Uhlig, 2018). Still, the threat of such an outcome brought the financial system to its knees because it revealed unanticipated vulnerabilities and hidden linkages in the system.

It is a general property of crises that they catch everyone by surprise. Systemic crises are typically unknown-unknowns. But this means that it is hard to specify the objectives the macro AI has to meet to control systemic risk. If every crisis has statistical patterns that make it essentially unique – the source of the surprise – the regulators only know what to guard against ex-post. All they can do ex-ante is to specify general objectives. While in principle feasible for human policymakers, ignoring the practical problem of what these objectives should be — financial stability, minimizing systemic risk, long-term economic growth — is not possible for AI. For it to know how to evaluate specific outcomes, the financial authorities have to specify the objectives of the AI engine in detail. It appears very unlikely that this can be done via high-level, abstract concepts, for example "keep the system safe". How should the macro AI, for example, trade off financial losses that occur at different points of the economy's wealth distribution? Furthermore, if you ask a financial authority what their objectives are, they will likely be vague — constructive ambiguity — a successful strategy to deal with the moral hazard of crisis interventions. A macro AI with fixed objectives cannot be ambiguous.

Human regulators cannot foresee the unknown-unknowns, but they are reasonably well equipped to respond to them. As the presence and importance of hitherto ignored factors become apparent, they can update their objectives, using established processes to respond. They have historical, contextual, and institutional knowledge; they reason well with theoretical concepts. Decisions are taken within modular organizational structures with formal and informal communication channels, with personnel selection based on education, experience and performance. It is not known how to replicate such decentralized objective formulating mechanisms for AI. That means that, while AI will assist in collecting information and modelling parts of the problems, crisis decision making will likely remain a human domain for the foreseeable future.

5. Practical consequences

As we start employing AI for risk management, micro regulations, and especially for macro regulations, the three conceptual challenges, how the system response to AI, the usefulness of data, and the interaction of unknown-unknowns with fixed objectives, lead to three practical consequences that are of particular concern: optimisation against the system, the need for trust, and procyclicality.

5.1. Optimisation against the system

The structure of the financial system is not static, evolving instead in a directed manner because of the endogenous interactions of the agents that make up the system. The financial authorities adapt the constraints imposed on the system to meet their policy objectives, while other economic agents, human or AI, see their objective as maximising expected profits subject to those regulatory constraints as well as self-imposed risk limits. In addition, competition introduces an important adversarial element into the financial games, and rules aimed at inhibiting risk-taking by financial entities often become obstacles to be overcome. All of these follow from how the financial system responds to AI, in the spirit of the Lucas (1976) critique (see Section 4.1).

Agents' optimisation against the system takes many forms. It could be innocent and inadvertent, like those institutions buying structured credit products in the years before 2008, the danger of which was only visible once the crisis was underway. Self-preservation, accepted and encouraged by the micro authorities, can lead financial institutions to coordinate in hoarding liquidity, leading to a credit crunch. Alternatively, speculators can exploit legal loopholes, destabilise the system in the name of profit maximisation without breaking the law. Examples include various carry trades, where strategic complementarities attract evermore traders in destabilising behaviour.

Coordination among economic agents can become more common in a system with pervasive use of AI when algorithms learn to cooperate. Whether that is good or bad depends on what equilibrium they coordinate on. Calvano et al. (2020), for example, show that independent reinforcement learning algorithms are very good at sustaining collusive equilibria in pricing games, keeping prices above competitive levels. For human actors, collusion is not only difficult to sustain as it can be strategically very complex, it may also be illegal. But the human owner only needs to instruct its pricing algorithm to maximise profit. The algorithm can implement an anti-competitive strategy without its human owner having told it to do so or even being aware of it. Such tacit collusion raises serious legal and practical concerns for the regulators.

Because of its fixed objectives, a macro AI will likely be more transparent than human regulators who can use strategic ambiguity in their communications. It will provide more publicly observable trigger points that market participants can use to coordinate their behaviour. Whether this improved ability to coordinate is problematic depends on the nature of the problem. Transparency can, for example, facilitate coordination in bank run scenarios and be detrimental to financial stability, a problem well understood in the context of regulatory stress tests (Bouvard et al., 2015).

The common factor here is profit maximisation. An alternative form of optimisation against the systems involves agents intent on damage, whether terrorists, rogue nations or criminals seeking a ransom in exchange for system stability. They actively search for vulnerabilities, and by solving the problem of double coincidence of stress — attacking the system at its weakest point when it is most vulnerable — can cause significant damage and even a systemic crisis.

AI has to content with all these categories of agents, which puts it at a disadvantage, especially the macro AI. It faces a highly complex computational problem as it has to monitor and control the entire system. The opponent only has to identify local loopholes that can be exploited. AI's intrinsic rationality amplifies this advantage. It makes the AI engine predictable, giving its adversaries an edge. Rational behaviour within a well defined environment allows for the reverse engineering of the AI's objectives via repeated interactions. Human drivers facing a self driving car, for example, can gain an advantage in traffic.

The system's complexity works against the macro AI, in a way consistent with the Lucas critique. The attackers only need to use machine learning to identify and exploit loopholes in the current system, while the macro AI needs to understand the system's reaction when it acts to close these loopholes. That is a much more challenging problem as the AI needs to understand the attackers' reaction function. The AI engine's problem is compounded by having to monitor the entire system, since, as we note in Section 4.2, data lives in silos, making monitoring across silos difficult.

Countermeasures certainly exist. The standard defence is for AI to react randomly in interactions with human beings or other AI, limiting their ability to game it, mimicking humans' natural defence – they create randomness, nuance, and interpretation, varying across individuals and time. However, in the context of financial policy, that can raise thorny legal issues if randomised responses have to be programmed into a regulatory AI.

5.2. Trusting the machine

If a regulatory AI is to act autonomously, humans will have to first fix its objectives. But a machine with fixed objectives, let loose on a highly complex environment, will have unexpected behaviour (Russel, 2019). The unknown-unknowns, inherent in such an environment, and the inability to specify fixed, comprehensive and immutable objectives, raise fundamental questions of trust for any financial authority, human or AI. But it is even more critical for AI, further amplified by lack of data reliability as discussed in Section 4.2.

In the 1980s, an AI decision support engine called EURISKO used a cute trick to defeat all of its human competitors in a naval wargame, sinking its own slowest ships to maintain manoeuvrability. This very early example of reward hacking[2] illustrates how difficult it is to trust AI. How do we know it will do the right thing? A human admiral doesn't have to be told that they can't routinely sink their own ships (if they do, it requires high-level political acquiescence.) Any current AI has to be told, but the real world is far too complex for us to pre-specify rules covering every eventuality, so AI will predictably run into cases where it will take 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 by the designers of the TCS simulation system." (Lenat, 1983, p 82). Each of EURISKO's three successive victories resulted in rules changes intended to prevent any repetition, but in the end, only telling Lenat that his presence was unwelcome proved effective.

The human decision maker and the government that employs her have well known strategies for coping with unforeseen contingencies. As the presence and importance of hitherto ignored factors become apparent, she can update her objectives, making use of established political processes to impose checks and balances on how such decisions are made. Trust in human decision making also comes from a shared understanding of values and a shared understanding of the environment. AI has no values, only objectives. And its understanding of the environment will not necessarily be intelligible to humans. We can run hypotheticals past an AI engine and observe its decisions but cannot easily ask for an explanation (Joseph, 2019). The longer we leave an AI engine successfully in charge of some policy function, the more it becomes removed from human understanding. Eventually, we might come to the point where neither its knowledge of the economic system nor possibly even its internal data representations will be intelligible to its human operators.

The inability of any current or foreseeable AI to have a useful model of high level policy decision making might mean that AI will be relegated to small and safe regulatory functions. We think this unlikely because trust creeps upon us. Most people would have baulked at AI managing their personal finances 20 years ago, or five years ago few would have trusted self driving cars. But we have no problem entrusting our lives to AI landing aircraft and AI controlling surgical robots. AI is proving its value in myriad daily applications, and as AI proves its value to policymakers, they will start trusting it. AI will do an excellent job in good times, probably for many years, and trust will increase. You might be willing to give up on explainability when you can see many examples of success, and the AI model of the objectives and constraints that matter during the good times will be highly refined and successful because they will receive frequent testing and evaluation.

[2] For a list of similar examples see Krakovna (2018).

The existing rulebook mostly specifies the objectives and rules for micro regulations, so there is little danger of AI making a seriously wrong decision. If not, then the short reporting timescales involved mean we realise the problem quickly, allowing us to react early. It is different with macro regulations. AI will run into cases where it takes critical decisions in a way that no human would — the financial version of sinking its own ships. AI will only become helpful for macro policy if its reasoning and assumptions can be effectively explained to human supervisors. For current generation AI, this poses enormous challenges. Go grandmasters debate AlphaZero's moves because AlphaZero cannot explain them.

The more we rely on AI, the harder it is for human regulators to take the reins when problems emerge. The AI's knowledge of the financial system and internal data representations will likely be unintelligible to humans. Human authorities that believe that the AI's model of their objectives and constraints is flawed but can only access information and exert control through the medium of this model are in a very difficult situation.

5.3. Procyclicality and risk monoculture

Procyclicality is a significant cause of financial instability. The day-to-day activities of financial institutions are inherently procyclical. Banks lend more freely in good times, amplifying credit booms and contract lending when things turn sour, leading to a credit crunch which drives the economy down (Schularick and Taylor, 2012). Risk sensitive capital exacerbates the procyclicality because of how the risk weights are calculated. Risk weights are based on defaults, and as loan defaults are low in the upturn and high in the downturn, so are the risk weights, a prime example of data problems (see Section 4.2), and the challenge of modelling risk in an environment of unknown-unknowns. Risk sensitive capital is consequently low in the up cycle and high in the down cycle. It amplifies the cycle.

As the measurement of risk is based on the financial institutions' perception of the current riskiness of an asset and the system, the more homogenous these risk measurements are, the more similarly financial institutions see the world. Most financial institutions have similar objective functions that determine their behaviour, expected profit maximisation subject to risk constraints. Hence, a harmonisation of risk perceptions inevitably makes them act more procyclically. A more powerful risk measurement system leads to more effective optimisation, taking each financial institution closer to its optimum portfolio and closer to other financial institutions' portfolios. The consequence is more similar perceptions of risk and more crowded trades. Heterogenous perceptions and objectives can act as a counterbalancing force and thereby stabilise the system.

All three drivers of procyclicality, risk appetite, measurement, and optimisation have been steadily harmonised over the past decades, driven by regulation, best practices, and more sophisticated risk measurement techniques. The growing dominance of a handful of AI-based risk management systems like BlackRock's Aladdin and MCSI's RiskMetrics will further strengthen this tendency. Microprudential regulations increasingly dictate the amount of risk allowed for banks, pension funds, insurance companies and other risk taking entities and how they are meant to manage that risk.

We contend that AI amplifies the inherent procyclicality of the financial system. AI will have a comprehensive knowledge of all publicly available data, state-of-the-art risk models and best practices. The various AI working in the private and public sectors are set to converge on how they perceive risk and best manage that risk. Even if the risk appetite of financial institutions remains different, how AI measures and manages risk will work to unify their behaviour. Furthermore, as AI will give the financial authorities better access to financial information, the regulators may increas-

ingly prefer to exercise control over private sector institutions' risk appetite, as they already do extensively.

We see the AI engines at both financial institutions and regulators in good times acting with increasing conformity, standardisation, and groupthink precisely because AI will do a better job on risk estimation and management than more primitive risk measurement systems. The higher degree of confidence in managing risk will likely coincide with an increased willingness to take on risk.

When new information arrives in the form of an unexpected shock, AI engines in both the private and public sectors will update their models similarly. This is because all see risk in the same way. Their comprehensive knowledge of all public data and shared risk management processes means they are surprised in the same way. The result is that the financial institutions' AIs will change their portfolios simultaneously and in the same way, potentially triggering destabilising dynamics such as fire sales. An example of such outcomes occurred in the summer of 2007 when AI-based quant fund algorithms unexpectedly coordinated on selling, leading to vicious feedback loops between algorithmic responses that caused the price of certain assets to fall significantly (Khandani and Lo, 2011). Furthermore, as we noted in Section 4.2, all crises are different in essential ways, so the macro AI will necessarily be confronted with new information in the form of unexpected shocks, exposing linkages that no AI engine, neither public nor private, has observed before.

In summary, the use of AI for risk control can be expected to lead to better performance in good times, driving its adoption. But, unfortunately, this comes at the cost of greater procyclicality and increased systemic risk.

6. Conclusion

AI is making increasing inroads in financial applications, driven by efficiency and cost savings, and it seems likely to be of substantial and growing benefits to micro problems. AI is most useful when it has clear rules and can observe repeated related. It has to know what it is allowed to do and must be able to infer meaningful associations and relationships embedded in the data. It helps if risks involved are known-unknowns and can be reasonably treated as exogenous. In such situations, standard models work well, and any problems should be relatively minor and quick to diagnose. These conditions apply well to both micro regulatory and financial institutions' AI, where AI can increase efficiency and reduce costs.

The same does not apply to the macro regulations concerned with the stability of the entire financial system. In a crisis, rules are broken or evolve. Data is scarce as historical data can become irrelevant, and associations might change overnight. Unknown-unknowns dominate and the endogenous nature of risk cannot be ignored. To operate effectively in this environment, AI would need to understand causality, reason on a global rather than local basis, and identify threats that have not yet resulted in adverse outcomes. These are all well beyond current capabilities.

Our ultimate conclusion is that there is a dichotomy between the macro and micro financial problems that directly affect how AI will be useful and should be implemented. The increased use of AI for micro prudential regulations and internal risk management will be, for most parts, socially beneficial. Increasing efficiency, fairness and robustness of the provision of financial services while significantly decreasing costs. The adverse consequences will mostly face employees now doing jobs that would be overtaken by AI, such as risk modelling and management and low level policy analysis. It is different with the macro concerned with the stability of the financial system and the prevention of large losses that threaten the solvency of pension funds, banks and insurance companies. Macro addresses threats years and usually decades into the future, the events are very rare and quite unique, raising severe issues of procyclicality, trust and ability to manipulate the control processes.

We furthermore suspect that regardless of the trajectory of technology, these problems will not be overcome. Consequently, the use of AI in macro prudential regulations and to critical private sector objectives should be heavily scrutinized and rejected if any of these issues become pertinent. The longer we leave a macro AI in charge, the harder it will be to switch it off. Its knowledge of the financial system and internal representation of data will become unintelligible to humans. Turning AI off risks disrupting the system in unforeseen ways.

Declaration of Competing Interest

None.

Acknowledgement

We thank the UKs Economic Research Council and Engineering and Physical Science Research Council for funding in a footnote on the title page (footnote denoted *) and list the respective grant numbers.

References

Angrist, J.D., Pischke, J.-S., 2008. Mostly harmless econometrics: An empiricist's companion. Princeton University Press.
Athey, S., 2017. Beyond prediction: using big data for policy problems. Science 355 (6324), 483–485.
Athey, S., Bayati, M., Imbens, G., Qu, Z., 2019. Ensemble methods for causal effects in panel data settings. AEA Papers and Proceedings 109, 65–70.
Bard, N., Foerster, J.N., Chandar, S., Burch, N., Lanctot, M., Song, H.F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., et al., 2020. The hanabi challenge: a new frontier for ai research. Artif Intell 280, 103216.
Benoit, S., Colliard, J.E., Hurlin, C., Perignon, C., 2017. Where the risks lie: a survey on systemic risk. Rev Financ 21, 109–152.
Bianchi, D., Büchner, M., Tamoni, A., 2020. Bond risk premia with machine learning. Review of Financial Studies 2 (32), 1046–1089.
BlackRock, 2019. Artificial intelligence and machine learning in asset management. Whitepaper. https://www.blackrock.com/corporate/literature/whitepaper/viewpoint-artificial-intelligence-machine-learning-asset-management-october-2019.pdf
Bouvard, M., Chaigneau, P., Motta, A.D., 2015. Transparency in the financial system: rollover risk and crises. J Finance 70 (4), 1805–1837.
Brown, N., Sandholm, T., 2019. Superhuman AI for multiplayer poker. Science 365 (6456), 885–890.
Brunnermeier, M.K., Pedersen, L.H., 2008. Market liquidity and funding liquidity. Review of Financial Studies 22, 2201–2238.
Bryzgalova, S., Pelger, M., Zhu, J., 2020. Forest through the trees: building cross-sections of stock returns. Available at SSRN 3493458.
Calvano, E., Calzolari, G., Denicolo, V., Pastorello, S., 2020. Artificial intelligence, algorithmic pricing, and collusion. American Economic Review 110 (10), 3267–3297.
Cong, L.W., Tang, K., Wang, J., Zhang, Y., 2020. Alphaportfolio for investment and economically interpretable AI. Available at SSRN 3554486.
Cowgill, B., Tucker, C.E., 2019. Economics, fairness and algorithmic bias. Journal of Economic Perspectives (forthcoming).
Daníelsson, J., James, K., Valenzuela, M., Zer, I., 2017. Can we prove a bank guilty of creating systemic risk? a minority report. Journal of Money Credit and Banking 48. https://ssrn.com/abstract=2692086
Daníelsson, J., Shin, H.S., 2002. Endogenous Risk. Modern Risk Management — A History. Risk Books. www.RiskResearch.org, http://www.RiskResearch.org.
Daníelsson, J., Valenzuela, M., Zer, I., 2018. Learning from history: volatility and financial crises. Review of Financial Studies 31, 2774–2805.
Ding, Y., Liu, W., Bian, J., Zhang, D., Liu, T.-Y., 2018. Investor-imitator: A framework for trading knowledge extraction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1310–1319.
Goodhart, C. A. E., 1974. Public lecture at the Reserve Bank of Australia.
Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev Financ Stud 33 (5), 2223–2273.
Hase, P., Bansal, M., 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?2005.01831.
Ip, G., 2015. Foolproof: Why safety can be dangerous and how danger makes us safe. Little, Brown and Company.
Joseph, A., 2019. Opening the machine learning black box. Bank of England underground blog, https://bankunderground.co.uk/2019/05/24/opening-the-machine-learning-black-box/.
Kelly, B., Pastor, L., Veronesi, P., 2016. The price of political uncertainty: theory and evidence from the option market. J Finance.

Khandani, A.E., Lo, A.W., 2011. What happened to the quants in august 2007? evidence from factors and transactions data. Journal of Financial Markets 14 (1), 1–46.
Klein, A., 2020. Reducing bias in AI-based financial services. Technical Report. Brookings.
Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2018. Human decisions and machine predictions. Q J Econ 133 (1), 237–293.
Knight, F., 1921. Risk, uncertainty and profit. Houghton Mifflin.
Krakovna, V., 2018. Specification gaming examples in AI. https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/.
Laeven, L., Valencia, F., 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. Artif Intell 21 (1–2), 61–98.
Liu, Y., Shaliastovich, I., 2021. Government policy approval and exchange rates. J financ econ.
Lucas, R.E., 1976. Econometric policy evaluation: A critique. In: Carnegie-Rochester conference series on public policy, Vol. 1. North-Holland, pp. 19–46.
Minsky, H., 1986. Stabilizing an unstable economy. Yale University Press.
Mullainathan, S., Obermeyer, Z., 2017. Does machine learning automate moral hazard and error? American Economic Review 107 (5), 476–480.
Nakamura, E., Steinsson, J., 2018. Identification in macroeconomics. Journal of Economic Perspectives 32 (3), 59–86.
Ng, S., 2017. Opportunities and challenges: Lessons from analyzing terabytes of scanner data. In: Advances in Economics and Econometrics: Eleventh World Congress. Cambridge University Press, pp. 1–34.
Norvig, P., Russell, S., 2010. Artificial intelligence: A modern approach:. Pearson.
Ospina, J., Uhlig, H., 2018. Mortgage-backed securities and the financial crisis of 2008: a post mortem. NBER Working Paper (24509).
Russel, S., 2019. Human compatible. Allen Lane.
Schularick, M., Taylor, A.M., 2012. Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870–2008. American Economic Review 102 (2), 1029–1061.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al., 2017. Mastering the game of go without human knowledge. Nature 550 (7676), 354–359.
Sutton, R.S., Barto, A.G., 2018. Reinforcement learning: An introduction. MIT Press.