BANK OF ENGLAND

Staff Working Paper No. 937
Comparing minds and machines: implications for financial stability

Marcus Buckmann, Andy Haldane and
Anne-Caroline Hüser

August 2021

Staff Working Papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the Bank of England or to state Bank of England policy. This paper should therefore not be reported as representing the views of the Bank of England or members of the Monetary Policy Committee, Financial Policy Committee or Prudential Regulation Committee.

BANK OF ENGLAND

Staff Working Paper No. 937
Comparing minds and machines: implications for financial stability

Marcus Buckmann,[1] Andy Haldane[2] and Anne-Caroline Hüser[3]

Abstract

Is human or artificial intelligence more conducive to a stable financial system? To answer this question, we compare human and artificial intelligence with respect to several facets of their decision-making behaviour. On that basis, we characterise possibilities and challenges in designing partnerships that combine the strengths of both minds and machines. Leveraging on those insights, we explain how the differences in human and artificial intelligence have driven the usage of new techniques in financial markets, regulation, supervision, and policy making and discuss their potential impact on financial stability. Finally, we describe how effective mind-machine partnerships might be able to reduce systemic risks.

Key words: Artificial intelligence, machine learning, financial stability, innovation, systemic risk.

JEL classification: B4, C45, C55, C63, C81.

(1) Bank of England. Email: marcus.buckmann@bankofengland.co.uk
(2) Bank of England. Email: andy.haldane@bankofengland.co.uk
(3) Bank of England. Email: anne-caroline.huser@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. This paper was prepared for publication in: *Oxford Review of Economic Policy*, Vol. 37 No. 3, Autumn 2021. We would like to thank David Bholat, Adam Brinley-Codd, Shiv Chowla, Mohammed Gharbawi, Andreas Joseph, Jack Meaning, Paul Nahai-Williamson, Daniel Susskind, Oliver Thew, Nick Vause, and David Vines for helpful comments and suggestions.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email enquiries@bankofengland.co.uk

© Bank of England 2021
ISSN 1749-9135 (on-line)

1 Introduction

Is human or artificial intelligence more conducive to a stable financial system? By comparing the building blocks of artificial and human intelligence, we identify key differences in the abilities of humans and machines. This enables us to assess which operations in the financial system can best be executed by humans and which by computers and which, crucially, are best served as the two acting in partnership.

Applications of artificial intelligence (AI) in the financial sector are increasing rapidly. In a joint survey by the Bank of England and UK's Financial Conduct Authority (Bank of England and Financial Conduct Authority, 2019), two thirds of the financial services industry report using machine learning (ML). The median firm uses ML applications in two business areas and expects to double the number of applications within three years. As Russell (2021) points out, AI is not a new technology. However, several recent developments have increased its number of applications and their effectiveness.

First, computers have become much more powerful, both in terms of processing power and memory capacity, and large-scale cloud computing has become easily accessible. Second, there has been an explosion in data that is systemically recorded and that learning algorithms can exploit. Third, open-source software has reduced the development cost of ML applications, often condensing the training of ML models to a few lines of code. Finally, new algorithmic innovations have allowed for ML solutions to challenging problems. The adoption of AI has been driven by both these supply factors and the demand-side needs of financial firms (Financial Stability Board, 2017).

The term AI is often used to describe computers that take actions associated with human cognitive functions. However, by defining AI anthropomorphically, contrasting fundamental differences between machine and human intelligence becomes difficult. We therefore use the AI definition of Russell and Norvig (2016) "the designing and building of intelligent agents that receive precepts from the environment and take actions that affect that environment", where "[...] an intelligent agent takes the best possible action in a situation."

The applications of AI discussed in this paper are all examples of *narrow AI*, an AI system that has been trained to master specific tasks such as playing board games or driving cars. In contrast, *Artificial General Intelligence* (AGI) or *strong AI* describes a hypothetical universal algorithm that can learn and act in any environment (Russell and Norvig, 2016). In a recent survey of 352 AI experts, the aggregated forecast suggests that an AGI that outperforms humans in any task will be created within 45 years with a 50% probability (Grace et al., 2018). But the current state of AI is very far from AGI (Russell, 2021) and the path to it is unknown (Chollet, 2019; Marcus, 2018). Nonetheless, if it ever materialises, AGI has the potential to completely reshape our world (Tegmark, 2017; Russell, 2019).

ML is a field in AI research using computer programs to learn from data without being explicitly programmed what to learn (Samuel, 1959). Not all AI is based on ML. For example, a TicTacToe agent that plays according to a fixed set of rules that were implemented by a human is an AI *expert system* that is not capable of learning. In contrast, a ML TicTacToe player learns not to lose the game from a database of games or by playing the game repeatedly. In more complex problems, programming a rule-based AI that anticipates all possible states in a system can quickly become infeasible.

ML approaches are often divided into three categories: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, a ML algorithm learns an association between input and output pairs from data. The output is the supervisory signal, which the model learns to infer from the inputs. For example, a computer learns from a dataset containing characteristics of borrowers (input) and whether they defaulted (output). The model is then used to predict whether a future borrower is likely to default.

In unsupervised learning, a ML model discovers patterns in the input data. No output (supervisory signal) is given. For instance, an unsupervised ML model clusters borrowers in groups according to their similarity or identifies anomalous data points with respect to the entire data set.

In reinforcement learning, a computer agent tries to identify the sequence of actions in an environment that lead to the maximum reward. The agent needs to explore the environment to learn the best strategy. For example, a reinforcement learning agent learns to master a board game by playing against itself many times. The system is designed such that the agent receives a reward when winning the game and is penalised when losing it. The agent is only programmed to strive for the reward—but is not equipped with any strategies before it starts to learn.

Many of the milestone successes of AI in recent years, such as a reinforcement learning agent playing Go (Silver et al., 2018), supervised models detecting skin cancer from images (Esteva et al., 2017), or the unsupervised language models that can write coherent text (Brown et al., 2020), are based on deep artificial neural networks, which are also referred to as *deep learning*. By passing the input data from layer to layer in the network, it is represented in an increasingly more abstract way. Provided enough data points are available from which to learn meaningful representations, deep learning models can extract signals from unstructured high dimensional data such as images, text, and sound. This is a task more traditional ML approaches struggle with.

In many AI applications, humans and machines work together to deliver a system that is both stable and efficient. The financial system is no exception. For the financial system, stability means the ability to absorb shocks while preventing disruption to the real economy (Schinasi, 2004). Many excellent papers have comprehensively reviewed applications of AI in the financial sector and several studies have focused on their implications for financial stability (Financial Stability Board, 2017; Danielsson et al., 2019; Gensler and Bailey, 2020). Our paper focusses on those applications—from trading and lending to regulation and policy making—which best illustrate some of the strengths and weaknesses of humans and machines.

For example, the use of AI in algorithmic trading has the clear advantage of speed of execution and the ability to consider a vast array of information simultaneously (Nevmyvaka et al., 2006; Bacoyannis et al., 2018). Further, an algorithmic trader is less likely to make mistakes or biased, irrational decisions (Jain et al., 2015). But most AI agents are confined to a narrow optimisation problem and do not have a broader understanding of the environment or the ability to reason about causality. This means that, when past data does not reflect the current economic situation, the model may behave unexpectedly leading to instability in trading behaviour. When it comes to lending, new data sources and ML algorithms extracting predictive patterns can improve credit risk models (Khandani et al., 2010; Lessmann et al., 2015), potentially lower borrowing costs and widening access to credit (Berg et al., 2019; Frost et al., 2020). However, an AI system optimising decisions of individual firms—for example, to cut lending in a downturn—could worsen system-wide stability by aggravating financial and economic crises.

In regulation, AI can help in automating processes and reduce reporting errors due to human deficiencies and increase the speed of regulation (Arner et al., 2016). While AI tools can collect and oversee rich information about a firm (Proudman, 2020), an AI system cannot understand a firm in the context of the economy, and thus can only support but not replace a supervisor. AI can also be used to inform economic policy making by improving macroeconomic models (Döpke et al., 2017; Zheng et al., 2020). However, compared to human policy makers, AI cannot be creative and propose new solutions to a problem outside the narrow framework it was calibrated on. This suggests that, while AI can help to calibrate existing policy tools, it will not invent an innovative tool that is more than a combination of existing measures.

Our review of the literature suggests that in many AI applications, a human-centred partnership is optimal, one which harnesses the analytical power of AI to collect and process data, while leaving a role for human judgements as an overlay, and safety-catch, on decision-making. However, in those applications where human intervention is either costly or impossible, extra care needs to be taken in the design and supervision of AI, to preserve trust and avoid systemic risks.

The paper is structured as follows. Section 2 compares human and artificial intelligence, both on a physical basis and in their decision-making processes and discusses the design of effective human-AI partnerships. Section 3 reviews applications of AI in financial markets and their implications for financial stability, including partnerships between human and artificial agents in trading, portfolio management, and lending. Section 4 reviews AI applications in supervision, regulation, and policy making. Section 5 concludes.

2 Comparing human and artificial intelligence

Analogies between the human brain and computers are long-standing. Alan Turing famously used an anthropomorphic analogy when defining his theoretical model of a computer in 1948: “A man provided with paper, pencil, and rubber, and subject to strict discipline, is in effect a universal machine” (Turing, 1948; Proudfoot, 2011). In cognitive science, the human brain has been regarded as an information processing device—a computer—for decades. Proponents of this analogy argue that, just like a computer, the brain processes (sensory) inputs using computations and representations to create an output, a thought or behaviour (Pinker, 2003; Boden, 2008).

However, many scholars question the usefulness of this metaphor for understanding the human brain because representations and computation in the brain and in computers are so different. Unlike a computer, our brain is not operating on abstract representations of the world. Rather, using our body, we directly interact with the world. We see a cat holistically and not a matrix of numbers that represents the image of a cat. When catching a frisbee, we do not solve differential equations but simply keep a constant angle when running towards it (Gigerenzer and Gray, 2017). Our brain is not an empty general purpose information processing machine (Epstein, 2016) like a a computer. Over our long evolutionary history, the brain has been *optimised* to control our body and survive in our world, which a computer-based AI cannot simply imitate.

At the physical level, computer hardware is faster than the human brain and can store information more effortlessly. The neurons in the brain fire between 1–200 times in a second (Bryant, 2013), whereas the computer fires at 2-3 gigahertz, which is millions times faster. But speed is not decisive for intelligence. "[I]t just means you get the wrong answer more quickly" (Russell, 2019). The first 100,000 digits of π just require a megabyte on a computer hard drive, but are almost impossible to memorise for humans.[1]

Our understanding of the neurological processes of human intelligence are limited. Despite innovations in techniques to measure brain activities, we still do not understand how the brain processes information and memorises (Adolphs, 2015). So to have an informed comparison of human and artificial intelligence, it is not enough to compare their physical basis; we need to compare their behaviours.

2.1 Reasoning and decision making

The human brain is better prepared for some tasks than others. Our evolutionary history has prepared the brain for the physical world. Humans are equipped with highly optimised unconscious sensory procedures, which makes object recognition tasks trivial and allows us to efficiently learn from only a few data points (Lake et al., 2015). A child can readily identify a giraffe having seen only a few pictures. In contrast, ML requires immense amounts of data and substantive computational resources for visual object recognition. The landmark successes of ML in recent years that learned from visual inputs (such as playing games, recognising faces or detecting cancer) were all based on millions of data points and millions of parameters calibrated during the optimisation.

Similarly, humans can predict a Jenga game because we intuitively understand gravity and statics. ML models can also master these kinds of task, but only by learning from thousands of examples of the specific game (Lerer et al., 2016). Usually, ML models are not equipped with any theoretical or intuitive understanding about the world (Russell, 2019). They have to learn the physics of a Jenga game by watching the tower collapse over and over again. Humans are not only intuitive physicists but also intuitive psychologists. We have a model of others' intents and beliefs (Premack and Woodruff, 1978) and expect them to act in a goal-directed way (Spelke and Kinzler, 2007). This means

[1] Akira Haraguchi recited these digits from memory in 16 hours in 2006. But he encoded them as a collection of stories rather than abstract numbers (Bellos, 2015).

we can naturally navigate social interactions, something an assistant robot would need to learn.

On the other hand, data-driven optimisation and, more generally, mathematical computations are much harder for the human mind than for an AI agent. These problems have only become relevant in our very recent history and our biological brain appears to be less optimised for these (Moravec, 1988). For example, often even experts fail to make the correct decision, despite data clearly pointing towards it. Goldberg (1970) has shown that diagnoses made by physicians can be improved by using a simple linear regression rather than letting the physicians aggregate pieces of information. Grove and Meehl (1996) and Grove et al. (2000) showed that simple algorithmic rules outperform human judgments in many different problems including the prediction of cancer recovery, parole violation and college grades. Yntema and Torgerson (1961) stated 60 years ago that "Men and computers could cooperate more efficiently [...] if a man could tell the computers how he wanted decisions made, and then let the machine make the decisions for him [p. 20]".

Given these computational limitations, human decisions are often not consistent. For example, given two identical medical records a doctor may make different diagnoses. Inconsistencies are also reinforced by boredom, fatigue and distractions (Goldberg, 1970). In contrast, algorithmic predictions are usually deterministic—if the input does not change, nor does the prediction.[2] Deterministic decisions make decision processes more reliable. However, they can also increase their vulnerability to attacks. Markose (2021) argues that the deterministic predictability of AI systems poses a substantial threat to their resilience. In contrast, humans react adaptively to attacks by changing their behaviour.

In the following we take a closer look at several dimensions along which human and artificial agents differ in their decision making.

2.1.1 Causal understanding

Our ability to understand intuitively cause and effect relationships has been considered a key breakthrough in the cognitive evolution of humans (Stuart-Fox, 2015). Lake et al. (2017) describes human learning as building a model of the world to explain what is and imagine what can and cannot happen. As intuitive physicists with rich prior knowledge about our world, we automatically understand causal relationships which are not obvious for computers. For instance, we understand that rainfall is the cause of a wet street.

Causal reasoning is not a prerequisite for accurate predictions. Even without any theoretical causal knowledge, a pattern recognition model can make accurate predictions. Consider precipitation forecasting. The standard models simulate the atmospheric physics and are computationally extremely costly to train and thus cannot produce forecasts for very short horizons. Agrawal et al. (2019) propose a theory-free ML approach to predict rainfall one hour ahead that compares favourably with the physical model.

[2] However, calibrating ML models often is not a deterministic process due to the stochastic nature of optimisation processes and the fact that a learned solution is often not certifiably optimal.

The only information the model is learning from are the pixels of radar images showing precipitation in the area. This works because data are abundant and the problem is stable.

However, if data are scarce, predictions without causal knowledge is challenging. For these scenarios, Einhorn and Hogarth (1985) and Armstrong et al. (2015) have stressed the importance of theoretical causal models for selecting relevant variables as predictors. Data driven prediction models that exploit many variables for prediction are more susceptible to changes in the data than models informed by theory that use only those variables that are believed to be relevant in the future (see also Nestor et al. (2019); Geirhos et al. (2020)). An example of this phenomenon is the Google Flu Trend model (Ginsberg et al., 2009), which nowcasted doctor-related flu visits from the frequency of Google searches correlated with the flu. While the model performed well after being introduced, its performance deteriorated over time and the model was eventually abandoned. Lazer et al. (2014) showed that a simple autoregressive model outperformed Google Flu Trends. Other empirical studies also suggest that simpler models can be more robust to changes over time in the data generating process than more complex ML approaches (Hand, 2006; Lee et al., 2017; Mushava and Murray, 2018; Wang and Perkins, 2019).

2.1.2 Transparency of the decision process

Humans can reason about their decision process and explain it to others, although their explanation may not always reflect the true drivers of the decision. Nisbett and Wilson (1977) argue that we have limited access to our cognitive processes and therefore are not aware of the factors driving our behaviour. Empirical evidence suggests that we often form an opinion quickly based on few pieces of information, even when we believe we have considered many pieces of information (Klein and O'Brien, 2018; Dhami and Ayton, 2001). Emotions, consciously or unconsciously, also contribute to our decision making (Lerner et al., 2015). Collectively, this shows how complex and often non-transparent human decision making can be.

ML models have been criticised for their lack of transparency. In the strictest sense, a model is transparent if a human can compute the predictions from the inputs and parameters of the model in a short amount of time (Lipton, 2018). While this is possible for a linear regression with few parameters or other simple models, it is not for more complex ML approaches. The output of a black box model, such as a neural network, is a complex function usually not expressible in a simple formula. While a neural network can accurately tell apart pictures of dogs and cats it is not clear how it does that exactly. There exist several frameworks for *explainable AI* (Ribeiro et al., 2016; Lundberg and Lee, 2017; Sundararajan et al., 2017; Joseph, 2020) which explain the decisions of ML models by decomposing individual predictions into the contribution of the individual variables. With these frameworks, we can depict marginal effects of individual variables and interactions, but there is no way to concisely present all complex interactions of variables a model might have learned such that its inner workings are fully understood.

2.1.3 Biased decisions

The human mind is often exposed to biases. These have been extensively studied (Tversky and Kahneman, 1974; Kahneman, 2011). Examples are the tendency to focus on information that confirms preconceptions (confirmation bias, Nickerson (1998)), or the tendency to be overconfident in the accuracy of our judgments (Pallier et al., 2002). Cognitive biases do not necessarily imply worse decisions as they may make decisions more robust under uncertainty (Gigerenzer and Brighton, 2009). Even in the ML literature, researchers have tried to mimic human biases to improve the performance of ML models (Taniguchi et al., 2018).

Algorithmic models can also be biased when trained on data which itself is biased. For example, ML models have been found to have racial and gender biases (Caliskan et al., 2017; Lee, 2018; Buolamwini and Gebru, 2018; Vincent, 2018). In the literature, a trade-off between the accuracy and fairness of an algorithm has been observed (Kleinberg et al., 2016; Zafar et al., 2017) but by controlling for the biases in training data, this trade-off can disappear (Wick et al., 2019).

Biases are easier to uncover from simple prediction models than from complex black box ML models. But even the latter can be rigorously queried. For instance, one can test how the prediction of a model changes, when only changing a particular attribute of a data point that the model is potentially biased against, such as the race of a loan applicant. In contrast, proving that a particular decision made by a human was biased is almost impossible as their decisions are not deterministic and humans are often not aware of the biases influencing their decision (Greenwald and Banaji, 1995; Payne and Gawronski, 2010).

2.1.4 Setting goals

An AI system does not set itself a goal. Rather it learns the behaviour that optimises a specific goal, the *payoff function*. For example, learning to play a board game means maximising the number of wins against an opponent and detecting skin cancer means minimising the number of false positive and false negative diagnoses. For many narrow problems such as these, formulating the payoff function is straightforward. But for more complex problems, with multiple competing objectives, this can be challenging.

First, it can be difficult to make a problem quantifiable (Lipton, 2018). Consider, for example, an AI system tasked with designing a fair tax policy. How do you quantify fairness?

Second, AI payoff functions often need to be flexible to be in accordance with our values in all situations the AI will encounter. Consider an autonomous vehicle that runs into a moral dilemma. Should it swerve into an elderly pedestrian to save three children playing on the street? This decision should be guided by our human moral compass. But for rarely observed situations such as this one, human preferences can be difficult to measure given intrapersonal inconsistencies and well as individual and cultural variations (Awad et al., 2018).

Third, in complex reinforcement learning problems, where an AI agent operates in many different states, it can be very challenging for the designer of the AI to anticipate negative side effects that an AI agent may cause when maximising the payoff (Hadfield-Menell et al., 2017; Hadfield-Menell and Hadfield, 2019). For example, a robot learned to move from A to B by being rewarded when reaching B. But on its way from A to B it destroys a vase that was on the direct path because the designer forgot to specify that this behaviour should be penalised (Amodei et al., 2016). Further, an AI agent may learn to game the payoff function. For instance, a cleaning robot, rewarded for reducing the mess, might learn to litter the floor before cleaning to increase its reward. Similarly, if rewarded for not seeing mess, it might close its eyes or look away (Amodei et al., 2016).

Humans are good at solving abstract goals because they can decompose a complex task into several smaller tasks. When planning a dinner we are aware that we need to choose a meal, buy the food, prepare and serve it. In contrast, today's AI will struggle to solve such a high-level task because identifying relevant subgoals from data is challenging for a machine that does not understand the world as we do (Russell, 2019).

2.1.5 Creativity

The success of the AI AlphaZero (Silver et al., 2018) in playing the board games Go and chess led to very enthusiastic reviews of its *creative* style (Sadler and Regan, 2019; Ingle, 2018). It surprised players by making innovative moves they had not expected and had not been seen in human players. But is this really creative behaviour? The renowned cognitive scientist Margaret Boden (Boden, 2004; Du Sautoy, 2020) defines creativity as "the ability to come up with ideas or artefacts that are new, surprising, and valuable". Following this definition, AlphaZero can be considered creative. However, Kelly (2019) argues that AlphaZero is not creative as it only performs optimisation inside the narrow framework of the rules of the board game. It does not reflect any *out-of-the-box* thinking.

Skills acquired by a narrow ML model often transfer poorly to other domains (Marcus, 2018; Chollet, 2019). Humans, on the other hand, can use their diverse experiences and knowledge to solve new problems. Further, with their diverse motives, including their curiosity to learn and explore and their striving to improve skills and knowledge (Reiss, 2002; Tay and Diener, 2011), they are naturally inclined to try something new.

State-of-the-art ML models for natural language processing give an idea how much a model can learn when it is able to explore a very rich and diverse world. The GPT-3 model (Brown et al., 2020) is a neural network with 175 billion free parameters. It was trained on a single task: predict the next word in a sentence in a very diverse text corpus, including archives of books and the whole Wikipedia. The model was tested on diverse tasks and shows impressive performance. It can, among other skills, write stories, poems, translate, do simple arithmetic, answer quiz questions and write news articles without ever been explicitly trained for these tasks. This shows how much an extremely large model with a simple payoff function can achieve when trained on huge amounts of diverse data. But it does not perform well in all tasks; in some, more specialised language models do better. Further, its writing is often incoherent and repetitive and the model struggles with some reasoning questions such as “If I put cheese into the fridge, will it melt?”. This suggest that GPT-3 has only a limited understanding of the world (Marcus and Davis, 2020) despite having read much more about it than any human could in a lifetime. Directly transferring a human knowledge base (such as solid substances melt at high temperatures) to an AI model could improve its understanding of the world. However, for most AI approaches, there do not exist straightforward approaches to do this (Marcus, 2018).

Understanding also ensures robustness. We understand other people even if we miss hearing some words and we can read bad handwriting because we can reconstruct meaning from the context. On the other hand, ML models can often be easily tricked by slightly altering the data. These alterations may remain unnoticed to the human eye or ear but hackers can use these to control the output of a model (Akhtar and Mian, 2018; Yuan et al., 2019).

2.2 Partnerships of human and artificial intelligence

Bringing together the insights from the above discussion, we will now characterise possibilities and challenges in designing a partnership combining the strengths of both minds and machines. An example of a successful human-AI partnership comes from the world of chess. In 2005, an online chess tournament encouraged players to play with the help of computers against other human-computer teams. Eight years after the computer Deep Blue had defeated the world champion Kasparov, teams of human players and ordinary computers generally dominated powerful supercomputers that were playing alone, showing the power of human-AI partnerships. Surprisingly, the tournament was not won by an experienced grand master and his computer but by two computer savvy amateur players that used three PCs. This shows, Kasparov (2017) argues, that the efficiency of the coordination between humans and computers is decisive for success, rather than their skills individually.

Designing an effective partnership is challenging. A key problem is that humans do not seem to evaluate the capabilities of AI in an unbiased way. On the one hand, humans tend to discount advice based on algorithmic systems (Önkal et al., 2009) and tolerate errors made by algorithms less than errors made by humans (Dietvorst et al., 2015). Experimental work by Dietvorst et al. (2018) suggests that this *algorithmic aversion* can be overcome if humans are not forced to follow the predictions of the algorithm but retain the control by having the possibility to modify its predictions. Even when the room for modification is only small, participants are more likely to use algorithmic predictions and are more satisfied with the decision process.

On the other hand, humans may too willingly hand over tasks to an AI system, “prematurely ceding authority to them far beyond their competence” (Dennett, 2020). High confidence in an AI model can reduce the focus on other relevant cues the AI does not track. Carr (2015) gives the example of a doctor focusing on the area of a medical image that was highlighted by the AI, thereby overlooking a malign abnormality in another part of the image.

A reason for overestimating the capabilities of AI is our tendency to anthropomorphise, i.e. attributing human traits or intentions to non-human entities (Brooks et al., 1999; Dennett, 2009; Proudfoot, 2011). For example, users of language generating models such as GPT-3 are especially prone to anthropomorphism as these models use our language almost perfectly which makes the user think it has the same level of understanding of the world as humans have. Even users of the trivial chatbot ELIZA from the 1960s, which only posed open-ended questions and paraphrased the conversational partner, ascribed human traits to it (Weizenbaum, 1976).

Acknowledging the different strengths of human and machine decision-making, Jarrahi (2018) argues for a partnership where AI systems are "designed with the intention of augmenting, not replacing human contributions" with AI harnessing its analytical power to collect and process data and the human to make judgements based on the collated evidence (see also Duan et al. (2019); Miller (2018)). In such human-centred partnerships, the AI is a tool for the human operator who retains total control.[3] For example, in the medical domain, where ML has proven to be effective for diagnostic purposes, Verghese et al. (2018) recommends that "clinicians should seek a partnership in which the machine predicts (at a demonstrably higher accuracy), and the human explains and decides on action." This illustrates the importance of transparent AI predictions, as it will not suffice to tell the patient "because the model said so".

Human-centred partnerships can be effective in many scenarios but not in those decision problems that require a higher degree of automation because manual labour is either too costly or too slow. In these cases, humans will need to monitor the AI which might fail or behave unexpectedly in situation not anticipated by the system designer. Unfortunately, humans are not naturally good at passive monitoring. If all intellectually engaging tasks are outsourced, their concentration drifts (Carr, 2015; Bainbridge, 1983) and a failure of the system may remain unnoticed (Endsley and Kiris, 1995; Gouraud et al., 2017). Further, humans tend to lose skills when tasks are transferred to a machine, which has, for instance, been shown for navigation (Dahmani and Bohbot, 2020) and flying (Ebbatson, 2009). To support the human supervisor, the AI needs to be programmed to provide meaningful signals, for instance by raising a flag when it senses a drift in the data (Amodei et al., 2016). Also, there need to be contingency plans in place for when the performance of the AI model deteriorates. Shutting down an AI agent without a replacement strategy cannot be an option in many applications. Large scale decision problems cannot be simply executed manually in the case of an AI failure.

3 AI applications in financial markets

We do not aim to comprehensively review AI applications in financial markets but rather focus on applications in trading, portfolio optimisation and lending that have implications for financial stability. In all three areas, using ML brings advantages due to the

[3] In that, the AI can be regarded as an extension of our mind (Clark and Chalmers, 1998; Hernández-Orallo and Vold, 2019). AI tools become seamlessly part of our cognitive processes, just like a pencil that we need when solving mathematical equations.

ability of ML models to make fast and accurate decisions, but at some potential cost. We can assess these costs and benefits through the lens of the differences between artificial and human intelligence.

Compared to trades executed by humans, the key advantages of algorithmic trading are the speed of execution and the ability to consider a vast array of information simultaneously for the decision. Further, an algorithmic trader is less likely to make mistakes or biased, irrational decisions (Jain et al., 2015). Given that, algorithmic trading has the potential to strengthen the information function of the financial system and lower transaction costs (Menkveld, 2016). For example, Chaboud et al. (2014) find that algorithmic trading in the foreign exchange market causes an improvement in price efficiency. Reinforcement learning methods seem especially apt for executing trades (Nevmyvaka et al., 2006; Bacoyannis et al., 2018; Wei et al., 2019). These models are capable of learning complex strategies and constantly adapt their strategy in response to the market.

Gu et al. (2020) demonstrate that ML methods also perform well in empirical asset pricing. Based on 920 variables, a ML forecast led to substantially higher gains to investors than a simple regression model. Bianchi et al. (2020) show that ML methods outperform conventional methods in predicting bond excess returns. In these applications, ML models can exploit complex associations in the data (e.g. nonlinearities and interactions) that a human would not be able to process manually. Automated investment services, often called *robo-advisors*, can lower the cost and increase the quality of investment advice (Liu et al., 2020). In contrast to human advisers that are known to be biased to sell products that come with a higher commission (Executive Office of the President, Council of Economic Advisors, 2015), robo-advisors are not biased per se.

ML can also improve credit risk models. Compared to linear models— the traditional approach to assess the creditworthiness of borrowers (Dastile et al., 2020)—ML models are more accurate in estimating default risk based on the credit history of borrowers (Khandani et al., 2010; Lessmann et al., 2015; Sirignano et al., 2016). The arguably more substantive advantage of ML is that it can detect patterns in rich alternative data that cannot easily be processed with simple models or swept through by humans. Fintech companies exploit data sources such as social media activity (Dorfleitner et al., 2016; Jagtiani and Lemieux, 2019) or other data users leave online (Berg et al., 2019). For example, the Fintech firm ZestFinance reported that it considered more than 10,000 diverse data points per loan applicant to estimate their creditworthiness compared to 10–15 data points on the applicant's credit history that classical lenders consider (Carney, 2013).

Using AI and ML in lending can reduce the costs and enhance the efficiency of financial services leading to lower fees and borrowing costs for customers. Further, new credit scoring models may help enable greater access to credit (Berg et al., 2019; Frost et al., 2020). For example, in China, Ant Financial provides millions of loans to small companies by estimating their default risk based on transaction data from the online retailer Alibaba, the parent company of AntFinancial (Zeng, 2018). For established banks, providing these small loans has not been profitable, given the limited credit history of the small companies and the costly manual assessment of their creditworthiness. Similarly, Frost et al. (2020) report that big technology firms serve unbanked borrowers in Argentina.

3.1 Risks stemming from using AI in financial markets

The financial market is a complex, dynamic domain in which many market participants interact with each other, and is influenced by economic, political, and societal events and developments. So it is very different from those narrow domains were AI agent reliably excel such as board games. Most risks AI poses in the financial sector stem from the fact that AI agents do not *understand* the world as humans do and thus will be ignorant of the societal context. For example, it will not understand what an election campaign or a terror attack are and what their economic effects could be.

3.1.1 Dynamic economy and shocks

If an abnormal economic situation arises, the strategy learned under stable economic conditions might no longer be appropriate (Yadav, 2015; Sherif, 2018). This is equivalent to introducing a new rule to a board game that the agent has no time to experiment with before being forced to make the next move. For example, in a survey of the UK banking sector, about 35% of the banks reported a decline in performance of their ML models due to the COVID crisis (Bholat et al., 2020). ML models may be especially susceptible to changes in the economy, if new alternative data sources are used that have not existed long enough to cover a full business cycle (Bazarbash, 2019). Pozen and Ruane (2019) argue that human asset managers with their intuition and general knowledge of the financial markets—two strengths the AI does not have—should take over when the present is unlike the past (see also Danielsson et al. (2019)).

Raman et al. (2020) empirically compared trades executed by humans and automated systems at the National Stock Exchange of India and show that human traders performed stronger during market stress. The authors observe that automated trading systems tend to reduce their participation in the market in turbulent times which can lead to feedback loops making the market more fragile. Yadav (2015) argues that it is only rational for automated trading systems to exit the market in response to a crisis instead of developing costly algorithms that deal with stressed conditions, as the respective data to calibrate the system is rare.

ML credit models may also be more sensitive to business cycles than traditional lending. Wang and Perkins (2019) show empirically that ML credit risk models are less robust to changes over time than a simple linear model. The former outperforms the latter only on shorter horizons. Further, in the face of an economic downturn, purely data-driven approaches may cut a large proportion of lending and thus magnify the downturn (Carstens, 2018; Bolton et al., 2016).

The opacity of ML-based trading systems makes it challenging to anticipate under which situations the model will fail and make irrational decisions. This will not be deducible from a look at thousands of parameters. In contrast, humans experience stress in unexpected or dangerous situations and effectually communicate their emotions and worries, for example by their facial expression. Computers on the other hand, do not have emotions and will only transmit warnings when explicitly programmed to do so. Further, research suggest that the self-report of a model's confidence is often not accurate (Guo et al., 2017; Hendrycks and Gimpel, 2016; Lakshminarayanan et al., 2017), which means, the model does not know when it is likely to be wrong.

3.1.2 Herding and collusion

If many market participants follow similar trading strategies, their correlated behaviour may generate financial stability risks.

Human traders may follow different schools of thought, have individual preferences and motives and are subject to different biases. ML methods on the other hands are likely to produce similar strategies when they are based on the same data and payoff function (Danielsson et al., 2019; Gensler and Bailey, 2020).

One example is the Quant meltdown in August 2007, where some of the most successful hedge funds in the industry suffered record losses. The losses seemed to be concentrated almost exclusively among quantitatively managed hedge funds which had similar portfolios (Lo, 2016). Chaboud et al. (2014) find evidence that strategies of algorithmic traders—not necessarily based on ML—are more correlated than those by human traders and Jain et al. (2016) show that the introduction of a high-speed trading platform by the Tokyo Stock Exchange increased several measures of system risks due to correlated trading.

Trading models might learn to coordinate with each other, thereby increasing the interconnectedness of the market (OECD, 2017). Calvano et al. (2020) showed in a simulation study that reinforcement algorithms learn tacitly to collude, without being programmed to do so and without communication. This illustrates how a simple payoff function—maximizing profit—can lead to undetected and unintended consequences in a dynamic system. An empirical analysis of the German gasoline retail market (Assad et al., 2020) suggest that algorithms in this market might have learned collusive strategies. Trading algorithms that maximise investment growth can also learn spoofing, a prohibited practise of placing orders to manipulate other market participants without actually executing them (Martínez-Miranda et al., 2016; Allison, 2016).

3.1.3 Susceptibility to attacks

AI algorithms may not only engage in unethical behaviour, but are also susceptible to adversarial attacks (Danielsson et al., 2019). The deterministic nature of their decisions make them more vulnerable to being gamed than humans who are better able to detect attacks and change their behaviour in response to them (Markose, 2021). Arnoldi (2016) provides evidence for algorithmic trading models being misled by targeted attacks of human traders. Nehemya et al. (2020) describe how an attacker can control the decisions of an algorithmic trader by manipulating the data the algorithm bases its decisions on. In their experiment, the attacks were successful even if the inner workings of the trading algorithm were not known to the attacker and the manipulation of the data were small and thus unlikely to be noticed by a monitoring system.

3.2 Human-AI interactions and partnerships

Next, we discuss how partnerships between human and artificial intelligence might operate in financial markets to reduce the risks of AI applications discussed above. We focus on applications in trading, portfolio management, and lending.

3.2.1 Trading

Human control over automatic trading is difficult because of the different speed at which decisions are made. A human needs 150–200 milliseconds to respond to a simple stimulus. In this time a trading system can have made thousands of trades. Research by Johnson et al. (2012) showed that stock price movements fall into two different regimes: one at a slower speed, where humans and machines interact, one at higher speed at which exclusively automated systems interact with each other and humans are unable to intervene or respond. Thus a trader supervising an automated system has to rely on aggregated data at a higher level of abstraction that comes with a time lag (Baxter and Cartlidge, 2013). The time for detecting a problem and addressing it can be considerable. In 2012, it took Knight Capital 45 minutes to stop an erroneous trading software, which burned US$ 440 million in that time (Nanex, 2012).

A human supervisor being degraded to passively monitor an AI trading system will find it difficult to focus on this tiring, monotonic task (Baxter and Cartlidge, 2013). Further, when only passively monitoring, the supervisors' mental models of the trading system will likely degrade over time and with that their ability to diagnose problems (Baxter and Cartlidge, 2013). The short lifespan of trading algorithms (Adler, 2012) and the fact that adaptive ML agents change their strategies over time makes monitoring even more difficult. Thus, there need to be measures in place to ensure that the monitoring is effective.

On top of that, there are several factors that can facilitate a better human-AI partnership in trading. First and most material, Haldane (2012b) discusses the possibility of introducing resting periods between trades. This would make monitoring of automated trades easier and allow for a more timely detection of problems. Equally important, it restores the possibility of collaboration and communication between human and AI traders at the same speed. Cartlidge and Cliff (2018) found in an experimental study that doing that can increase the efficiency of the market.

Second, regulators need to understand the underlying algorithms and need to ensure that those humans monitoring the AI agent while it is trading do understand its inner workings as well. These requirement are not straightforward if the algorithm is a complex black box that bases its decision on diverse data.

Finally, financial firms need to define when and how to stop an automated trading system and need to plan how to quickly replace a failing algorithm to avoid liquidity shocks (Lee, 2020).

3.2.2 Portfolio management

Robo-advisors directly interact with customers. This interaction needs to be carefully designed as it influences the decision of the customers. Several studies have shown that anthropomorphising robo-advisors—e.g. by giving them a face and name and making them converse—can increase the perceived trustworthiness and the amount of money the customers are willing to invest in the recommendations of the AI (Morana et al., 2020; Hodge et al., 2018; Adam et al., 2019; Hildebrand and Bergner, 2020). One could think of more unethical nudges that designers implement to steer the customers towards certain financial products. However, this behaviour can be detected and prevented by the regulator, if the underlying algorithm, data and interface is made available to auditors (Baker and Dellaert, 2017). In contrast, it is more difficult to hold human advisors accountable for individual cases of misleading or self-serving financial advice.

Compared to a robo-advisor, its human counterpart is better positioned to coach the clients and help with long-term financial planning. Aware of the limits, firms offer hybrid advice, which showcases an effective partnership of the different strength of human and artificial advisors. The automated investment is guided by algorithms whereas humans take over the relationship management (Baker and Dellaert, 2017; Miller, 2018) and can build trust. Surveys suggest that customers prefer hybrid advice over human or computer only services (Financial Planning Association, 2017; Accenture, 2017).

3.2.3 Lending

ML models may cut lending during a crisis magnifying the downturn. This is in contrast to relationship banking, where financial institutions establish long-term cooperation with their customers based on soft information that are difficult to measure and use in a computational model (Lončarski and Marinč, 2020). Several studies show that relationship banking protects customers from cutting back lending in response to a crisis (Bolton et al., 2016; Beck et al., 2018). Similarly, it also protects banks from runs during a crisis (Iyer and Puri, 2012). Relationship lending only works when the loan officers have flexibility and discretion—something a deterministic algorithmic system does not have. However, as Jakšič and Marinč (2019) points out, this discretion can also lead to unfairness or increased risk taking when lending decision are biased by career concerns (Cole et al., 2015), the mood of the loan officer (Cortés et al., 2016) or the characteristics of the applicant (Ravina, 2008).

The different strength of (human) relationship lending and automated lending suggest that these two approaches should complement each other in a partnership of human and artificial intelligence (Bartoli et al., 2013; Jakšič and Marinč, 2019). Mocetti et al. (2017) show synergetic effects of this partnership across a set of 300 Italian banks. With a higher adoption of information technology, loan officers have more time to use soft information and are given more decision-making autonomy. By including human decision making based on soft information instead of exclusively relying on quantitative information processed by similar AI systems, lending behaviour can be effectively diversified, which can reduce the degree of correlation in the credit market.

4 AI applications in regulation, supervision and policy making

The fundamental differences between human and artificial intelligence not only help to understand how applications of AI affect financial stability risks in the financial market, but also in regulating and supervising the financial sector, as well as in economic policy making more generally.

4.1 Regulation and supervision

Financial regulation is complex (Haldane, 2013; Gai et al., 2019; Herring, 2018). For example, in 2017, the prudential rules for banks operating in the UK contained over 720,000 words (Amadxarif et al., 2019), more than Tolstoy's War and Peace. The cost of regulation has grown substantially with the increase in regulation and regulatory reporting after the financial crisis in 2008 (Haldane, 2012a; Arner et al., 2016).

In response to this development, technological regulatory approaches (RegTech) are being developed to automate regulatory processes. RegTech reduces manual work and the associated reporting errors due to human deficiencies. By streamlining the reporting and monitoring of regulatory data, the speed of the regulatory process can greatly increase and the cost can be reduced. Regulators can obtain real-time insights into the markets with the help of AI methods. This can help to shift the focus of regulation from reacting to problems after the fact to anticipating and preventing problems before they happen (Arner et al., 2016).

4.1.1 Machine readable regulatory rules

A step towards automated regulatory reporting is making regulatory rules machine-readable. With that, Micheler and Whaley (2019) suggest, a regulatory change "could become as simple as installing a software update." However, this is challenging because of the differences in human and artificial intelligence. Current regulation in natural language is often abstract and vague (Amadxarif et al., 2019). While humans naturally deal with this kind of information, the current capabilities of AI do not allow flexible reasoning on abstract and vague concepts. Thus, making regulation machine readable requires recoding existing regulation into precise rules.[4] However, precise technical rules are often less general and flexible than rules expressed in legal terms of our natural language (Micheler and Whaley, 2019). Exactly reproducing legal regulation in code may even prove infeasible so that the translation of ambiguous legal terms into computer code implies changes in regulation.

Micheler and Whaley (2019) state that "a good quality regulatory regime achieves more than technical compliance" (p.365) and quote Black (2015): "Conduct should be in accordance with the principles and purposes of the rules, not the letter." Technical compliance to a set of precise rules might obscure the overarching aims and motives of the regulator. For instance, the *fundamental rules* of the PRA rulebook (e.g. "A

[4] Language models such as GPT-3 could help with this task; it has been shown that GPT-3 is able to transform text describing simple programs into the respective computer code (Vu, 2020).

firm must conduct its business with integrity", "A firm must act in a prudent manner") communicate generalised expectations of the regulator that are difficult to formalise using computer code.

Further, precise regulation is easier to game than a more flexible regulation with a human interpreter of legal rules in the loop. Once the rules are written in code, an AI system might even assist in regulatory arbitrage, i.e. finding loopholes. A human interpreter, on the other hand, could always judge these practices as breaking fundamental rules such as those above.

4.1.2 Supporting supervisors

AI applications can help supervisors to work more efficiently and better discover risks buried in data. Searching for information in texts is one of the most time-consuming manual activities (Proudman, 2020) and supervisors do not have the capacity to carefully study all reports and minutes that are potentially relevant. AI can help to identify relevant documents by scanning for topics or producing summaries of large documents such as management and board information provided by the firms (El-Haj et al., 2019).

Supervisors use formal rating systems to assess a firm's financial soundness to ensure high standards and avoid subjective judgements and biases. Predictive ML can help to complement, refine and back-test these models. Several studies have shown that financial distress and failure of banks can be predicted using ML methods on the firms' financial information (Carmona et al., 2019; Gogas et al., 2018; Suss and Treitel, 2019). Further, using ML to extract information from texts such as news, auditor reports, or management statements can improve the prediction of distress compared to relying on financial variables alone (Cerchiello et al., 2017; Matin et al., 2019). These models are most useful to the supervisor if they are transparent and explain their predictions and thus can be critically assessed and compared to the supervisors' own judgement.

4.2 Policy making

AI can inform economic policy making and help detecting financial stability risks by improving forecasts and macroeconomic models. Due to AI's limited creativity and ability to think out-of-the-box, it is unlikely to invent innovative policy tools that are more than a combination of existing measures.

4.2.1 Monitoring the financial sector and the economy

AI can be used for forecasting and systemic risk assessment. A major advantage of ML models over human reasoning (and classical statistical models) is their ability to identify patterns from big data of different types including text and images. For example, Nyman et al. (2018) observe that a narrative consensus and high excitement in texts about financial markets can signal systemic stress. Kalamara et al. (2020) use ML on newspaper texts to forecast macroeconomic variables such as GDP, inflation and unemployment, finding that these models perform significantly better than those only built on macroeconomic indicators. Another advantage of alternative data sources is their timeliness. While official statistical indicators (CPI, GDP, unemployment, etc.) are subject to a reporting delay, policy makers can, for example, nowcast economic developments by analysing satellite images (Huang, 2018).

Even when making forecasts based on numeric data, ML models can be superior to more standard statistical models. Examples are the prediction of financial crises (Alessi and Detken, 2018; Bluwstein et al., 2020; Tölö, 2020) and recessions (Ng, 2014; Döpke et al., 2017). Further, ML model can easily deal with many variables and infer the kind of relationships between variables (including nonlinearities and interactions) from data. This can protect predictive models from a biased manual selection of key predictors and assumed forms of relationships between variables. For instance, it was widely known before the global financial crises of 2008 that a high growth in debt is often a crucial risk factor for crises (Minsky, 1976; Kindleberger and Manias, 1978). Models embodying credit booms would have signalled an increased risk of a bust in the run-up to the crisis in 2008 (Jordà et al., 2011). However, the dangers of unsustainable credit growth were explained away by stating that "this time is different" (Reinhart and Rogoff, 2009), including arguments that financial innovations such as securitisation decreased the probability of a crash (da Silva and von Peter, 2018). As standard ML models do not easily allow the modeller to impose theoretical assumptions such as causal relationships, they are less susceptible to modellers imposing their prior beliefs into the model. On the other hand, when little data is available—e.g. when predicting rare crises—theoretical knowledge on which variables are important can be essential for obtaining meaningful forecasts.

History never repeats itself and while common vulnerabilities may be discovered by analysing patterns in historic data, AI cannot anticipate unpredictable shocks, such as the COVID-19 crisis in 2020. Even with data on previous epidemics, producing a point estimate of the economic impact of a pandemic is unlikely to be meaningful, as these extreme events do not follow a statistical distribution that allows meaningful forecasts (Taleb et al., 2020). In this and other situations of *radical uncertainty* (Kay and King, 2020) statistical models as well as ML models are likely to give a false sense of confidence.

We cannot expect humans to be much better at predicting shocks like COVID-19 but their judgement is likely to be superior when responding to these extreme situations. Humans' rich historical, contextual, and theoretical understanding helps us to deal with these unexpected situations (Danielsson et al., 2019). As the language model GPT-3 illustrates, just feeding an AI system with abundant historic knowledge does not make it a problem solver on a par with humans in situations of radical uncertainty.

4.2.2 Simulating the economy

Standard macroeconomic models such as dynamic stochastic general equilibrium (DSGE) models are used by policy makers to understand the economy. They often assume that agents in the economy have rational or near-rational expectations and are homogeneous. While mathematically convenient, these simplifying assumptions have been criticised because they fail to capture real-world behaviours. This is supported by the fact that these models did not foresee the global financial crisis of 2008 (Stiglitz, 2018).

Agent-based models, on the other hand, can more easily account for heterogeneous agents with bounded rationality (Haldane and Turrell, 2019; Dosi et al., 2020). Using ML techniques can make these agent-based models more realistic by making them adaptively learn in their environment to optimise their individual goals (Busoniu et al., 2008). For example, researchers from the cloud-based software company Salesforce simulated an economy to design taxation policies (Zheng et al., 2020). Using reinforcement learning, individual agents with different skills learn to collect resources, trade with other agents, or build houses in a two dimensional toy world. At the same time, a policy making agent learns a tax policy which optimises a trade-off between equality and productivity. The abundance of economic data in our digital economy can help to refine the simulated economy (Engler, 2020). Data on the real behaviour of human agents can be exploited to create more human-like agents in the simulation.

More generally, machine learning can be used to estimate macroeconomic models that are computationally so demanding to solve that traditional approaches rely on simplifying assumptions to make the estimation tractable (Hill et al., 2021; Fernández-Villaverde and Guerrón-Quintana, 2021).

4.3 A human-centred partnership

AI is very unlikely to replace regulators, supervisors, or policy makers in the near future, for reasons of robust decision-making in the face of uncertainty. However, they stand to profit from AI tools that can assist them in making their decisions (Wall, 2018; Proudman, 2020; Bauguess, 2017; Lagarde, 2018).

A fully-automated regulatory process based on machine-readable rules would bear great risks and human involvement is required to interpret legal requirements in a flexible manner and keep a focus on the overarching goal of a stable financial system. This point has been made more generally in the legal literature, suggesting AI cannot take over the role of lawyers and judges but can rather support them in their decision making (Markovic, 2019; Wu, 2019).

AI tools can help to spot patterns and outliers in regulatory trends as well as in macroeconomic data. But bigger and more timely data does not necessarily make the financial system by itself more stable. The more data, the higher the level of abstraction at which humans can parse it and the greater the reliance on automated tools for preprocessing, aggregating data and identifying relevant anomalies or trends. Our discussion of the limits of ML methods in Section 2 does not suggest that AI will master this task without problems. The value of a general economic and causal understanding of the system, together with the brittleness of AI models when they lack robustness to unforeseen changes in the data, all speak to a partnership with human judgement being a core ingredient. That is why judgement-based supervision lies at the heart of the Bank of England's, and other supervisors', approach.

In an effective partnership, supervisors, regulators, and policy makers must not be threatened in their autonomy and should always be able to overrule an algorithmic assessment. Further, it is crucial that supervisors are made aware of common misconceptions when interacting with AI such as anthropomorphism: even if an AI model is able to summarise a text accurately, it cannot *understand* patterns as humans do—e.g. explaining why a drop in capital ratio is not worrying for one firm but is for another. Supervisors understand the financial health of a firm because they have a holistic view on a firm from regulatory information, financial statements and texts produced by the firm and meetings with the firm in the context of the current economic conditions. AI can supplement, but is unlikely any time soon to supplant that information set and decision-making capacity.

5 Conclusion

The increasing use of AI is reshaping the financial system. The implications of the manifold AI applications for financial stability depend on *how* they are implemented and regulated. An AI application is not just an algorithmic model that can be considered detached of its designers and those parties it interacts with. Rather, it should be analysed holistically. In this study, we have discussed how partnerships based on the complementary strengths of human and artificial intelligence can reduce financial stability risks for several applications of AI in the financial markets, regulation and policy making.

While we have identified different risks that can arise from AI applications in the financial sector, we try to summarise succinctly these here by analysing their effects on trust in the financial system. Our financial system is based on trust (Tonkiss, 2009; Morris and Vines, 2014). Financial firms need to gain the trust of the people by acting in a trustworthy fashion (O'Neill, 2016). After the financial crisis in 2008, Joseph Stiglitz stated that "financial markets hinge on trust, and that trust has eroded" (Stiglitz, 2008).

Mayer et al. (1995) discusses three dimensions of trustworthiness: ability, benevolence and integrity. We have provided ample evidence that AI agents have the ability to take over relevant tasks in the market—often performing better than humans. However, ability implies a high degree of dependability (Aitken et al., 2020) and we have shown many examples, where AI fails to deliver that including unpredictable behaviour when the environment is changing and the fact that some AI applications can easily be gamed and exploited by adversaries.

Benevolence is the degree to which an agent is believed to do good to the person affected by its decision. Integrity means that an agent's actions are aligned with the values of that person. These aspects of trustworthiness are a major challenge for AI systems as they do not know what is good and what is not and have no inherent values. This makes them prone to unethical behaviour. The fact that AI agents collude or may discriminate are examples for this. It is the role of the designer of the AI to ensure that it consistently acts for the good of the people and incorporates the required values into the decision making. Using unbiased data to calibrate the models and defining an appropriate payoff function are necessary but often not sufficient to achieve this goal. Further, human values are not easily translatable into a quantitative measure that an

AI model can optimise. This, and more generally the fact that AI today is not able to pursue abstract goals, limits its applications if benevolence and integrity are not be undermined. For example, AI system should not be responsible for risk management in a financial firm, design economic policies, or autonomously supervise a financial firm.

Transparency plays a role for trust as well. It helps to assess the benevolence and integrity of an AI (Aitken et al., 2020). Neither the trustworthiness of the developer nor that of the algorithm are sufficient for an AI system to be trustworthy. Rather the trustworthiness has to be reviewed holistically. This argument is in line with Ananny (2016), who argues that the ethics of an algorithm can only be evaluated in the assemblages of computer code, human practices, and norms.

References

Accenture (2017) "The new face of wealth management." https://www.accenture.com/us-en/insights/capital-markets/new-face-wealth-management.

Adam, Martin, Jonas Toutaoui, Nicolas Pfeuffer, and Oliver Hinz (2019) "Investment decisions with robo-advisors: the role of anthropomorphism and personalized anchors in recommendations," in *Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden.*

Adler, Jerry (2012) "Raging bulls: How Wall Street got addicted to light-speed trading," *Wired Magazine*, Vol. 20, No. 9.

Adolphs, Ralph (2015) "The unsolved problems of neuroscience," *Trends in Cognitive Sciences*, Vol. 19, No. 4, pp. 173–175.

Agrawal, Shreya, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey (2019) "Machine Learning for Precipitation Nowcasting from Radar Images," *arXiv preprint arXiv:1912.12132*.

Aitken, Mhairi, Ehsan Toreini, Peter Carmichael, Kovila Coopamootoo, Karen Elliott, and Aad van Moorsel (2020) "Establishing a social licence for financial technology: Reflections on the role of the private sector in pursuing ethical data practices," *Big Data & Society*, Vol. 7, No. 1, pp. 1–15.

Akhtar, Naveed and Ajmal Mian (2018) "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, Vol. 6, pp. 14410–14430.

Alessi, Lucia and Carsten Detken (2018) "Identifying excessive credit growth and leverage," *Journal of Financial Stability*, Vol. 35, pp. 215–225. Network models, stress testing and other tools for financial stability monitoring and macroprudential policy design and implementation.

Allison, Ian (2016) "When intelligent algorithms start spoofing each other, regulation becomes a science," *International Business Times*. https://www.ibtimes.co.uk/machine-learning-markets-when-intelligent-algorithms-start-spoofing-each-other-regulation-becomes-1567986.

Amadxarif, Zahid, James Brookes, Nicola Garbarino, Rajan Patel, and Eryk Walczak (2019) "The language of rules: textual complexity in banking reforms," *Bank of England Staff Working Paper*, No. 834.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané (2016) "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565.*

Ananny, Mike (2016) "Toward an ethics of algorithms: Convening, observation, probability, and timeliness," *Science, Technology, & Human Values*, Vol. 41, No. 1, pp. 93–117.

Armstrong, J. Scott, Kesten C. Green, and Andreas Graefe (2015) "Golden rule of forecasting: Be conservative," *Journal of Business Research*, Vol. 68, No. 8, pp. 1717–1731.

Arner, Douglas W/, Janos Barberis, and Ross P. Buckey (2016) "FinTech, RegTech, and the reconceptualization of financial regulation," *Northwestern Journal of International Law & Business*, Vol. 37, No. 3, pp. 371–413.

Arnoldi, Jakob (2016) "Computer algorithms, market manipulation and the institutionalization of high frequency trading," *Theory, Culture & Society*, Vol. 33, No. 1, pp. 29–52.

Assad, Stephanie, Robert Clark, Daniel Ershov, and Lei Xu (2020) "Algorithmic pricing and competition: Empirical Evidence from the German retail gasoline market," *CESifo Working Paper*, No. 8521.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan (2018) "The moral machine experiment," *Nature*, Vol. 563, No. 7729, pp. 59–64.

Bacoyannis, Vangelis, Vacslav Glukhov, Tom Jin, Jonathan Kochems, and Doo Re Song (2018) "Idiosyncrasies and challenges of data driven learning in electronic trading," *arXiv preprint arXiv:1811.09549.*

Bainbridge, Lisanne (1983) "Ironies of automation," in *Analysis, design and evaluation of man-machine systems. Proceedings of the IFAC/IFIP/IFORS/IEA Conference, Baden-Baden, Federal Republic of Germany, 27–29 September 1982*, pp. 129–135.

Baker, Tom and Benedict Dellaert (2017) "Regulating robo advice across the financial services industry," *Iowa Law Review*, Vol. 103, pp. 713–750.

Bank of England and Financial Conduct Authority (2019) "Machine learning in UK financial services," https://www.bankofengland.co.uk/-/media/boe/files/report/2019/machine-learning-in-uk-financial-services.pdf.

Bartoli, Francesca, Giovanni Ferri, Pierluigi Murro, and Zeno Rotondi (2013) "SME financing and the choice of lending technology in Italy: Complementarity or substitutability?" *Journal of Banking & Finance*, Vol. 37, No. 12, pp. 5476–5485.

Bauguess, Scott W (2017) "The role of big data, machine learning, and AI in assessing risks: a regulatory perspective," *Machine Learning, and AI in Assessing Risks: A Regulatory Perspective (June 21, 2017). SEC Keynote Address: OpRisk North America.*

Baxter, Gordon and John Cartlidge (2013) "Flying by the seat of their pants: What can high frequency trading learn from aviation?" in *Proceedings of the 3rd International Conference on Application and Theory of Automation in Command and Control Systems*, pp. 56–65.

Bazarbash, Majid (2019) "Fintech in financial inclusion: machine learning applications in assessing credit risk," *International Monetary Fund Working Paper*, No. 19/109.

Beck, Thorsten, Hans Degryse, Ralph De Haas, and Neeltje Van Horen (2018) "When arm's length is too far: Relationship banking over the credit cycle," *Journal of Financial Economics*, Vol. 127, No. 1, pp. 174–196.

Bellos, Alex (2015) "He ate all the pi : Japanese man memorises π to 111,700 digits," *The Guardian*. https://www.theguardian.com/science/alexs-adventures-in-numberland/2015/mar/13/pi-day-2015-memory-memorisation-world-record-japanese-akira-haraguchi.

Berg, Tobias, Valentin Burg, Ana Gombović, and Manju Puri (2019) "On the rise of FinTechs: Credit scoring using digital footprints," *The Review of Financial Studies*, Vol. 33, No. 7, pp. 2845–2897, 09.

Bholat, David, Mohammed Gharbawi, and Oliver Thew (2020) "The impact of Covid on machine learning and data science in UK banking," *Bank of England Quarterly Bulletin, Q4*.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni (2020) "Bond risk premia with machine learning," *The Review of Financial Studies*, Vol. 34, No. 2, pp. 1046–1089.

Black, Julia (2015) "Regulatory styles and supervisory strategies," in Niamh Moloney, Eilis Ferran, and Jennifer Payne eds. *The Oxford Handbook of Financial Regulation*: Oxford University Press, pp. 217–253.

Bluwstein, Kristina, Marcus Buckmann, Andreas Joseph, Miao Kang, Sujit Kapadia, and Özgür Simsek (2020) "Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach," *Bank of England Staff Working Paper*, No. 848.

Boden, Margaret A. (2004) *The creative mind: Myths and mechanisms*: Psychology Press.

——— (2008) *Mind as machine: A history of cognitive science*: Oxford University Press.

Bolton, Patrick, Xavier Freixas, Leonardo Gambacorta, and Paolo Emilio Mistrulli (2016) "Relationship and transaction lending in a crisis," *The Review of Financial Studies*, Vol. 29, No. 10, pp. 2643–2676.

Brooks, Rodney Allen et al. (1999) *Cambrian intelligence: The early history of the new AI*: MIT press.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al. (2020) "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33*.

Bryant, Astra (2013) "Ask a Neuroscientist! - What is the synaptic firing rate of the human brain?," http://www.neuwritewest.org/blog/4541.

Buolamwini, Joy and Timnit Gebru (2018) "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 77–91.

Busoniu, Lucian, Robert Babuska, and Bart De Schutter (2008) "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 38, No. 2, pp. 156–172.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (2017) "Semantics derived automatically from language corpora contain human-like biases," *Science*, Vol. 356, No. 6334, pp. 183–186.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello (2020) "Artificial intelligence, algorithmic pricing, and collusion," *American Economic Review*, Vol. 110, No. 10, pp. 3267–97.

Carmona, Pedro, Francisco Climent, and Alexandre Momparler (2019) "Predicting failure in the US banking sector: An extreme gradient boosting approach," *International Review of Economics & Finance*, Vol. 61, pp. 304–323.

Carney, Michael (2013) "Flush with 20M from Peter Thiel, ZestFinance is measuring credit risk through non-traditional big data," https://pando.com/2013/07/31/flush-with-20m-from-peter-thiel-zestfinance-is-measuring-credit-risk-through-non-traditional-big-data.

Carr, Nicholas (2015) *The glass cage: Where automation is taking us*: Random House.

Carstens, Agustín (2018) "Big tech in finance and new challenges for public policy." Speech at the FT Banking Summit.

Cartlidge, John and Dave Cliff (2018) "Modelling Complex Financial Markets Using Real-Time Human–Agent Trading Experiments," in *Computing in Economics and Finance*, pp. 35–69, Springer.

Cerchiello, Paola, Giancarlo Nicola, Samuel Ronnqvist, and Peter Sarlin (2017) "Deep learning bank distress from news and numerical financial data," *arXiv preprint arXiv:1706.09627*.

Chaboud, Alain P., Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega (2014) "Rise of the machines: Algorithmic trading in the foreign exchange market," *The Journal of Finance*, Vol. 69, No. 5, pp. 2045–2084.

Chollet, François (2019) "The Measure of Intelligence," *arXiv preprint arXiv:1911.01547*.

Clark, Andy and David Chalmers (1998) "The extended mind," *Analysis*, Vol. 58, No. 1, pp. 7–19.

Cole, Shawn, Martin Kanz, and Leora Klapper (2015) "Incentivizing calculated risk-taking: Evidence from an experiment with commercial bank loan officers," *The Journal of Finance*, Vol. 70, No. 2, pp. 537–575.

Cortés, Kristle, Ran Duchin, and Denis Sosyura (2016) "Clouded judgment: The role of sentiment in credit origination," *Journal of Financial Economics*, Vol. 121, No. 2, pp. 392–413.

da Silva, Luiz Awazu Pereira and Goetz von Peter (2018) "Financial instability: can Big Data help connect the dots?." Speech at the Nonth European Central Bank Statistic Conference on "20 years of ESCB statistics: what's next?", Frankfurt am Main, 11 July 2018.

Dahmani, Louisa and Véronique D. Bohbot (2020) "Habitual use of GPS negatively impacts spatial memory during self-guided navigation," *Scientific Reports*, Vol. 10, No. 1, pp. 1–14.

Danielsson, Jon, Robert Macrae, and Andreas Uthemann (2019) “Artificial intelligence and systemic risk: Market developments and potential financial stability implications,” *Available at SSRN 3410948*.

Dastile, Xolani, Turgay Celik, and Moshe Potsane (2020) “Statistical and machine learning models in credit scoring: A systematic literature survey,” *Applied Soft Computing*, Vol. 91.

Dennett, Daniel (2009) “Intentional systems theory,” in Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter eds. *The Oxford Handbook of Philosophy of Mind*: Oxford University Press Oxford, pp. 339–350.

Dennett, Daniel C (2020) “The age of post-intelligent design,” in Steven S. Gouveia ed. *The Age of Artificial Intelligence: An Exploration*: Vernon Press, pp. 27–62.

Dhami, Mandeep K. and Peter Ayton (2001) “Bailing and jailing the fast and frugal way,” *Journal of Behavioral Decision Making*, Vol. 14, No. 2, pp. 141–168.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey (2015) “Algorithm aversion: People erroneously avoid algorithms after seeing them err.,” *Journal of Experimental Psychology: General*, Vol. 144, No. 1, p. 114.

——— (2018) “Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them,” *Management Science*, Vol. 64, No. 3, pp. 1155–1170.

Döpke, Jörg, Ulrich Fritsche, and Christian Pierdzioch (2017) “Predicting recessions with boosted regression trees,” *International Journal of Forecasting*, Vol. 33, No. 4, pp. 745–759.

Dorfleitner, Gregor, Christopher Priberny, Stephanie Schuster, Johannes Stoiber, Martina Weber, Ivan de Castro, and Julia Kammler (2016) “Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms,” *Journal of Banking & Finance*, Vol. 64, pp. 169–187.

Dosi, Giovanni, Mauro Napoletano, Andrea Roventini, Joseph E Stiglitz, and Tania Treibich (2020) “Rational heuristics? Expectations and behaviors in evolving economies with heterogeneous interacting agents,” *Economic Inquiry*, Vol. 58, No. 3, pp. 1487–1516.

Du Sautoy, Marcus (2020) *The Creativity Code: Art and Innovation in the Age of AI*: Harvard University Press.

Duan, Yanqing, John S. Edwards, and Yogesh K. Dwivedi (2019) “Artificial intelligence for decision making in the era of Big Data—evolution, challenges and research agenda,” *International Journal of Information Management*, Vol. 48, pp. 63–71.

Ebbatson, Matthew (2009) “The loss of manual flying skills in pilots of highly automated airliners,” Ph.D. dissertation.

Einhorn, Hillel J. and Robin M. Hogarth (1985) “Prediction, diagnosis, and causal thinking in forecasting,” in George Wright ed. *Behavioral Decision Making*: Springer, pp. 311–328.

El-Haj, Mahmoud, Paul Rayson, Martin Walker, Steven Young, and Vasiliki Simaki (2019) “In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse,” *Journal of Business Finance & Accounting*, Vol. 46, No. 3-4, pp. 265–306.

Endsley, Mica R. and Esin O. Kiris (1995) "The out-of-the-loop performance problem and level of control in automation," *Human Factors*, Vol. 37, No. 2, pp. 381–394.

Engler, Alex (2020) "Can AI model economic choices?" *Brookings Institute AI Governance Series*. https://www.brookings.edu/research/can-ai-model-economic-choices/.

Epstein, Robert (2016) "The empty brain," *Aeon, May 18, 2016*. https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun (2017) "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, Vol. 542, No. 7639, pp. 115–118.

Executive Office of the President, Council of Economic Advisors (2015) "The effects of conflicted investment advice on retirement savings." https://obamawhitehouse.archives.gov/sites/default/files/docs/cea_coi_report_final.pdf.

Fernández-Villaverde, Jesús and Pablo A Guerrón-Quintana (2021) "Estimating DSGE models: Recent advances and future challenges," *Annual Review of Economics*, Vol. 13.

Financial Planning Association, Investopedia (2017) "High-tech and high-touch: Investors make the case for converging automated investing platforms and financial planning," http://i.investopedia.com/dimages/graphics/fpa_research_report_v207.pdf.

Financial Stability Board (2017) "Artificial intelligence and machine learning in financial services: Market developments and financial stability implications," *FSB Publication, November*.

Frost, Jon, Leonardo Gambacorta, Yi Huang, Hyun Song Shin, and Pablo Zbinden (2020) "BigTech and the changing structure of financial intermediation," *Economic Policy*, Vol. 34, No. 100, pp. 761–799, 01.

Gai, Prasanna, Malcolm Kemp, Antonio Sánchez Serrano, Isabel Schnabel et al. (2019) "Regulatory complexity and the quest for robust regulation," *European Systemic Risk Board Reports of the Advisory Scientific Committee*, No. 8.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann (2020) "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, Vol. 2, No. 11, pp. 665–673.

Gensler, Gary and Lily Bailey (2020) "Deep learning and financial stability," *Available at SSRN 3723132*.

Gigerenzer, Gerd and Henry Brighton (2009) "Homo heuristicus: Why biased minds make better inferences," *Topics in Cognitive Science*, Vol. 1, No. 1, pp. 107–143.

Gigerenzer, Gerd and Wayne D. Gray (2017) "A simple heuristic successfully used by humans, animals, and machines: The story of the RAF and Luftwaffe, hawks and ducks, dogs and frisbees, baseball outfielders and sidewinder missiles-oh my!," *Topics in Cognitive Science*, Vol. 9, No. 2, pp. 260–263.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant (2009) "Detecting influenza epidemics using search engine query data," *Nature*, Vol. 457, No. 7232, pp. 1012–1014.

Gogas, Periklis, Theophilos Papadimitriou, and Anna Agrapetidou (2018) "Forecasting bank failures and stress testing: A machine learning approach," *International Journal of Forecasting*, Vol. 34, No. 3, pp. 440–455.

Goldberg, Lewis R (1970) "Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences.," *Psychological Bulletin*, Vol. 73, No. 6, pp. 422–432.

Gouraud, Jonas, Arnaud Delorme, and Bruno Berberian (2017) "Autopilot, mind wandering, and the out of the loop performance problem," *Frontiers in Neuroscience*, Vol. 11, No. 541.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans (2018) "When will AI exceed human performance? Evidence from AI experts," *Journal of Artificial Intelligence Research*, Vol. 62, pp. 729–754.

Greenwald, Anthony G. and Mahzarin R. Banaji (1995) "Implicit social cognition: attitudes, self-esteem, and stereotypes.," *Psychological Review*, Vol. 102, No. 1, pp. 4–27.

Grove, William M. and Paul E. Meehl (1996) "Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy.," *Psychology, Public Policy, and Law*, Vol. 2, No. 2, p. 293.

Grove, William M., David H. Zald, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson (2000) "Clinical versus mechanical prediction: a meta-analysis," *Psychological Assessment*, Vol. 12, No. 1, pp. 19–30.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020) "Empirical asset pricing via machine learning," *The Review of Financial Studies*, Vol. 33, No. 5, pp. 2223–2273.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger (2017) "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330.

Hadfield-Menell, Dylan and Gillian K Hadfield (2019) "Incomplete contracting and AI alignment," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 417–422.

Hadfield-Menell, Dylan, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan (2017) "Inverse reward design," in *Advances in Neural Information Processing Systems 30*.

Haldane, Andrew G. (2012a) "The Dog and the Frisbee." Jackson Hole Speech given at the Federal Reserve Bank of Kansas City's 36th Economic Policy Symposium 'The Changing Policy Landscape'.

——— (2012b) "The race to zero," in Joseph E. Stiglitz, R. Gordon, and J. Fitoussi eds. *The Global Macro Economy and Finance*: Palgrave Macmillan, pp. 245–270.

——— (2013) "Capital discipline." Speech given at the American Economic Association, Denver, 9 January.

Haldane, Andrew G. and Arthur E. Turrell (2019) "Drawing on different disciplines: macroeconomic agent-based models," *Journal of Evolutionary Economics*, Vol. 29, No. 1, pp. 39–66.

Hand, David J. (2006) "Classifier technology and the illusion of progress," *Statistical Science*, Vol. 21, No. 1, pp. 1–14.

Hendrycks, Dan and Kevin Gimpel (2016) "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*.

Hernández-Orallo, José and Karina Vold (2019) "AI extenders: the ethical and societal implications of humans cognitively extended by AI," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 507–513.

Herring, Richard J. (2018) "The evolving complexity of capital regulation," *Journal of Financial Services Research*, Vol. 53, No. 2-3, pp. 183–205.

Hildebrand, Christian and Anouk Bergner (2020) "Conversational robo advisors as surrogates of trust: onboarding experience, firm perception, and consumer financial decision making," *Journal of the Academy of Marketing Science*, pp. 1–18.

Hill, Edward, Marco Bardoscia, and Arthur Turrell (2021) "Solving Heterogeneous General Equilibrium Economic Models with Deep Reinforcement Learning," *arXiv preprint arXiv:2103.16977*.

Hodge, Frank D., Kim I. Mendoza, and Roshan K. Sinha (2018) "The effect of humanizing robo-advisors on investor judgments," *Contemporary Accounting Research*, Vol. 38, No. 1, pp. 770–792.

Huang, Zheping (2018) "Doubtful of China's economic numbers? Satellite data and AI can help," *QUARTZ, August 16, 2018*. https://qz.com/1251912/doubtful-of-chinas-economic-numbers-satellite-data-and-ai-can-help/.

Ingle, Sean (2018) ""Creative" AlphaZero leads way for chess computers and, maybe, science," *Guardian, December 11, 2018*. https://www.theguardian.com/sport/2018/dec/11/creative-alphazero-leads-way-chess-computers-science.

Iyer, Rajkamal and Manju Puri (2012) "Understanding bank runs: The importance of depositor-bank relationships and networks," *American Economic Review*, Vol. 102, No. 4, pp. 1414–45.

Jagtiani, Julapa and Catharine Lemieux (2019) "The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform," *Financial Management*, Vol. 48, No. 4, pp. 1009–1029.

Jain, Ravindra, Prachi Jain, and Cherry Jain (2015) "Behavioral biases in the decision making of individual investors," *IUP Journal of Management Research*, Vol. 14, No. 3, pp. 7–27.

Jain, Pankaj K., Pawan Jain, and Thomas H. McInish (2016) "Does high-frequency trading increase systemic risk?" *Journal of Financial Markets*, Vol. 31, pp. 1–24.

Jakšič, Marko and Matej Marinč (2019) "Relationship banking and information technology: The role of artificial intelligence and FinTech," *Risk Management*, Vol. 21, No. 1, pp. 1–18.

Jarrahi, Mohammad Hossein (2018) "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Business Horizons*, Vol. 61, No. 4, pp. 577–586.

Johnson, Neil, Guannan Zhao, Eric Hunsader, Jing Meng, Amith Ravindar, Spencer Carran, and Brian Tivnan (2012) "Financial black swans driven by ultrafast machine ecology," *arXiv preprint arXiv:1202.1448.*

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor (2011) "Financial crises, credit booms, and external imbalances: 140 years of lessons," *IMF Economic Review*, Vol. 59, No. 2, pp. 340–378.

Joseph, Andreas (2020) "Parametric inference with universal function approximators," *arXiv preprint arXiv:1903.04209.*

Kahneman, Daniel (2011) *Thinking, fast and slow*: Macmillan.

Kalamara, Eleni, Arthur Turrell, George Kapetanios, Sujit Kapadia, and Chris Redl (2020) "Making text count for macroeconomics: What newspaper text can tell us about sentiment and uncertainty," *Bank of England Staff Working Paper*, No. 865.

Kasparov, Garry (2017) *Deep thinking: where machine intelligence ends and human creativity begins*: Hachette UK.

Kay, John and Mervyn King (2020) *Radical Uncertainty: Decision-Making Beyond the Numbers*: WW Norton & Company.

Kelly, Sean (2019) "A philosopher argues that an AI can't be an artist," *MIT Technology Review, February 21, 2019.*

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo (2010) "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, Vol. 34, No. 11, pp. 2767 – 2787.

Kindleberger, Charles P. and Panics Manias (1978) *Crashes: a history of financial crises*: New York: Basic Books.

Klein, Nadav and Ed O'Brien (2018) "People use less information than they think to make up their minds," *Proceedings of the National Academy of Sciences*, Vol. 115, No. 52, pp. 13222–13227.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016) "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807.*

Lagarde, Christine (2018) "Central banking and fintech: A brave new world," *Innovations: Technology, Governance, Globalization*, Vol. 12, No. 1-2, pp. 4–8.

Lake, Brenden M, Ruslan Salakhutdinov, and Joshua B Tenenbaum (2015) "Human-level concept learning through probabilistic program induction," *Science*, Vol. 350, No. 6266, pp. 1332–1338.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman (2017) "Building machines that learn and think like people," *Behavioral and Brain Sciences*, Vol. 40.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (2017) "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems 31.*

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014) "The parable of Google Flu: traps in big data analysis," *Science*, Vol. 343, No. 6176, pp. 1203–1205.

Lee, Nicol Turner (2018) "Detecting racial bias in algorithms and machine learning," *Journal of Information, Communication and Ethics in Society*, Vol. 16, No. 3, pp. 252–260.

Lee, Joseph (2020) "Access to Finance for Artificial Intelligence Regulation in the Financial Services Industry," *European Business Organization Law Review*, Vol. 21, No. 4, pp. 731–757.

Lee, Michael D., Gabrielle Blanco, and Nikole Bo (2017) "Testing take-the-best in new and changing environments," *Behavior Research Methods*, Vol. 49, No. 4, pp. 1420–1431.

Lerer, Adam, Sam Gross, and Rob Fergus (2016) "Learning physical intuition of block towers by example," in *Proceedings of The 33rd International Conference on Machine Learning*, pp. 430–438.

Lerner, Jennifer S, Ye Li, Piercarlo Valdesolo, and Karim S Kassam (2015) "Emotion and decision making," *Annual Review of Psychology*, Vol. 66, pp. 799–823.

Lessmann, Stefan, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas (2015) "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, Vol. 247, No. 1, pp. 124–136.

Lipton, Zachary C. (2018) "The mythos of model interpretability," *Queue*, Vol. 16, No. 3, pp. 31–57.

Liu, Che-Wei, Mochen Yang, and Ming-Hui Wen (2020) "Resilience in the storm: adaptive robo-advisors outperform human investors during the COVID-19 financial market turmoil," *Available at SSRN 3737821.*

Lo, Andrew W. (2016) "Moore's Law vs. Murphy's Law in the financial system: who's winning?" *BIS Working Paper*, No. 564.

Lončarski, Igor and Matej Marinč (2020) "The political economy of relationship banking," *Research in International Business and Finance*, Vol. 51.

Lundberg, Scott M and Su-In Lee (2017) "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30.*

Marcus, Gary (2018) "Deep learning: A critical appraisal," *arXiv preprint arXiv:1801.00631.*

Marcus, Gary and Ernest Davis (2020) "GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about," *MIT Technology Review.*

Markose, Sheri M (2021) "Novelty production and evolvability in digital genomic agents: logical foundations and policy design implications of complex adaptive systems," in Euel Elliott and L. Douglas Kiel eds. *Complex Systems in the Social and Behavioral Sciences: Theory, Method and Application*: University of Michigan Press.

Markovic, Milan (2019) "Rise of the robot lawyers," *Arizona Law Review*, Vol. 61, pp. 325–350.

Martínez-Miranda, Enrique, Peter McBurney, and Matthew J. W. Howard (2016) "Learning unfair trading: A market manipulation analysis from the reinforcement learning perspective," in *2016 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pp. 103–109, IEEE.

Matin, Rastin, Casper Hansen, Christian Hansen, and Pia Mølgaard (2019) "Predicting distresses using deep learning of text segments in annual reports," *Expert Systems with Applications*, Vol. 132, pp. 199–208.

Mayer, Roger C., James H. Davis, and F. David Schoorman (1995) "An integrative model of organizational trust," *Academy of Management Review*, Vol. 20, No. 3, pp. 709–734.

Menkveld, Albert J. (2016) "The economics of high-frequency trading: taking stock," *Annual Review of Financial Economics*, Vol. 8, pp. 1–24.

Micheler, Eva and Anna Whaley (2019) "Regulatory technology: replacing law with computer code," *European Business Organization Law Review*, pp. 349–377.

Miller, Steven M. (2018) "AI: Augmentation, more so than automation," *Asian Management Insights*, Vol. 5, No. 1, pp. 1–20.

Minsky, Hyman P. (1976) *John Maynard Keynes*: Springer.

Mocetti, Sauro, Marcello Pagnini, and Enrico Sette (2017) "Information technology and banking organization," *Journal of Financial Services Research*, Vol. 51, No. 3, pp. 313–338.

Morana, Stefan, Ulrich Gnewuch, Dominik Jung, and Carsten Granig (2020) "The effect of anthropomorphism on investment decision-making with robo-advisor vhatbots.," in *Proceedings of European Conference on Information Systems, Marrakech, Morocco.*

Moravec, Hans (1988) *Mind children: The future of robot and human intelligence*: Harvard University Press.

Morris, Nicholas and David Vines (2014) *Capital failure: Rebuilding trust in financial services*: Oxford University Press.

Mushava, Jonah and Michael Murray (2018) "An experimental comparison of classification techniques in debt recoveries scoring: Evidence from South Africa's unsecured lending market," *Expert Systems with Applications*, Vol. 111, pp. 35–50.

Nanex (2012) "Knightmare on Wall Street: what really happened, or how to test your new market making software and lose a pile of money, fast," http://www.nanex.net/aqck2/3522.html.

Nehemya, Elior, Yael Mathov, Asaf Shabtai, and Yuval Elovici (2020) "When Bots Take Over the Stock Market: Evasion Attacks Against Algorithmic Traders," *arXiv preprint arXiv:2010.09246*.

Nestor, Bret, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi (2019) "Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks," in *Proceedings of the Machine Learning for Healthcare Conference*, pp. 381–405.

Nevmyvaka, Yuriy, Yi Feng, and Michael Kearns (2006) "Reinforcement learning for optimized trade execution," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 673–680.

Ng, Serena (2014) "Boosting recessions," *Canadian Journal of Economics*, Vol. 47, No. 1, pp. 1–34.

Nickerson, Raymond S (1998) "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of General Psychology*, Vol. 2, No. 2, pp. 175–220.

Nisbett, Richard E and Timothy D Wilson (1977) "Telling more than we can know: verbal reports on mental processes," *Psychological Review*, Vol. 84, No. 3, pp. 231–259.

Nyman, Rickard, Sujit Kapadia, David Tuckett, David Gregory, Paul Ormerod, and Robert Smith (2018) "News and narratives in financial systems: exploiting big data for systemic risk assessment," *Bank of England Staff Working Paper*, No. 704.

OECD (2017) "Algorithms and collusion: Competition policy in the digital age," https://www.oecd.org/daf/competition/Algorithms-and-colllusion-competition-policy-in-the-digital-age.htm.

O'Neill, Onara (2016) "What is banking for." Remarks given at the Federal Reserve Bank of New York.

Önkal, Dilek, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock (2009) "The relative influence of advice from human experts and statistical methods on forecast adjustments," *Journal of Behavioral Decision Making*, Vol. 22, No. 4, pp. 390–409.

Pallier, Gerry, Rebecca Wilkinson, Vanessa Danthiir, Sabina Kleitman, Goran Knezevic, Lazar Stankov, and Richard D Roberts (2002) "The role of individual differences in the accuracy of confidence judgments," *The Journal of General Psychology*, Vol. 129, No. 3, pp. 257–299.

Payne, B. Keith and Bertram Gawronski (2010) "A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going," in Bertram Gawronski and Keith Payne eds. *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*: Guildford Press, pp. 1–15.

Pinker, Steven (2003) *How the mind works*: Penguin UK.

Pozen, Robert C. and Jonathan Ruane (2019) "What machine learning will mean for asset managers," *Havard Business Review*.

Premack, David and Guy Woodruff (1978) "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, Vol. 1, No. 4, pp. 515–526.

Proudfoot, Diane (2011) "Anthropomorphism and AI: Turing's much misunderstood imitation game," *Artificial Intelligence*, Vol. 175, No. 5-6, pp. 950–957.

Proudman, James (2020) "Supervisor-centred automation – the role of human-centred automation in judgement-centred prudential supervision." Speech to have been given at the conference "Impact of AI and Machine Learning on the UK economy" at the Bank of England.

Raman, Vikas, Michel A Robe, and Pradeep K. Yadav (2020) "Man vs. machine: liquidity provision and market fragility," http://dx.doi.org/10.2139/ssrn.3757848.

Ravina, Enrichetta (2008) "Love & loans: The effect of beauty and personal characteristics in credit markets," http://dx.doi.org/10.2139/ssrn.1101647.

Reinhart, Carmen M. and Kenneth S Rogoff (2009) *This time is different: Eight centuries of financial folly*: Princeton University Press.

Reiss, Steven (2002) *Who am I?: 16 basic desires that motivate our actions define our persona*: Penguin.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016) “” Why should I trust you?” Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Russell, Stuart (2019) *Human compatible: Artificial intelligence and the problem of control*: Penguin.

——— (2021) “The history and Future of AI,” *Oxford Review of Economic Policy*, Vol. 37, No. 3.

Russell, Stuart and Peter Norvig (2016) *Artificial intelligence: a Modern Approach*: Pearson, 3rd edition.

Sadler, Matthew and Natasha Regan (2019) *Game Changer: AlphaZero’s Groundbreaking Chess Strategies and the Promise of AI*: New in Chess.

Samuel, Arthur L. (1959) “Some studies in machine learning using the game of checkers,” *IBM Journal of Research and Development*, Vol. 3, No. 3, pp. 210–229.

Schinasi, Garry J (2004) “Defining financial stability,” *IMF working paper*, No. 04/187.

Sherif, Nazneen (2018) “Quants warn over flaws in machine learning predictions,” https://www.risk.net/derivatives/5451231/quants-warn-over-flaws-in-machine-learning-predictions.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel et al. (2018) “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play,” *Science*, Vol. 362, No. 6419, pp. 1140–1144.

Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke (2016) “Deep learning for mortgage risk,” *arXiv preprint arXiv:1607.02470*.

Spelke, Elizabeth S and Katherine D Kinzler (2007) “Core knowledge,” *Developmental Science*, Vol. 10, No. 1, pp. 89–96.

Stiglitz, Joseph (2008) “The fruit of hypocrisy,” *The Guardian, September 16, 2008*.

——— (2018) “Where modern macroeconomics went wrong,” *Oxford Review of Economic Policy*, Vol. 34, No. 1-2, pp. 70–106.

Stuart-Fox, Martin (2015) “The origins of causal cognition in early hominins,” *Biology & Philosophy*, Vol. 30, No. 2, pp. 247–266.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017) “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328.

Suss, Joel and Henry Treitel (2019) “Predicting bank distress in the UK with machine learning,” No. 831.

Taleb, Nassim Nicholas, Yaneer Bar-Yam, and Pasquale Cirillo (2020) "On single point forecasts for fat-tailed variables," *International Journal of Forecasting*.

Taniguchi, Hidetaka, Hiroshi Sato, and Tomohiro Shirakawa (2018) "A machine learning model with human cognitive biases capable of learning from small and biased datasets," *Scientific Reports*, Vol. 8, No. 1, pp. 1–13.

Tay, Louis and Ed Diener (2011) "Needs and subjective well-being around the world.," *Journal of Personality and Social Psychology*, Vol. 101, No. 2, pp. 354–365.

Tegmark, Max (2017) *Life 3.0: Being human in the age of artificial intelligence*: Allen Lane.

Tölö, Eero (2020) "Predicting systemic financial crises with recurrent neural networks," *Journal of Financial Stability*, Vol. 49, No. 100746.

Tonkiss, Fran (2009) "Trust, confidence and economic crisis," *Intereconomics*, Vol. 44, No. 4, pp. 196–202.

Turing, Alan Mathison (1948) "Intelligent machinery," National Physical Laboratory, UK, Technical Report.

Tversky, Amos and Daniel Kahneman (1974) "Judgment under uncertainty: Heuristics and biases," *Science*, Vol. 185, No. 4157, pp. 1124–1131.

Verghese, Abraham, Nigam H. Shah, and Robert A. Harrington (2018) "What this computer needs is a physician: humanism and artificial intelligence," *JAMA*, Vol. 319, No. 1, pp. 19–20.

Vincent, James (2018) "Amazon reportedly scraps internal AI recruiting tool that was biased against women," *The Verge*. https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report.

Vu, Kevin (2020) "Exploring GPT-3: A new breakthrough in language generation," https://www.kdnuggets.com/2020/08/exploring-gpt-3-breakthrough-language-generation.html.

Wall, Larry D. (2018) "Some financial regulatory implications of artificial intelligence," *Journal of Economics and Business*, Vol. 100, pp. 55–63.

Wang, J. Christina and Charles B. Perkins (2019) "How magic a bullet is machine learning for credit analysis? An exploration with FinTech lending data," *Working Paper Series, Federal Reserve Bank of Boston*, No. 19-16.

Wei, Haoran, Yuanbo Wang, Lidia Mangu, and Keith Decker (2019) "Model-based reinforcement learning for predictions and control for limit order books," *arXiv preprint arXiv:1910.03743*.

Weizenbaum, Joseph (1976) *Computer power and human reason: From judgment to calculation*: WH Freeman & Co.

Wick, Michael, Swetasudha Panda, and Jean-Baptiste Tristan (2019) "Unlocking fairness: a trade-off revisited," in *Advances in Neural Information Processing Systems 32*.

Wu, Tim (2019) "Will artificial intelligence eat the law? The rise of hybrid social-ordering systems," *Columbia Law Review*, Vol. 119, No. 7, pp. 2001–2028.

Yadav, Yesha (2015) "How algorithmic trading undermines efficiency in capital markets," *Vanderbilt Law Review*, Vol. 68, pp. 1607–1671.

Yntema, Douwe B. and Warren S. Torgerson (1961) "Man-computer cooperation in decisions requiring common sense," *IRE Transactions on Human Factors in Electronics*, Vol. 2, No. 1, pp. 20–26.

Yuan, Xiaoyong, Pan He, Qile Zhu, and Xiaolin Li (2019) "Adversarial examples: attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 30, No. 9, pp. 2805–2824.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi (2017) "Fairness constraints: mechanisms for fair classification," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA*, pp. 962–970.

Zeng, Ming (2018) "Alibaba and the future of business," *Harvard Business Review*, Vol. 96, No. 5, pp. 88–96.

Zheng, Stephan, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C. Parkes, and Richard Socher (2020) "The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies," *arXiv preprint arXiv:2004.13332*.