Artificial intelligence and systemic risk*

Jón Daníelsson, Robert Macrae and Andreas Uthemann
Systemic Risk Centre
London School of Economics

This version June 2019

Abstract

Artificial intelligence (AI) is rapidly changing how the financial system is operated, taking over core functions because of cost savings and operational efficiencies. AI will assist both risk managers and microprudential authorities. It meanwhile has the potential to destabilise the financial system, creating new tail risks and amplifying existing ones due to procyclicality, endogenous complexity, optimisation against the system and the need to trust the AI engine.

*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. Updated versions of this paper can be downloaded from our website `www.riskresearch.org`.

Contents

1 Introduction

Artificial intelligence (AI) is rapidly changing how financial institutions are operated and regulated, and it will likely increase efficiency and reduce costs.[1] AI will be particularly beneficial to the financial regulators, helping them, to cope with the infinite complexity of the financial system. At the same time, the increased use of AI also threatens the financial system, increasing financial systemic risk.

The financial system has traditionally been the most heavily regulated part of the economy. A well functioning financial system is essential for economic growth and finance is very profitable to governments. Meanwhile, it is easy for banks to exploit their clients, and financial crises are extremely costly. Not surprisingly, every aspect of the financial system is under close government scrutiny, all the way from rules protecting unsophisticated bank clients with standards on font sizes in documentation, what is known as micro rules, to regulations aiming to ensure banks behave prudently, thus lowering systemic risk, the macro rules.

Micro regulations lend themselves easily to AI as the regulators oversee many similar events, facilitating the use of machine learning. AI will rapidly take over banks' financial risk functions as the tasks are relatively straightforward, and data to train on is ample. Once we have AI in charge of risk management and regulations, the supervision of banks will be executed by the AI engines in banks and regulatory agencies talking to each other, ensuring compliance, efficiency, lower costs and fewer mistakes. What stands in the way are cultural, political and legal considerations, not technical ones.

At the other extreme of the regulatory spectrum lies systemic risk, issues relating to the stability of the entire financial system. Here, the challenges are very different as financial crises are both rare and unique, frustrating standard machine learning techniques. Even then, AI will be irresistible to the macro authorities. Banks are complex institutions and report large amounts of data the authorities have limited ability to process, particularly challenging for the macro authorities that not only need to find risks emanating from individual banks; they also have to identify risks arising from all the interactions between banks. It is the aggregation of information to the level of the

[1] A 2017 Financial Stability Board finds the impact of AI on the practice of finance to be broadly positive.

entire system that is particularly challenging, severely limiting their ability to understand systemic risk. AI can help with that, but also poses particular dangers that any implementation needs to take into account. We identify four particularly pertinent issues.

First, AI is unable to reason about events it has not seen. Yet, such events are the root causes of most financial crises. When faced with new situations, human beings draw on their experience in making decisions, taking into account ethical, political and social considerations. AI, at least in the present and any foreseeable forms, cannot do that. When faced with its first crisis, the AI engine will perform unpredictably and can even contribute to instability. To keep society safe, we will need a kill switch for AI.

Second, it is difficult, to the point of impossible, to understand how AI makes its decisions. Human decision makers can explain their reasoning with reference to their education and life experience. We can ask how they would arrive at conclusions in hypothetical scenarios before putting them in charge. It is not possible to do the same with AI.

Third, AI is more likely to amplify economic and financial cycles than current human regulators. Automation favours standardised, best-of-breed and hence homogeneous methodologies, implying monoculture. Monocultures are more fragile than heterogeneous ecosystems, in the financial system just as in nature, as a single threat impacts all members in a similar way. This creates pro-cyclicality as more participants will update their behaviour in lockstep, increasing systemic risk.

Finally, it will be easier for market participants to exploit a financial system regulated by AI for private gain, with systemic financial crises a possible outcome. Market regulations are an adversarial game where the objectives of the regulator and regulated are not aligned. The high rationality and predictability of AI, when coupled with the requirement for transparent and consistent implementation of regulations, help hostile agents more than the current human-centred setup. Transparency and fair play requirements prevent the regulatory AI from employing standard defences against hostile agents, like randomise responses, making it easy for opponents to optimise against it.

2 Financial regulations

Finance is essential for the economy. 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 companies to make multi-decade investments. Finance is also dangerous and exploitative. Banks fail, financial crises happen, and banks exploit their clients. The response of society is to enjoy the benefits of the financial system while also regulating it heavily.

Some regulations deal with the day-to-day activities of banks, micro regulations. They are hands-on and prescriptive, designed to prevent large losses or fraudulent behaviour, mandating and restricting how a bank should operate, what it can and cannot do, codified in the *rulebook*. While the rulebook was once in paper form, nowadays it is increasingly expressed as computerised logic, allowing programmatic access.

Regulated entities are responsible for following the rules, and the supervisor monitors compliance in various ways, ranging from analysing reported data to on-site inspections. The authority has extensive access to the internal information held by banks and considerable power to change bank behaviour if required. However, the financial system is vast, and supervisors can only monitor a small part of all the information reported to them.

Not surprisingly, financial regulations are extremely costly, and ultimately, those costs are paid for by the clients of banks; borrowers pay higher interest on loans while depositors get lower interests on deposits. It is hard to open bank accounts, and banking services require considerable paperwork. Any gains in efficiency benefit society, and here AI promises significant benefits, lowering costs and making cumbersome processes more efficient.

Longer-term objectives such as systemic risk and financial stability are the focus of macro regulations. Here, the job of the regulator is much harder because crises are rare and unique. The typical OECD country only has one crisis every 43 years,[2] so one is not usually around the corner. Crises are the outcomes of decisions made years and decades earlier, typically in times when all outward signs point to stability, so taking on more risk is not seen as problematic. However, "stability is destabilising" as noted by Minsky (1992).

[2]From the IMF crisis database, Laeven and Valencia (2018).

The very fact that the financial system is stable, incentivises economic agents to take more risk, creating future instability. That means the regulator needs to identify the buildup of risks 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.

3 AI and financial complexity

Artificial intelligence (AI) refers to a computer program that executes tasks that one would usually expect to be done by an intelligent human being. These tasks often require a structured representation of the environment within which the AI has to operate, knowledge of rules that have to be followed and a formal specification of objectives to be achieved.

3.1 Risk, riskometers and asteroids

For AI to be useful to the financial system, it has to understand risk, the core domain of finance. Taking on risk is a natural consequence of making investments, yet too much risk can lead to unsustainable losses and even system-wide crises. Unfortunately, there is no single concept called risk. It is a latent variable that has to be inferred from observed outcomes. Consequently, every measurement of risk is subjective, based on some mathematical model that has to be assumed, with many models to choose from, each providing very different answers, where it is hard to impossible to discriminate between them. Just one example is provided by Daníelsson (2015b), who shows that the most commonly used "riskometers" provided vastly different predictions of the likelihood of a large market outcome, when the Swiss appreciated its currency in January 2015.

The job of any financial regulator, human or AI, is frustrated by the fact that the easiest type of risk to measure, exogenous risk, is the least useful, while the hard to measure endogenous risk what matters most, a classification that comes from Danielsson and Shin (2002). Exogenous risk arrives in the financial system like an asteroid might hit the City of London, while endogenous risk is created by the interaction of the human beings that make up the financial system. Almost all financial risk of note is endogenous while

in practice, nearly all models for forecasting and managing risk assume risk is exogenous. This assumption is undoubtedly convenient for statistical purposes, and most of the time, harmless, when aggregate behaviour can be captured as an exogenous stationary random process. That holds when the entity making the risk estimate is small relative to the markets as a whole and other entities can be expected to behave independently of each other.

AI is ideally suited for measuring and managing exogenous risk because it can make use of large data samples, well-established statistical techniques, and many repeated events to train on. All of these facilitate machine learning, allowing for classification and extrapolation. Examples include the modelling of risk in financial derivatives and the management of the risk of derivative portfolios. 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 gain considerable.

Endogenous risk arises when economic agents stop behaving independently and start synchronising their behaviour. This happens in times of stress when constraints on their behaviour, such as best practices or capital and margin requirements, bind. The consequence can be a vicious feedback loop between market stress, binding constraints and harmonised behaviour, ultimately culminating in a major stress event or crisis as shown by the theoretical models of Brunnermeier and Pedersen (2008) and Daníelsson et al. (2011).

The micro authorities are, for the most part, concerned with exogenous risk, why AI will be useful to them. It is not so for the macro authorities, because the risk they care about is endogenous risk. For AI to be of help, it needs to understand that endogenous risk, as well as being able to reason and act strategically, taking into account how market participants will react to hitherto unseen events.

3.2 AI playing games

Some interactive tasks are naturally suited for AI, such as strategic games like 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. This allows powerful reinforcement learning algorithms[3] to be deployed; the strategic situation is given by the current state of play, say the board position, and is fed as an input into a flexibly parameterised function, typically a deep neural network, that outputs both suggestions for next moves and an evaluation of the current situation in terms of probability of winning. Play generated from these suggestions creates data on which the AI engine can be trained on.

When playing games, AI also benefits from knowing that its opponents think and act like itself, which allows it to generate training data via self-play. This last assumption becomes problematic in games of incomplete information, where the AI is uncertain about the types of opponents it faces. In poker, for example, we do not know the opponents' hands. For this class of problems, reinforcement learning algorithms currently do much worse than in games like Go and do not get close to human-like performance. The AI needs to be endowed with a more sophisticated *Theory of Mind* than merely thinking of opponents as its clones. This is particularly challenging for tasks that are not purely adversarial, zero-sum games but require some cooperation among players, like the game Diplomacy. Here, the benefits of coordination among players can lead to multiple local optima, which creates additional problems for a learning algorithm (Bard et al., 2019).

3.3 Games regulators play

In contrast to such idealised strategic games, most real world problems are much more complex and unstructured. Financial market participants operate in highly uncertain social environments in which even the game that is being played changes over time, and the market participants are often able to change the rules to their advantage. The reason is that the financial system is not invariant under observation, instead changing when studied. The rule of law attempts to codify and regularise large areas of social activity so that plans can be made and bargains struck. But rules and laws are the result of a political process, and they evolve, as witnessed, for example, by the vast number of new financial regulations created over the last decade in response

[3]Sutton and Barto (2018)

to the crisis in 2008.

If regulators want to use machine learning to the problems of financial regulations, they face a series of serious challenges. First, they have to describe the strategically relevant situation, in particular, they have to identify the relevant variables that summarise the current state of the system. Unlike in strategic games, these state variables are rarely obvious a priori. They either have to be provided to the AI based on, say, existing economic theories or the AI has to learn them.

Assuming this can be done, the AI would then have to learn how to behave optimally given the strategic situation it faces and the objectives it aims to achieve. Generating training data from self-play is unlikely to be appropriate given the asymmetry between regulators and regulated entities. Historical data could be used to simulate the behaviour of market participants. However, the system undergoes a regular structural change; new types of market participants enter the game all the time; others drop out, financial innovations open up new moves for market participant. This reduces the value of historical data for simulations.

A particularly tricky challenge for in the regulator is when the global environment suddenly changes, exposing hitherto unknown vulnerabilities. One example was in 1914 when after Archduke Ferdinand was assassinated in Sarajevo on June 29, the global financial system immediately got into a crisis, described by Roberts (2013). The reason was that the financial authorities across Europe anticipated a future war and hence decided to protect the financial markets from cross-border shocks. That, in turn, precipitated rapid deleveraging, resulting in the most significant financial crisis the world has ever seen, before and since. Neither the regulators nor anybody else knew the network structure of the financial system and had to draw on their experiences in other domains when fighting the crisis.

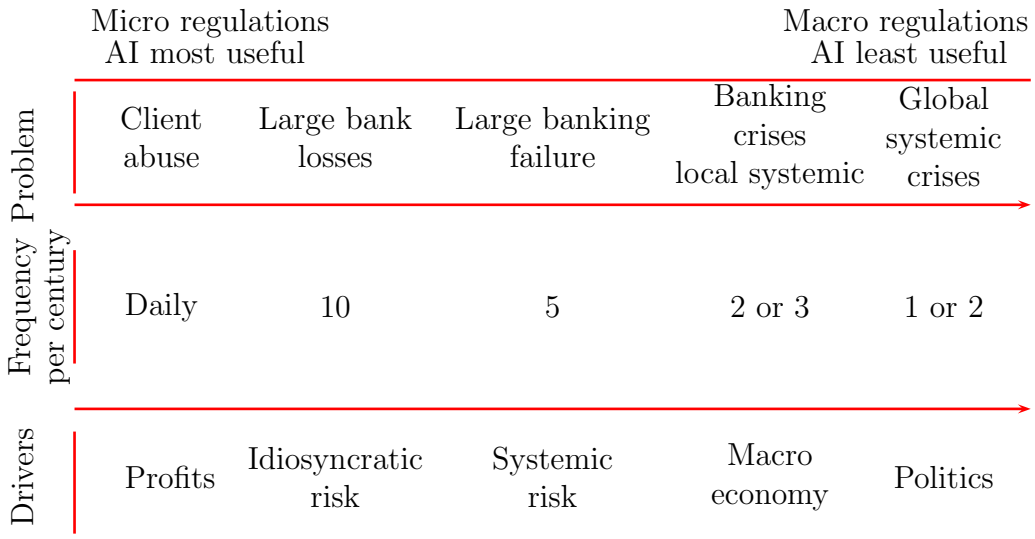
While the specific situation was unprecedented, the general vulnerabilities could be analysed in the context of historical episodes and understanding of the fragility of banks. By combining experiences from several different areas, the financial authorities were able to respond appropriately to what was to all appearances new event, without precedence.

3.4 Every minute, every hour, once in a lifetime

One of the biggest hurdles for the use of AI in financial regulation is data availability. A priori, the financial system might seem to be an ideal case for AI; after all, it generates almost infinite amounts of data to train on. Every minute decision is recorded, trades are stamped to the microsecond. Emails, messages and phone calls of traders and important decision makers' interactions with clients are recorded. The information leading to financial crises and misconduct is somewhere in this sea of data, and most of it is available to the financial authorities.

The nature of the problem dictates the frequency at which relevant events can be observed, as highlighted in Figure 1.

Figure 1: Artificial intelligence and the time dimension



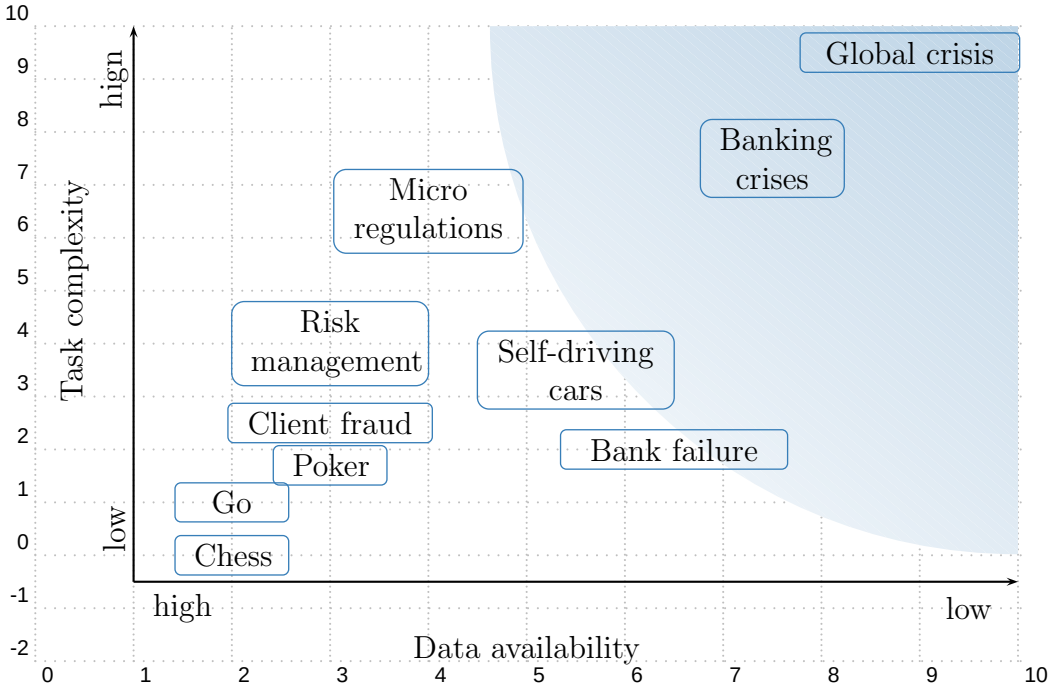
However, a large pool of financial data by itself is not sufficient. It has to be of the right kind, relevant to the objective and has to have supported over the range of possible outcomes. You could collect all financial information available over the period 1990 to 2007 and an AI engine would learn nothing about the crisis that was about to hit because even though all the data existed, the knowledge on how to connect the dots was not there. At the core of the crisis were subprime mortgages originated and regulated in US states, sold to federally regulated investment banks in New York, insured

by a New York State regulated insurance company via a London-based and regulated subsidiary of a wholly owned French-based bank, and sold on to municipalities throughout Europe. The specific fault line was collateralised debt obligations, CDOs, with embedded liquidity guarantees.

The root cause of the crisis in 2008 was politics, just like in every other crisis, with origins in the 1977 Community Reinvestment Act, as noted by Calomiris and Haber (2014), that set in motion events that culminated in a crisis 30 years later. The information of all of this was out there, but nobody connected the dots until it was too late. AI could not have done so either.

A useful comparison of the various tasks asked of AI can be seen in Figure 2. The x-axis shows the amount of data available to train on, and the y-axes the complexity of the task.

Figure 2: Data for AI



At one extreme, we have a game like chess, an easy task with ample data, while at the other, we have extreme global crises with very little available data while being extremely complex. In the middle, bank failures are not very complicated, and with the right data, are easily prevented, but the right data is hard to get. Micro regulations have ample data and an intermediate

level of complexity.

4 OK Computer?

> *Douglas Adams' "A Hitchhikers Guide to the Galaxy" features an exchange between a spaceship computer and galactic president Zaphod Beeblebrox. At one point, the computer informs Beeblebrox of its inability to defend the ship from a missile attack, to which Beeblebrox responds: "OK, computer, I want full manual control now".*

There are four main problems in using AI for financial regulation: Unknown unknowns, procyclicality, the need for trust, and the possibility to optimise against the system.

4.1 Unknown unknowns

The former US Secretary of Defense Donald Rumsfeld classified events into three categories. Known knowns, certainty. The known unknowns, something we anticipate might happen, and the unknown unknowns, what hits us out of the blue as a complete surprise.

Known unknowns do not cause crises as we anticipate such events 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 downward move of about 23%, implying losses in the US of about $600 billion, or $1.2 trillion in today's dollars and global losses exceeding $3 trillion in today's dollars, had little impact on financial markets and practically no impact on the real economy.

In the financial crisis of 2008, US subprime mortgages played a key role. What is surprising is how small the losses were. The overall subprime market was less than $1 trillion, and if half of the mortgage holders had defaulted with assumed recovery rates of 50%, the ultimate losses would have amounted to less than $250 billion. And that is an extreme scenario. Actual losses were smaller. Still, the mere threat of such an outcome brought the financial

system to its knees. The reason is that what materialised was an unknown unknown. The crisis caught everybody by surprise, and hence, nobody was prepared.

The ability to successfully scan the financial system for systemic risk hinges on where the vulnerabilities lie. Financial crises are driven by common factors well-founded in economic theory. Yet, the underlying details are usually unique to each event. After each crisis, regulators and financial institutions learn, adapt processes, and tend not to repeat precisely the same mistakes. When we examine the details of past crises, it is both clear that each had unique aspects, and that most of these were missed at the time. Indeed it is almost definitional that each crisis triggers a sudden and painful re-evaluation of previously comfortable assumptions.

While human risk managers and supervisors also miss the unknown unknowns, human regulators are more likely to know how to respond optimally to a crisis of unknown unknowns. They have historical, contextual and institutional knowledge, reason well with theoretical concepts and consequently have tools to handle it in a way that AI may not. That also means that human regulators may spot warning signs AI will miss. As the regulatory AI is most likely to focus on known unknowns, the danger from the AI is that it will focus on the least important types of risk, those that are readily measured while missing out on the more dangerous hidden connections in the system, more so than a human authority. In effect, it will automate and reinforce the adoption of mistaken assumptions that are already a central part of current crises. In doing so, it will make the resulting complacency even more likely to build up over time.

4.2 Trusting the machine

If we were to hire a policymaker in charge of financial stability, we can ask her how she would react to hypothetical situations and describe her reasoning. It is not possible to do the same with AI.

If AI is to make inroads into policymaking beyond micro policy, it is crucial to correctly and exhaustively specify its objectives, both intermediate and ultimate, to prevent undesirable outcomes. Suppose I tell the machine to minimise $f(x)$. My true objective function is $U(x, z) = f(x) + z$, but either I am, ex-ante, unaware of z or it is simply too complicated to spell out. The

AI engine might opt to minimise $f(x)$ but at the cost of maximising z. A human regulator with the same initial objectives will find out along the way that z also matters and update its objective function accordingly.

But what about the machine? When the EURISKO AI entered a naval war game in 1981, it easily beat all of its human competitors. All it did was to sink its own slowest ships so that it could outmanoeuvre all the other navies. A human being knows that is not an acceptable solution. AI has to be told. 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). The following year the rules of the war game were updated in response to EURISKO unwanted success, only for EURISKO to win again uncovering new loopholes ...

We have frequently seen the adverse consequences of ignoring important factors in past crises. During the Great Depression, the Federal Reserve was focused on moral hazard and inflation, ignoring the danger from deflation and failing banks, why a financial crisis and recession became a Depression, as noted by Friedman and Schwartz (1963). Similarly, central banks before 2007 were primarily concerned with the immediate objectives of monetary policy, neglecting financial stability.

Even so, the human decision maker has 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 the way such decisions are made. While AI might be able to do the same, we would have to trust it to make decisions in line with the objectives of its human operators.

This question of trust is fundamental. The longer we leave an AI engine successfully in charge of some policy function, the more it becomes removed from human understanding and the more we need to rely on trust. 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.

Paradoxically, as trust in an AI engine increases, so does the possibility of a catastrophic outcome when, eventually, the machine is forced to reason about

an unforeseen contingency. While AI will come up with some course of action, its analysis and conclusions might not agree with our human objectives. The consequences could be disastrous, perhaps a Minsky moment. This does not necessarily have to be the case. But we have no obvious way of entering into a dialogue with the AI engine in the same way a financial stability committee would consult with its experts. We might be forced to take its reasoning on faith, an outcome that is unlikely to be acceptable to the financial authorities.

A government entrusting AI with regulating the financial system will, therefore, want to do the same as President Zaphod Beeblebrox in the quote at the start of the section. It needs a kill switch, be able to turn off the AI engine if it poses a threat to society and be able to act independently of its advice if necessary.

The issue of trust is more relevant for the macro regulators than the micro authorities. The latter mostly execute low-level functions with clear objectives and limited damages in case of failure. With macro, the underlying problem is highly complex, the objectives are ill-defined, and the cost of failure potentially catastrophic, all characteristics that make AI not only less suitable but also more dangerous.

4.3 Procyclicality and risk monoculture

Procyclicality is a major cause of financial instability. When things are good, banks lend freely, amplifying a boom, and when things turn sour, they contract lending, creating a credit crunch, driving the economy down. While procyclicality is inherent in finance, data-driven approaches for measuring and managing risk exasperate the problem. Price data tends to be more stable and diversified in upturns than in downturns. Any backward looking data driven process, including machine learning, will identify risk as low in quiet times and high after a crisis.

The degree to which markets and regulators react in a procyclical manner is strongly influenced by how diverse their perceptions and objectives are. Diverse views and objectives dampen the impact of shocks and act as a stabilising force, because some will be right and some wrong, and many will update their expectations and hence portfolios in differing ways. This reduces systemic risk. Increased homogeneity in beliefs and actions, by contrast, amplifies systemic risk. Financial regulations, standardised risk management

practices and improved risk estimation all increase homogeneity for different but related reasons.

Financial regulations constrain financial activity, channelling it into specific areas or silos. Those within the same silo are similar to each other, and there seems little reason to expect the silos to diverge from each other by an amount sufficient to counteract this effect. Indeed much of the thrust of the last decade of regulation has been in precisely the opposite direction. One example is the realisation of the financial authorities that asset managers and insurance companies may pose systemic risk. Unfortunately, there is no body of research on why such institutions fail, unlike the two centuries of research on banking fragilities. The response of the authorities has been to apply the knowledge of bank fragilities to asset managers and insurance companies, thereby making the system in aggregate more homogeneous.

Standardised risk management practices within silos are also a problem. From the point of view of a micro regulator, it may seem desirable that all risk models provide the same evaluation for some given test portfolio, but enforcing this homogeneity comes at a cost in systemic fragility, as noted by Daníelsson (2015a). As every participant is forced to update their model identically in response to new information, all participants will wish to buy or sell assets at the same time.

Even improved risk estimation is a problem. After all, the direction of improvement is likely to be towards a common optimum and, as a consequence, all participants' behaviour will move closer to each other as noted by Watkins (2008). AI seems certain to improve risk estimation and so will contribute to this source of homogeneity of risk management techniques.

All of these developments, no matter how well-intentioned and otherwise efficient, increase pro-cyclicality and hence systemic risk. And, while many of them are very human failures, AI will make them worse rather than better by increasing the reliance on historical data and increasing the homogeneity of response.

4.4 Optimisation against the system

> *"Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."*

Goodhart's Law, Charles Goodhart 1974.

An AI engine in charge of financial stability might be quite effective in minimizing systemic risk if the structure of the financial system remained static or evolved in an exogenously determined stochastic manner. Then, the problem facing AI is one of sufficient data and computational resources. But the structure of the financial system is certainly not static. Instead, it evolves in a directed and often adversarial manner because of the endogenous interactions of the agents that make up the system. Economic agents are profit seeking and any rule that aims to reduce or inhibit risk taking activity will meet resistance from the agents whose preferred level of risk-taking is being obstructed.

There are many ways for economic agents to bypass financial regulations and even act maliciously. A simple way is to create new types of financial instruments or structures that have the potential to amplify risk across apparently distinct parts of the system. Any rule that restricts risk taking must be continually defended against new channels of risk transfer that attempt to profit by circumventing or attenuating it. The rules of the game evolve in response to players' behaviour rendering their motivation and action space endogenous, by which we mean being both a cause and a consequence of the regulatory activity. The complexity of the financial system itself becomes endogenous.

Any attempt of regulatory control will then have to take the reaction of the regulated entities into account. Merely extrapolating from historical data assuming market participants' behaviour will be invariant to the intended policy will not work, an implication of Goodhart's Law. To evaluate the potential effects of a planned policy, one needs to have historical data on previous implementations of the policy under similar circumstances. In the absence of historical data, economic models that establish causal channels from policy to reaction can provide guidance. A regulatory AI thus will need a model of causality and economic theory. Standard data-driven machine learning techniques will not be enough to identify successful regulatory policies. Furthermore, to learn, the AI will need to be able to experiment with policies which will involve some degree of random behaviour by the AI. But are we comfortable with machines conducting independent policy experiments on the financial system? Legal considerations, alone, suffice to prevent that.

Randomization by the AI might also be necessary to make it less vulnerable to adversarial attacks. Most AI systems are designed to be so opaque that outside agents cannot figure out how it operates, where the engine is further continually evolving to present a dynamic target. Meanwhile, the engine is allowed to experiment on users to identify their behaviour and randomize rules. These strategies are of limited use in financial supervision because rules have to be transparent, fair and relatively stable. These restrictions mean the AI engine has almost no flexibility in how it can respond to market participants.

That predictability and stability make it easier to optimize against the authority. Of course, human supervisors face the same problem. However, AI may be easier to predict than current human counterparts. If, for example, an AI regulator can be queried as to regulatory questions at a rate measured in seconds rather than days then agents will be able to map out its responses in much greater detail and so perform a more accurate optimisation. Furthermore, human regulators will exhibit random behaviour simply by virtue of their humanity. It will likely be unacceptable to program that into their AI counterpart.

Paradoxically, known rationality in strategic settings often constitutes vulnerability, and this may apply more to AIs than to human regulators. While the human regulators will have the same ultimate objectives as AI, their reactions may be harder to predict because of the complex social and policy structure that conditions their behaviour, especially under the extreme stress of financial crises.

5 Conclusion

AI is making increasing inroads in financial regulations, driven by efficiency and cost savings. While it is likely to benefit micro regulators, and can even be put in charge of important functions, its situation is different for macro regulations concerned with the stability of the entire financial system.

For AI to be useful for financial regulations, it has to meet several conditions. It needs clear rules and data on repeated outcomes. It has to know what it is allowed to do and be able to use machine learning to identify patterns in the financial system and causal relationships embedded in the data.

The risks need to be for the most part exogenous. The more important endogenous risk is for the problem at hand, the worse AI will perform.

The AI will also constantly be called on to make judgment decisions on a course of action not seen in the data or the rules. It has to interpret the rules in a way that is intelligible to sceptical human observers, including law courts.

It is inevitable that AI makes mistakes, just like the AI in Boeing 737 Max aircraft led to the deaths of 400 passengers. But, the costs of getting things wrong must be contained. Making mistakes is acceptable, causing a global financial crisis not.

Because it is impossible to trust the AI to respond appropriately, if it is in charge of important bits of financial regulations, we will have to have a kill switch, disconnect it if it behaves incorrectly in times of crises.

To be effective, a macro AI needs to be able to exercise control and share data across national borders and the silos inherent in regulations. To fend off attacks and malicious agents, it further needs to be able to randomize responses and even create rules in a nontransparent way. All of these are unacceptable to any government.

The AI further needs to be able to understand causality, be able to reason on a global rather local basis, and identify threats that have not yet resulted in adverse outcomes. These are all beyond current capabilities.

References

Adams, D. (1978). The hitchhiker's guide to the galaxy. BBC broadcast.

Bard, N., J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, et al. (2019). The hanabi challenge: A new frontier for ai research. *arXiv preprint arXiv:1902.00506*.

Brunnermeier, M. and L. Pedersen (2008). Market liquidity and funding liquidity. *Review of Financial Studies 22*, 2201–2238.

Calomiris, C. W. and S. H. Haber (2014). *Fragile by Design: The Political Origins of Banking Crises and Scarce Credit*. Princeton University Press.

Daníelsson, J. (2015a). Towards a more procyclical financial system. *VoxEU.org*. `voxeu.org/article/towards-more-procyclical-financial-system`.

Daníelsson, J. (2015b). What the swiss fx shock says about risk models. *VoxEU.org*. `www.voxeu.org/article/what-swiss-fx-shock-says-about-risk-models`.

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

Daníelsson, J., H. S. Shin, and J.-P. Zigrand (2011). Balance sheet capacity and endogenous risk.

Financial Stability Board (2017). Artificial intelligence and machine learning in financial services Market developments and financial stability implications. Technical Report November, Financial Stability Board (FSB).

Friedman, M. and A. Schwartz (1963). *A monetary history of the United States : 1867-1960*. Princeton Univ. Press.

Goodhart, C. A. E. (1974). Public lecture at the reserve bank of Australia.

Laeven, L. and F. Valencia (2018). Systemic banking crises revisited. Technical report, IMF. 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.

Minsky, H. (1992). The financial instability hypothesis. Yale University, Mimeo.

Roberts, R. (2013). *Saving the City: the great financial crisis of 1914*. Oxford University Press.

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis (2017). Mastering the game of go without human knowledge. *Nature*. www.nature.com/nature/journal/v550/n7676/full/nature24270.html.

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

Watkins, C. (2008, Sept). Selective breeding analysed as a communication channel: Channel capacity as a fundamental limit on adaptive complexity. In *2008 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pp. 514–518.