Intelligent financial system: how AI is transforming finance

I Aldasoro
BIS

L Gambacorta
BIS & CEPR

A Korinek
Univ. of Virginia & GovAI

V Shreeti
BIS

M Stein
Univ. of Oxford

June 20, 2024

Abstract

At the core of the financial system is the processing and aggregation of vast amounts of information into price signals that coordinate participants in the economy. Throughout history, advances in information processing, from simple book-keeping to artificial intelligence (AI), have transformed the financial sector. We use this framing to analyse how generative AI (GenAI) and emerging AI agents as well as, more speculatively, artificial general intelligence will impact finance. We focus on four functions of the financial system: financial intermediation, insurance, asset management, and payments. We also assess the implications of advances in AI for financial stability and prudential policy. Moreover, we investigate potential spillover effects of AI on the real economy, examining both an optimistic and a disruptive AI scenario. To address the transformative impact of advances in AI on the financial system, we propose a framework for upgrading financial regulation based on well-established general principles for AI governance.

JEL Codes: E31, J24, O33, O40.
Keywords: artificial intelligence, generative AI, financial system, financial institutions.

We thank seminar participants at ECB and GovAI seminars and Douglas Araujo, Fernando Perez-Cruz, Fabiana Sabatini, David Stankiewicz and Andrew Sutton for helpful comments and suggestions, and Ellen Yang for research assistance. Contact: Aldasoro (inaki.aldasoro@bis.org), Gambacorta (leonardo.gambacorta@bis.org), Korinek (anton@korinek.com), Shreeti (Vatsala.Shreeti@bis.org) and Stein (stein.merlin@gmail.com).

1 Introduction

Like the brain of a living organism, the financial system processes vast amounts of dispersed information and aggregates it into price signals that facilitate the coordination of all the players of the economy and guide the allocation of scarce resources. It not only enables the efficient flow of capital but also contributes to the economic system's overall health by managing risk, maintaining liquidity and supporting stability. Financial markets and intermediaries, when they function well, are a fundamental source of progress and welfare. Conversely, the role of financial policy and regulation is to correct instances of "brain malfunction" and to instead harness the intelligence of the financial system to enhance social welfare.

Processing all the necessary information and coordinating the actions of numerous participants in the economy is a notably complex problem. As the brain of the economy, financial markets and intermediaries have played this role for a long time. At any given point in time, their capacity to do so was shaped in large part by the information processing technology available. For example, over the years, technological advancements like telecommunications and the internet have continuously enhanced the capacity of financial markets to solve economic problems: a brain that can process more information more efficiently is better suited to solving increasingly complex tasks. It is then no surprise that financial markets have been a magnet for both cutting-edge information processing technology and for sophisticated human talent. Most recently, the information processing capabilities of the financial system have been enhanced by fast-paced advancements in artificial intelligence (AI).

In this paper, we describe the evolution of the financial sector through the prism of advancements in information processing, with a special focus on AI. We evaluate the opportunities and challenges created for the financial sector from different generations of AI, including machine learning (ML), generative AI (GenAI), and the emergence of AI agents. We also provide a discussion of the effects of AI on financial stability and the risk of real sector disruptions caused by AI. In light of these insights and increasing AI adoption we discuss the implications for the regulation of the financial sector.

Over the course of human history, the trajectory of methods for information processing – of which AI is part – has been closely linked with developments in commerce, trade and finance. In section 2 we describe this trajectory in detail. History is replete with examples of the financial system either sparking a change

in the arc of technological development, or itself being an early adopter of technology. From the abacus of the ancient Sumerians to double-entry book-keeping, the evolution of information processing technology and finance has often gone hand in hand. Over the last century, the most significant advance in the realm of information processing was the invention of computers in the 1950s. This allowed for the automation of many analytic and accounting functions that were very useful for the functioning of the financial system. As computational power increased over time, more sophisticated technologies emerged that allowed for the processing of non-traditional data, like machine learning models and, most recently, GenAI.

Each generation of information processing technology has had a large footprint on the financial system, and has opened new doors for efficiency and innovation. We discuss some of these in Section 3. In general, AI has enhanced the ability of the financial system to process information, analyse data, identify patterns and make predictions. Early rule-based systems were already deployed for automated trading and fraud detection. As technology advanced, the use cases of AI in the financial sector became more complex. Machine learning and deep learning models are used extensively in asset pricing, credit scoring, and risk analysis. While GenAI is nascent, the financial system is already adopting it to enhance back-end processing, customer support and regulatory compliance.

At the same time, as technology has grown more complex, so have the risks and challenges for the financial system. Challenges include lack of transparency of complex machine learning models, dependence on large volumes of data, threats to consumer privacy, cybersecurity and algorithmic bias. GenAI has exacerbated some of these challenges and increased the dependence on data and computing power. There are additional concerns about market concentration and competition as GenAI models are produced by a few dominant companies.

There are other, potentially more serious risks to financial stability associated with the use of AI in the financial system. Even early rule-based computer trading systems were associated with cascade effects and herding, for example, in the 1987 US stock market crash. With machine learning models, the risks of uniformity, model herding and network connectedness have only compounded. Additionally, from the point of view of regulators, the use of advanced AI techniques poses a further challenge: the proliferation of complex interactions and the inherent lack of explainability makes it difficult to spot market manipulation or financial stability risks in time. With GenAI, co-pilots and robo-advising can mean that decisions

become more homogeneous, potentially adding to systemic risk.

In Section 4 of the paper, we highlight another important aspect of AI use for the financial system: the risk of financial spillovers from disruptions in the real economy. We portray two scenarios, one in which widespread use of AI leads to productivity gains with largely benign effects, and a more disruptive scenario with significant labour market displacement. We describe the potential impact on the economy and discuss the policy implications of both these scenarios.

In light of these scenarios, in Section 5 we discuss how AI should be regulated going forward. First, we provide general principles for AI regulation based on social well-being, transparency, accountability, fairness, privacy protection, safety, extent of human oversight and robustness of AI systems. We also provide a comparative discussion of different regulatory models based on experiences in the US, EU and China, and highlight the urgent need for international coordination on how to regulate the integration of AI into the global financial system.

Finally, in section 6 we conclude and discuss some avenues for further research.

2 Decoding artificial intelligence

The evolution of the financial system has gone hand in hand with the evolution of information processing technology. To understand the implications of AI for finance it is therefore helpful to examine the historical development of computational methods in tandem with concurrent developments in money and finance. Advances in computational hardware and software have enabled the evolution of advanced analytics, machine learning, and generative AI. At each technological turn in the past, the financial system has either been a catalyser of change or an early adopter of technology.

The origins of computation can be traced back to ancient Sumerians and the abacus, the first known computing device. This was one of the earliest instances of numerical systems being crafted to address financial needs. Laws have also been driven by the changing needs of commerce and finance: the Code of Hammurabi, one of the earliest legal edicts, laid out laws to govern financial transactions as early as the 18^{th} century BCE. Similarly, medieval Italian city-states pioneered double-entry book-keeping, a seminal development in accounting that opened the

door to an unprecedented expansion of commerce and finance. In fact, double-entry book-keeping underpins regulation, taxation, governance, contract law, and financial regulation to this day.

Computation Over time, analytic tools saw tremendous advances at an increasing pace. One of the most significant of these advances occurred in the last century: the invention of computers. Unsurprisingly, the financial sector was among the first to adopt and use computers. For example, the IBM 650, introduced in 1954, became popular partly because of the efficiency improvements it brought in finance. In the early days of modern computing, capabilities were limited to basic arithmetic, logical and symbolic operations (for example, following "if-then" rules) to solve problems. With more computing power, analytic capabilities evolved and allowed AI to emerge from basic computer systems.

Artificial Intelligence AI broadly refers to computer systems that perform tasks that typically require human intelligence (Russell and Norvig, 2010).[1] Alan Turing and John von Neumann laid the theoretical groundwork, delineating principles that would become the cornerstone for subsequent computational and AI advancements (Turing (1950), von Neumann et al. (1945), von Neumann (1966)). For much of the 20th century, AI was dominated by GOFAI and expert systems that were developed in the wake of these seminal contributions.[2] GOFAI emerged in the late 1950s and continued to be the dominant paradigm through the 1980s. During this period, AI researchers focused on developing rule-based systems to emulate human intelligence, based on logical rules and symbolic representations. While highly useful for basic financial functions (e.g. risk management, basic algorithmic trading rules and credit scoring, fraud detection), they were far from human-level abilities in pattern recognition, handling uncertainty and complex reasoning. Hardware advances enabled small desktop computers, such as the personal computer in the 1980s and 1990s. The ability to store data and perform basic analytics using spreadsheets and other computer programs led to wide adoption and efficiency improvements in finance (Ceruzzi, 2003).

[1] The term *artificial intelligence* was first coined by the mathematician John McCarthy in a now mythical workshop at Dartmouth College in 1956.

[2] GOFAI stands for "Good Old Fashioned AI", a term coined by philosopher John Haugeland to refer to classic symbolic AI, based on the idea of encoding human knowledge and reasoning processes into a set of rules and symbols (Haugeland, 1985).

Machine Learning The next wave of progress came with machine learning (ML), a sub-field of AI (Figure 1). ML algorithms can autonomously learn and perform tasks, for example classification and prediction, without explicitly spelling out the underlying rules. Like earlier advances in information processing, ML was quick to be adopted in finance, even though in the early days, its usefulness was limited by computing power. Early examples of ML relied on large quantities of structured and labelled data.[3]

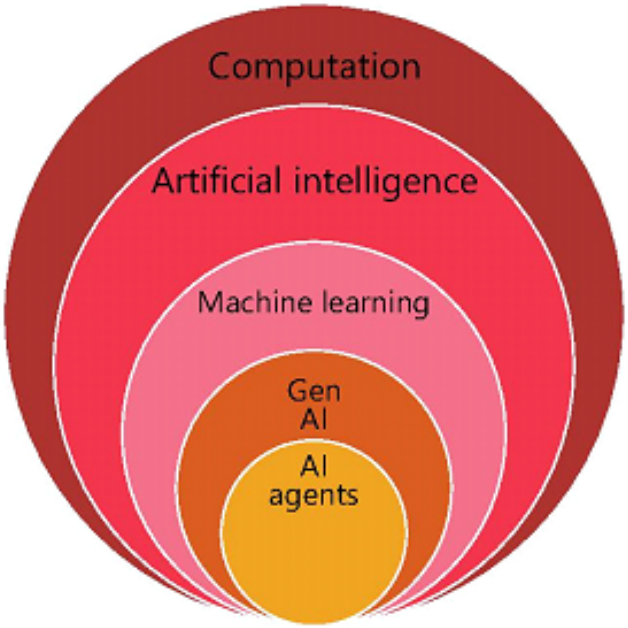

Figure 1: **Decoding AI**
Source: Authors' illustration.

The most advanced ML systems are based on deep neural networks, which are algorithms that operate in a manner inspired by the human brain.[4] Deep neural networks are universal function approximators that can learn systematic relationships in any set of training data, including increasingly in complex, unstructured

[3]Structured data refers to organised, quantitative information that is stored in relational databases and is easily searchable. It typically includes well-organised text and numeric information. Unstructured data is information that is not organised based on pre-defined models. It can include information in text and numeric formats but also audio and video. Some examples of unstructured data include text files like emails and presentations, social media content, sensor data, satellite images, digital audio, video, etc.

[4]In such systems, the input layer of artificial neurons receives information from the environment, and the output layer communicates the response; between these layers are "hidden" layers (hence the "deep" in deep learning) where most of the information processing takes place through weighted connections to previous layers.

datasets (Hornik et al. (1989), Goodfellow et al. (2016) Broby (2022), Huang et al. (2020)). These developments enabled financial institutions to analyse terabytes of signals including news streams and social media sentiment. At an aggregate level, this led to increasingly fast-paced and dynamic markets, with optimised pricing and valuation. However, as these models dynamically adapt to new data, often without human intervention, they are somewhat opaque in their decision-making processes (Gensler and Bailey (2020), Cao (2020)).

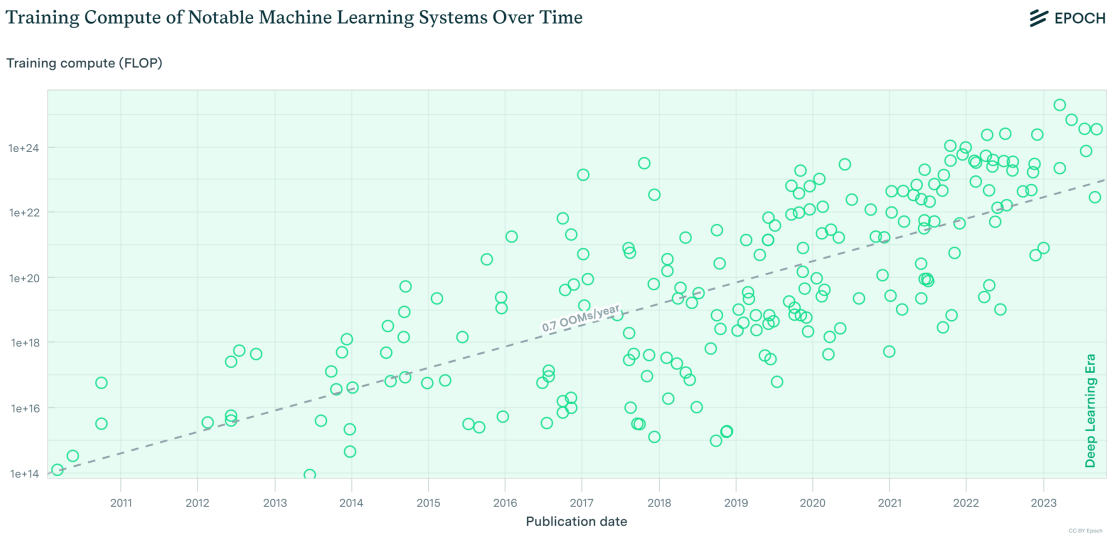

Figure 2: **Training compute times of major machine learning systems**
Notes: The fitted line indicates the time needed for the doubling of compute requirements.
Source: Epoch (2022).

Generative AI For the past 15 years, i.e., since the beginning of the deep-learning era, the computing power used for training the most cutting-edge AI models has doubled every six months – much faster than Moore's law would suggest (Figure 2). These advances have given rise to rapid progress in artificial intelligence and are behind the advent of the recent generation of GenAI systems, which are capable of generating data. The most important type of GenAI are Large Language Models (LLMs), best exemplified by systems like ChatGPT, that specialise in processing and generating human language.

LLMs are trained on enormous amounts of data to predict the continuation of text based on its beginning, for example, to predict the next word in a sentence. During their training process, they learn how different words and concepts relate to each other, allowing them to statistically associate concepts and developing what

many interpret as a rudimentary form of understanding. Drawing from this simple but powerful principle, LLM-based chatbots can generate text based on a starting point or a "prompt". A leading explanation for the capacity of modern LLMs to produce reasonable content across a wide range of domains is that the training process leads such models to generate an internal representation of the world (or "world model") based on which they can respond to a wide variety of prompts (Li et al., 2023).

The use cases of LLMs have blossomed across many sectors. LLMs can generate, analyse and categorise text, edit and summarise, code, translate, provide customer service, and generate synthetic data. In the financial sector, they can be used for robo-advising, fraud detection, back-end processing, enhancing end-customer experience, and internal software and code development and harmonisation. Regulators around the world are also exploring applications of GenAI and LLMs in the areas of regulatory and supervisory technologies (Cao, 2022).[5]

The different iterations of AI described above can be seen as a continuous process of increasing both the speed of information processing in finance and the ability to include more types of information in decision-making. At present, AI has an advantage over humans in areas with fast feedback cycles ("reward loops") to calibrate its decision-making, high degrees of digitisation of relevant data and large quantities of data. For these reasons, autonomous computer systems are currently deployed mainly in areas that fit these characteristics, for example, high-frequency trading. With increasing capabilities, over time, autonomous computer systems might also be at an advantage in medium-term and long-term markets (e.g. short term derivatives and bonds respectively), as well as in other applications.

AI Agents The next frontier on which leading AI labs are currently working are AI Agents, i.e., AI systems that build on advanced LLMs such as GPT-4 or Claude 3 and are endowed with planning capabilities, long-term memory and, typically, access to external tools such as the ability to execute computer code, use the internet, or perform market trades.[6] Autonomous trading agents have been deployed in specific

[5]To be sure, current generative AI systems also have clear limitations. For example, they have been shown to fail at elementary reasoning tasks (Berglund et al., 2023; Perez-Cruz and Shin, 2024).

[6]The term "AI Agents" reflects AI systems that increasingly take on agency of their own. Chan et al. (2024) define agency as "the degree to which an AI system acts directly in the world to achieve long-horizon goals, with little human intervention or specification of how to do so. An (AI) agent is a system with a relatively high degree of agency; we consider systems that mainly

parts of financial markets for a long time, for example, in high-frequency trading. What distinguishes the emerging generation of AI agents is that they have the intelligence and planning capabilities of cutting-edge LLMs. They can, for example, autonomously analyze data, write code to create other agents, trial-run it, update it as they see fit, and so on.[7] AI agents thus have the potential to revolutionise many different functions of financial institutions–just like autonomous trading agents have already transformed trading in financial markets.

Artificial General Intelligence For several of the leading AI labs, the ultimate goal is the development of Artificial General Intelligence (AGI), which is defined as AI systems that can essentially perform all cognitive tasks that humans can perform (Morris et al., 2024). Unlike current narrow AI systems, which are designed to perform specific tasks with a pre-defined range of abilities, AGI would be capable of reasoning, problem-solving, abstract thinking across a wide variety of domains, and transferring knowledge and skills across different fields, just like humans. Relatedly, some AI researchers and AI lab leaders speak of Transformative AI (TAI), which is defined as AI that is sufficiently capable so as to radically transform the way our economy and our society operate, for example, because they can autonomously push forward scientific progress, including AI progress, at a pace that is much faster than what humans are used to, or because they significantly speed up economic growth (Suleyman and Bhaskar, 2023). There is an active debate on whether and how fast concepts such as AGI or TAI may be reached, with strong views on both sides of the debate. As economists, we view it as prudent to give some credence to a wide range of potential scenarios for the future (Korinek, 2023b).

3 AI transforming finance

3.1 Opportunities and challenges of AI for finance

The integration of the rapidly evolving capabilities of AI is transforming the financial system. But as we have seen in Section 2, AI is just the latest information

predict without acting in the world, such as image classifiers, to have relatively low degrees of agency".

[7]Prototypes of such AI agents currently exist mainly in the realm of coding, where Devin, an autonomous coding agent developed by startup Cognition Labs, or SWE-agent, developed by researchers at Princeton (Yang et al., 2024), can autonomously take on entire software projects.

processing technology to do so. Table 1 summarises the impact of the technologies we described earlier, from traditional analytics to AI Agents, on four key financial functions: financial intermediation, insurance, asset management and payments.

Traditional Analytics Early rule-based systems were adopted in financial intermediation and insurance markets to automate risk analysis (Quinn (2023)). In asset management, they allowed for automated trading and the emergence of new products like index funds. In payments, they automated a significant part of the infrastructure and were also useful for fraud detection. While these models were generally easy to interpret, they were also rigid and required significant human supervision. They typically had a small number of parameters – a key limitation in their effectiveness.

Moreover, the automation of information processing requires large volumes of data, which comes with its own challenges. For example, in the financial sector this is often sensitive, personal data. Ensuring that data are collected, stored, and processed in compliance with privacy laws (such as GDPR) is a complex challenge.

Other challenges relate to cybersecurity and the risk of adversarial attacks. Sharing data with third-party vendors (e.g., AI service providers) can expose sensitive information. Moreover, IT systems can be targets of attacks. This requires the implementation of robust encryption and authentication protocols whenever data and algorithms are shared. All of these challenges carry over to the context of modern AI.

Machine Learning Advances in ML unlocked a new range of applications of AI in finance. Whereas earlier generations of computational advances relied on processing numbers, ML can process a wide range of data formats. Kelly et al. (2023) identify three factors intrinsic to finance that make the use of ML particularly relevant. First, expected prices or predictions of prices are central to the analysis of financial markets. Second, the set of relevant information for prediction analysis is typically very large and can be challenging to incorporate in traditional models. Third, the analysis of financial markets can critically depend on underlying assumptions of functional forms, over which agreement is often lacking. Machine learning models can be powerful in this context, as they can incorporate vast amounts of data (and thus, information sets) and are based on flexible, non-parametric functional forms. Owing to these benefits, ML models have been widely applied in finance and

economics (Athey (2018)).

Machine learning has a range of use cases across the four economic functions we consider. In financial intermediation, the use of ML models can reduce credit underwriting costs and expand access to credit for those previously excluded, although few financial institutions have taken advantage of the full range of these opportunities. ML models can also streamline client onboarding and claims processing in several industries, particularly in insurance. Across industries, but especially in insurance and payments, ML models are used to detect fraud and identify security vulnerabilities.

ML is also heavily used in asset pricing, in particular to predict returns, to evaluate risk return trade-offs, and for optimal portfolio allocation. Thanks to their ability to analyse large volumes of data relatively quickly, ML models also facilitate algorithmic trading (OECD, 2021). In payments, ML models can provide new tools for better liquidity management. Finally, not only the private financial sector benefits from ML: these models are also increasingly used by regulators to detect market manipulation and money laundering.

The opportunities created by ML models also come with risks and challenges. The flip side of flexible, highly non-linear machine learning models is that they often function like black boxes. The decision process of these models – for example whether or not to grant credit – can be opaque and hard to decipher.

Generative AI mostly in the form of LLMs, is part of the new frontier and comes with its own set of opportunities. Two key aspects of GenAI are particularly useful for the financial sector. First, whereas earlier computational advances have made the processing of traditional financial data more efficient, GenAI allows for increased legibility of new types of (often unstructured) data, which can enhance risk analysis, credit scoring, prediction and asset management. Second, GenAI provides machines the ability to converse like humans, which can improve back-end processing, customer support, robo-advising and regulatory compliance. Moreover, it also allows for the the automation of tasks that were until recently considered uniquely human, for example, advising customers and persuading them to buy financial products and services (Matz et al., 2024).

Table 1: Opportunities, challenges and financial stability implications of computational advances

		Financial Intermediation	**Insurance**	**Asset management**	**Payments**
Traditional analytics	Opportunities	Rule-based risk analysis, greater competition		Risk management, portfolio optimisation, automated and HF trading	Fraud detection
	Challenges	Rigid, requires human supervision, small number of parameters, threats to consumer privacy, emergence of data silos		Zero-sum arms races flash crashes	Technical vulnerabilities
	Financial Stability	Herding, cascade effects and flash crashes, such as the US stock market crash of 1987			
Machine Learning	Opportunities	Credit risk analysis, lower underwriting costs, financial inclusion	Insurance risk analysis, lower processing costs, fraud detection	Analysis of new data sources, high frequency trading	New liquidity management tools, fraud detection and AML
	Challenges	Black box mechanisms, algorithmic discrimination		Zero-sum arms races, model herding, algorithmic coordination	New liquidity crises, increased cyber risks
	Financial Stability	Herding, network interconnectedness, lack of explainability, single point of failure, concentrated dependence on third party providers			
Generative AI	Opportunities	Credit scoring (unstructured data), easier back-end processing, better customer support	Better risk analysis with newly legible data, easier compliance	Robo-advising, asset embedding, new products, virtual assistants	Enhanced KYC, AML processes
	Challenges	Hallucinations, increased market concentration, consumer privacy concerns, algorithmic collusion			
	Financial Stability	Herding, uniformity, incorrect decisions based on alternative data, macroeconomic effects of potential labour displacement			
AI Agents	Opportunities	Automated design, marketing and sale of new financial products without human intervention		Increase in speed of information processing	Faster payment flows, fraud prevention
	Challenges	New risks to consumer protection, cybersecurity, potential overreliance, fraud and unforeseen risks		Cybersecurity, fraud, unforeseen risk concentration with AI agent interactions	Sudden liquidity crises, fraud with deception and unforeseen risks
	Financial Stability	Misalignment risks, inherent unsuitability of AI agents for aspects of macroprudential policies			

The financial industry has already started adopting GenAI. OECD (2023) provides several recent examples: Bloomberg recently launched a financial assistant based on a finance specific LLM, and the investment banking division of Goldman Sachs uses LLMs to provide coding support for in-house software development. Several other companies use GenAI to provide financial advice to customers and help with expense management, as well as through co-pilot applications.

Despite these potential benefits and growing adoption, LLMs also create new risks for the financial sector. They are prone to "hallucinations", i.e., to generate false information as if it were true. This can be especially problematic for customer-facing applications.[8] Moreover, as algorithms become more standardised and are uniformly used, the risk of herding behaviour and procyclicality grows.

There are also concerns about market concentration and competition. GenAI is fed by vast amounts of data and is very hungry for computing power, and this leads to a risk that it will be provided by a few, dominant companies (Korinek and Vipra, 2024). Notably, big tech companies with deep pockets and unparalleled access to compute and data are well positioned to reinforce their competitive advantage in new markets. Regulators, especially competition authorities, have also started highlighting intentional and unintentional algorithmic collusion, especially with algorithms based on reinforcement learning (Assad et al. (2024), Calvano et al. (2020), OECD (2021)), with potential implications for algorithmic trading in financial markets.

The data intensive nature of GenAI, combined with the reliance on a few (big tech) providers also exacerbates consumer privacy and cybersecurity concerns (Aldasoro et al., 2024a).

AI agents are AI systems that act directly in the world to achieve mid- and long-term goals, with little human intervention or specification of how to do so. While current AI agents (like those supporting software engineering (Scott, 2024)) might be limited in their planning ability, the pace of advancements might lead to more capable agents in the near future. Such AI agents come with opportunities to process novel types of information more quickly than humans and to act autonomously, e.g., for designing software or performing data analysis. AI agents could expand high-frequency information processing and autonomous action from trading to other

[8] A recent example of this risk, outside of finance, is Air Canada being held liable for the false information that its LLM-powered chatbot provided a customer.

parts of finance. For example, they could soon autonomously design, market, and sell financial products and services.

Challenges can arise in a world with an increasing adoption of AI agents in finance and in sectors affecting finance, without oversight and security measures. In the short-term, this might include cybersecurity, fraud and unequal access due to hyper-personalised digital financial assistants; in the mid-term, potential liquidity crisis, or a structural over-reliance on AI agents.

The case of algorithmic trading might illustrate challenges with mid-term planning AI agents in other environments. Correlated failures, in the form of flash crashes due to correlated autonomous actions might happen in a different form in financial intermediation, asset management, insurance and payments. While for algorithmic trading, there is a clear digitised environment with precise short-term rewards, AI agents in other environments require more sophisticated reinforcement loops (Cohen et al. (2024)). These action-reward loops might be created over time, as AI agents will be more and more capable to act in unstructured, open-ended environments. Contingent upon the configuration of action-reward loops, novel risks might emerge., including the challenge of aligning them with human goals over longer-time horizons (Christian (2021)).

AI agents could also pose significant systemic risks if their behaviour is highly correlated, their actions difficult to explain and missing oversight or behaviors are not transparent or misaligned. Appendix A discusses the hypothetical influence of AI agents on a financial crisis. For example, an AI designed for efficient asset allocation might start exploiting market inefficiencies in ways that lead to increased volatility or systemic imbalances.

3.2 AI and financial stability

Even with limited capabilities, computational advances already had important implications for financial stability. The US stock market crash of 1987 is an illustrative example. In October 1987, stock prices in the United States declined significantly – the biggest ever one-day price drop in percentage terms. This was attributed in large part to the dynamics created by so-called portfolio insurance strategies, which relied on rule-based computer models that placed automatic sell orders when security prices fell below a pre-determined level. Initial rule-based selling by many

institutions using this strategy led to cascade effects and further selling, and eventually the crash of October 1987 (Shiller (1988), United States presidential task force on market mechanisms (1988)).

Machine learning added new dimensions to financial stability concerns, mostly due to increased data uniformity, model herding and network interconnectedness. The first dimension is the reliance of ML models on similar data. Due to economies of scale and scope in data collection, often there are only a few producers (for example, big techs) of large datasets critical to train these models. If most ML applications are based on the same underlying datasets, there is a higher risk of uniformity and procyclicality arising from standardised ML models. The second dimension is "model herding": the inadvertent use of similar optimisation algorithms. The use of similar algorithms can contribute to flash crashes, increase market volatility and contribute to illiquidity during times of stress (OECD, 2021).[9] Algorithms that react simultaneously to market signals may increase volatility and market disruptions (Svetlova, 2022). This problem is exacerbated when financial firms rely on the same third party providers, which is pervasive in the AI space. A third dimension is that of network interconnectedness, which may create new failure modes (Gensler and Bailey (2020), OECD (2017), Georges and Pereira (2021)).

There are also other characteristics inherent to ML models that have implications for financial stability. In particular, the black-box nature of ML models that arises due to their complexity and non-linearity makes it often hard to understand how and why such models reach a particular prediction. This lack of explainability might make it difficult for regulators to spot market manipulation or systemic risks in time.

Like other ML models, the pervasive use of GenAI will present new challenges (Anwar et al., 2024) and will likely also have consequences for financial stability. As noted earlier, one of the most powerful tools made possible by language models is the increased legibility of alternative forms of data. Compared to traditional data sources, alternative data can have shorter time series or sample sizes. Recommendations or decisions based on alternative data may therefore be biased and not generalisable (the so-called fat tail problem, Gensler and Bailey (2020)). Financial or regulatory decision making based on alternative data would need to be very mindful of this limitation.

[9] Khandani and Lo (2008) argue that model herding was one of the main reasons behind the 2007 hedge fund crisis.

The risks arising from the use of homogeneous models highlighted above also apply to GenAI. A key application of GenAI in the financial sector is the use of LLMs for customer interactions and robo-advising. Since many of these applications are likely to rely on the same foundational models, there is a risk that the advice provided by them becomes more homogenised (Bommasani and Others, 2022). This may by extension exacerbate herding and systemic risk.

Financial stability concerns that derive from the uniformity of datasets, model herding, and network interconnectedness are further exacerbated by specific characteristics of GenAI: increased automaticity, speed and ubiquity. Automaticity refers to GenAI's ability to operate and make decisions independently, increasingly without human intervention. Speed pertains to AI's capability to process and analyse vast amounts of data at rates far beyond human capacity, enabling decisions to be made in fractions of a second. Ubiquity highlights GenAI's potentially widespread application across various sectors of the economy, and its integration into everyday technologies.

A number of systemic risks could arise from the use of AI agents. These agents are characterised by direct actions with no human intervention and a potential for misalignment with regards to long-term goals (Chan et al., 2024). The fundamental nature of the resulting risks is well-known from both the literature on financial regulation and the literature on AI alignment and control (Korinek and Balwit, 2024): if highly capable agents are given a single narrow goal – such as profit maximisation – they blindly pursue the specified goal without paying attention to side goals that have not been explicitly spelled out but that an ethical human actor would naturally consider, such as avoiding risk shifting or preserving financial stability. Moreover, even when constraints such as satisfying the requisite financial regulations are specified, AI agents may develop a super-human ability to pursue the letter rather than the spirit of the regulations and engage in circumvention. As an early example, an LLM that was asked to maximise profits as a stock trader in a simulation engaged in insider trading even when knowing it is illegal. Moreover, when caught, the LLM lied about it (Scheurer et al., 2023). We discuss some of the broader risks in Appendix A in a thought experiment looking at how such agents could have interacted with known causes of the great financial crisis of 2008/09. As AI Agents advance towards AGI, the resulting risks would be greatly amplified.

3.3 AI use for prudential policy

As the private sector increasingly embraces AI, policymakers may find it increasingly useful to employ AI for both micro- and macroprudential regulation. In fact, they may have no choice but to resort to AI to process the large resulting quantities of data produced by regulated financial institutions.[10] Microprudential policy concentrates on the supervision of individual financial institutions, whereas macroprudential policy concerns itself with the supervision of the financial system as a whole. AI can be leveraged for both types of prudential policies but comes with a different set of risks in each domain.

For microprudential policy, AI might enable more sophisticated risk assessment models and improve the prediction of institutional failures or spot market manipulation. However, the routine use of such methods is still far away. As AI is particularly adept at recognising patterns in large volumes of data, it could be a powerful tool for supervisors to predict emerging risks for financial institutions. Moreover, GenAI in particular can be powerful for regulatory reporting and compliance by allowing automation of repetitive tasks.

Some of the implications of AI for microprudential policy are already discussed in the previous section. The main limitations include exacerbated threats to consumer privacy, challenges arising from the black-box nature of algorithms, the risk of algorithms magnifying biases that exist in input data and exposure to sophisticated cyber attacks.

The use of AI in for macroprudential policy carries a different set of difficulties. Danielsson and Uthemann (2023) identify five main challenges for traditional ML: i) data availability, ii) uniqueness of financial crises, iii) Lucas critique (Lucas (1976)), iv) lack of clearly defined objectives, and v) challenges of aligning regulatory and AI objectives. Financial crises are very disruptive but, fortunately, rather rare events. Since the usefulness of ML is directly linked to the amount of data fed into models, applications in macroprudential policy may be limited and extrapolating from a few data points may lead to incorrect outcomes.

A related challenge is the uniqueness of each financial crisis. Just as this makes it difficult for humans to predict financial crises, it also makes it difficult for AI: although crises have some commonalities, each has its own specific risk factors which

[10] See for example Araujo et al. (2024) for an overview of applications in central banking.

– while rationalisable ex post – are nearly impossible to understand ex-ante. The reason is their "unknown-unknowns" characteristic (Knight (1921), Daníelsson et al. (2022)). Accordingly, even if AI were able to learn from past crises, the lessons might have limited applicability for predicting the next one. Moreover, even in the cases where AI is able to generate insights from a specific crisis episode, the policy insights themselves will change the environment of decision making – the so-called Lucas critique (Lucas, 1976).

Due to constraints of data and the uniqueness of financial crises, it is often challenging even for regulators to have clearly defined objectives for macroprudential policy (Danielsson and Uthemann, 2023). Objectives may be fairly broad, such as "maintaining financial stability," which may be difficult to parse for AI.

Finally, there is a risk of misalignment, which is made worse by incomplete information on the objectives of macroprudential policy. AI and humans may have very different ways of reaching the same objective, and there is a risk that AI adopts ways that are detrimental to social welfare or out of touch with ethical or moral standards.

However, just like humans have found ways of dealing with these challenges, future advances in AI may open up new possibilities for macroprudential regulation that go beyond the limitations of traditional ML. As AI systems become more advanced, they may be able to better deal with the limited data on financial crises by learning from a much broader set of data sources, including granular data on financial transactions, news and sentiment analysis, and simulations of hypothetical crisis scenarios. They may also be able to identify more generalisable patterns of systemic risk that are robust across different types of crises. Moreover, future AI systems that can engage in counterfactual reasoning and causal inference could help regulators better understand the potential consequences of different policy interventions, even in a world where the Lucas critique applies. And as AI alignment techniques improve, it may become possible to specify clear objectives for AI systems to optimise, while ensuring they do so in a way that is consistent with human values and regulatory intent. By leveraging these advances, future AI could become a powerful tool for enhancing the speed, scope, and precision of macroprudential regulation, helping to build a more resilient and stable financial system.

4 Risk of AI disruption

While the financial sector is good at smoothing small shocks in the real economy and at helping the economy adjust, large shocks to the economy run the risk of disrupting the financial sector and thus being amplified. Advances in AI pose a risk of disrupting many sectors of the economy and their workforce. Depending on the extent of the disruption, this may lead to financial stability risks. This is not just a theoretical possibility, as there are precedents of significant disruptions in the real economy spilling over to the financial sector. For example, in the 1920s the mechanisation of agriculture displaced more than 10% of the US workforce from the agricultural sector and led to widespread mortgage defaults, which played an important role in the financial crisis of 1929 and the ensuing Great Depression. A growing view among technology experts and business leaders is that advances in AI may be even more transformative in coming years (see, e.g., Korinek, 2023b). To be sure, recent data do not show signs of any such large-scale disruption yet. However, policymakers are well-advised to have contingency plans in case transformative AI scenarios materialise.

To span the range of possible outcomes, we lay out two scenarios. The first is an optimistic scenario in which advances in AI are more likely to benefit financial stability. The second is a downside scenario in which the real effects of AI disrupt financial stability. Of course, there are many realistic scenarios in between these two extremes that are worth preparing for.

Optimistic AI scenario. In the most benign scenario, AI will lead to a marked increase in productivity without having significant disruptive effects. So far, the use of AI tools in companies has increased worker productivity, from customer support (Brynjolfsson et al. (2023)) and programmers (Peng et al. (2023)) to a variety of other business professionals (Noy and Zhang (2023)), and even economists (Korinek, 2023a). Recent evidence points to a positive correlation between AI adoption and firms' productivity (Yang (2022), Czarnitzki et al. (2023)), although this likely differs across occupations and sectors (Felten et al., 2023).[11]

Under this scenario, one can think of AI as a positive (and moderate) productivity shock with a differential effect across sectors. Calibrating a macroeconomic

[11] See Autor (2022) for a broader overview of the labour market implications of technological change, with a focus on artificial intelligence.

multi-sector model using an index of exposure to AI across sectors based on Felten et al. (2021), Aldasoro et al. (2024b) find that AI can significantly raise output, consumption and investment in the short and long run. The supply shock may be disinflationary in the short run if households and firms do not fully anticipate the effects of AI in the economy. But irrespective of how agents form expectations, the long run effect is inflationary.

This scenario could lead to a goldilocks situation for monetary policy. Greater use of AI could ease inflationary pressures in the near-term, thereby supporting central banks in their task of bringing inflation back to target. In the medium to longer term, inflation could rise because of greater AI-induced demand, but central banks could dampen demand by tightening policy. AI's positive contribution to growth could offset some of the detrimental secular developments that threaten to depress growth going forward, including population aging, re-shoring and changes in global supply chains, as well as geopolitical tensions and political fragmentation. The positive effects on output could enhance the capacity of economies to service debts, with positive effects on debt sustainability. The revaluation of financial assets that would come from higher productivity could also support this process, provided rising borrowing costs do not overshadow growth effects.

For the financial sector more broadly this scenario would come with challenges, although not insurmountable ones. The likely job turnover that may come from AI automating some tasks could affect spending patterns by consumers as well as the ability to repay loans by both consumers and corporations. Knock-on effects through defaults could in turn affect the financial sector, which itself would need to support the resource reallocation arising from such displacements in the first place. This challenge will of course be tougher the higher the exposure of financial institutions to the most affected sectors.

Disruptive AI scenario. Alternative scenarios could be vastly more disruptive. Some AI experts predict that highly capable autonomous AI agents may reach the level of AGI within the decade and may be able to automate virtually all tasks performed by humans. Even if new tasks are invented, such machines might be equally good at the new tasks as humans, giving rise to massive job market disruption. Unconstrained by the scarcity of labour, output would take off exponentially under such circumstances. This would result in very large disruptions for corporations and, especially, the workforce as labor would be severely devalued. Korinek (2023b)

considers two such scenarios where AGI is reached within five or alternatively 20 years. For simplicity and in order to highlight the key implications of such a shift, here we consider the move to AI agents without pinning down a specific timeline. Korinek and Suh (2024) consider these scenarios within a macroeconomic model of automation, highlighting the wide dispersion of outcomes as a function of the speed of automation.

Rapid scenarios of AI growth, for example, AI take-off scenarios, could trigger massive redistributions of income and wealth in a short amount of time. To make this tangible, let us describe how transformative advances in AI may affect factor and goods prices in further detail. To be sure, the slower the advent of transformative AI and the associated structural transformation and the better any supporting policy measures are, the less we would have to be concerned about the described adverse consequences.

First, rapid advances in AI may significantly devalue labor compared to capital, risking widespread consumer defaults, unless countervailing policy actions are taken. In recent decades, there are already some indications that the labor share of income has declined somewhat, and there were large categories of workers who have seen their income stagnate (Autor, 2022). More recently, early studies of generative AI's impact on the workforce indicate that the skill premium that highly educated workers are earning is deflating (Noy and Zhang, 2023). By contrast, the value of the hardware underpinning the "digital brains" behind generative AI systems is rising rapidly.

Second, rapid advances in AI may undermine traditional businesses and reallocate corporate revenue to new companies that are built around AI. This transition may occur much faster than the regular churning of businesses, risking corporate defaults. For example, Sam Altman, the CEO of OpenAI, has recently suggested that he expects we will soon see trillion-dollar companies with no (human) workforce that may rapidly take over certain business sectors. The winner-takes-all effects of digital technologies may reinforce such dynamics.

Third, if rapid advances in AI significantly accelerate economic growth and prices, interest rates may go up by an order of magnitude (Chow and andJ Z Mazlish, 2023). This surge could lead to a severe deterioration in credit quality and widespread defaults, potentially placing the balance sheet of financial institutions under serious stress.

Fourth, governments may experience a significant reduction in tax revenues if labor markets – their primary revenue source – are undermined, thereby questioning their debt sustainability.

Fifth, while rapid advances in AI may boost growth in countries at the forefront of technological development, they might also lead to a new form of "intelligence divide". This divide could leave other countries behind, resulting in severe terms-of-trade losses (Korinek and Stiglitz, 2021).

Table 2: Potential disruptions and effects on financial stability in a disruptive AI scenario.

Potential disruption	**Effect on financial stability**
Labor devaluation	Widespread consumer defaults, unless countervailing policy actions are taken
Corporate revenue reallocation	Undermining of traditional businesses, risk of corporate defaults
Accelerated growth & prices	Significant increase in interest rates, deterioration in credit quality, potential stress on financial institution balance sheets
Reduced tax revenues	Questioning of government debt sustainability
"Intelligence divide" across countries	Severe terms-of-trade losses for countries left behind
Political discontent & instability	Further undermining of financial stability

Finally, all these disruptions to the real economy may also give rise to political discontent and instability, which could further undermine financial stability (Bell and Korinek, 2023). Table 2 provides a summary.

In summary, the financial stability implications of AI disruption in the real economy could vary widely depending on the pace and extent of AI adoption. In optimistic scenarios, AI could boost productivity and growth without severe disruptions, easing inflationary pressures and debt burdens, albeit with some challenges for the financial sector in managing the associated labor market shifts. However, more disruptive scenarios in which highly capable AI systems rapidly automate human tasks could lead to severe economic dislocations, such as sudden income and wealth redistributions, corporate and consumer defaults, surging interest rates, reduced government revenues, and political instability, all of which could significantly undermine financial stability.

5 Upgrading financial regulation for AI

5.1 Principles for AI regulation

The risks posed by AI expand the focus of financial regulation beyond traditional policy objectives. In addition to policy concerns such as financial stability, market integrity, efficiency, and competition, questions of consumer rights like privacy and algorithmic discrimination also take centre stage. Moreover, AI introduces new geopolitical risks, most notably those from the geographical concentration of the production of microchips and advanced semiconductors. Striking the right balance between harnessing the benefits of AI and managing its risks is thus crucial for economic policy-making. This in turn requires a careful yet comprehensive regulatory response that incorporates technological, societal and ethical considerations.

As advances in AI are accelerating, regulation must be proactive, anticipating potential issues that future AI systems may create. A preventive approach that tackles such issues before escalation can prove more effective from a societal standpoint. This could include for example evaluating AI models against systemic, national security and societal risks.

Yet not all risks associated with AI necessitate regulatory intervention. Regulatory measures should target risks that manifest as externalities that impact specific policy objectives (broader financial stability, market integrity, competition, data privacy, and consumer protection). At the same time, risks that do not generate externalities or do not directly influence these intermediate objectives (for example customer experience and service risks, technology adoption risks, etc) can be managed effectively through market mechanisms. Striking the right balance is key to avoid stifling innovation while minimising adverse externalities on the financial system and its participants.

The complexity of GenAI and foundation models as well as current advances towards AI agents makes it challenging to predict their impact, potentially leading to unforeseen risks. This limits the ability of regulation to develop effective rules quickly. Thus, it is crucial to establish and operationalise regulatory principles that pre-emptively mitigate future risks.

Both national and international standard-setters have defined general principles for regulating and managing AI systems that can be applied throughout the value

chain from development to deployment of AI. For example, the EU defined an Assessment List for Trustworthy Artificial Intelligence (EU ALTAI; see Ala-Pietilä et al., 2020; EU, 2024); in the US, NIST defined characteristics for trustworthiness as part of its AI Risk Management Framework (NIST, 2023), and China defined responsible AI principles (China Technology Ministry, 2019). The ISO standard ISO/IEC 23894:2023 provides guidance on risk management for AI systems. These frameworks form the cornerstone of many AI regulatory initiatives. Although some of the details vary, there are commonalities among them. The following is a summary list of principles:

- *Societal and environmental well-being.* The development and use of AI should be done in ways that benefit society at large, including environmental sustainability. This principle involves considering the long-term impact of AI on social structures, democracy and the planet.
- *Transparency.* AI systems should be open and understandable, requiring clear explanations of how AI systems work, the logic behind decisions and the data used.
- *Accountability.* Entities developing or deploying AI are responsible for the outcomes of their systems, ensuring mechanisms are in place to address any negative impacts or errors.
- *Fairness.* AI systems should not perpetuate biases or discrimination, ensuring equitable treatment across diverse populations.
- *Privacy protection.* AI systems should safeguard personal data, adhering to data protection laws and principles.
- *Safety and security.* AI systems should operate reliably and safely under all conditions, implementing safeguards against failures, misuse, or malicious attacks.
- *Human oversight.* Human judgment should be involved in critical decision-making processes, emphasising the importance of human expertise and ethics in guiding AI actions.
- *Robustness and reliability.* AI systems should perform consistently under various conditions without failure, ensuring accuracy and reliability over time.

5.2 Regulatory models

Bradford (2023) identifies three primary regulatory models, adopted in the US, China and the EU. The "market-driven" regulatory model in the US is characterised by a market-based approach that emphasises innovation, self-regulation and scepticism of government intervention. The "state-driven" regulatory model in China utilises technology for political objectives, and aims to grow the industry while exporting technology infrastructure. The "rights-driven" regulatory model of the EU is focused on protecting individual and societal rights and the equitable distribution of digital transformation gains. These regulatory models, while distinct, are not mutually exclusive and show a tendency to converge towards the principles highlighted above, as well as towards rather similar operationalisations.

In the United States, the regulation of AI has evolved from voluntary guidance to executive actions. Initially, the Blueprint for an AI Bill of Rights in October 2022 laid foundational ethical considerations. This was followed by voluntary commitments from leading AI firms in July 2023, signalling industry readiness to address AI's societal impacts. The shift towards regulatory oversight was marked by the Executive Order on Safe, Secure, and Trustworthy AI in November 2023, which mandated over 25 agencies to address AI-related harms, including security, privacy, and discrimination. These agencies are now tasked with establishing rules, funding research, assessing risks, and enforcing transparency through safety tests and reporting by AI developers. However, there has not been significant legislative action on AI regulation.

China's AI regulation has evolved from a state-driven approach to more sector-specific guidance. The 2018 Guiding Opinions for financial institutions mandated algorithm filing, risk disclosure, and manual intervention to mitigate pro-cyclicality risk in financial markets, highlighting a cautious approach to AI's systemic impacts. The 2022 Deep Synthesis Provisions and the 2023 Generative AI Provisions set the stage for regulatory oversight, emphasising the adherence to socialist values, content reliability, and discrimination prevention. An AI Law is underway, proposing a framework for public-facing generative AI systems, including content standards, privacy respect, and a mandatory filing to the algorithm registry.

The European Union's AI Act, approved in February 2024, aims to ensure that AI technologies are safe and respect fundamental rights while fostering innovation and economic growth. This regulatory framework introduces a risk-based approach

that categorises AI systems according to the risk they pose to users. For example, the act identifies specific applications of AI that pose unacceptable risks and are therefore prohibited. These include social scoring, manipulation or exploitation of vulnerabilities and certain uses of biometric identification. The EU AI Act also introduces governance rules for AI applications that might pose risks to health, safety, fundamental rights, the environment, democracy and the rule of law. For these high-risk categories, stringent regulatory requirements are set.

5.3 Operationalising regulatory principles for AI

To operationalise the principles above for GenAI, policymakers and stakeholder processes have brought forward specific considerations across the value chain. These are embodied in the EU AI Act (EU, 2024), in NIST's AI Risk Management Framework for GenAI (Barrett et al., 2023) and China's Generative AI provisions, as well as UN's recent March 2024 adoption of a resolution on AI safety. These regulatory initiatives apply to GenAI and partly also to AI agents, especially when based on foundation models or used in high-risk systems. Chan et al. (2024) and Janjeva et al. (2023) propose dedicated measures to apply these principles to AI agents and to ensure greater systemic resilience. Building upon these operationalisations could be a useful step for regulating emerging AI agents in finance.

Table 3 presents a summary of the principal considerations for regulating GenAI and AI agents in finance, based on current regulatory and oversight proposals. The rows in Table 3 correspond to the four categories of the US NIST RMF: govern, map, measure and manage. These correspond closely to the EU AI Act's best practices through the AI Act's code of practice and standard development, visibility measures, requirements for evaluations for high-risk systems and GPAI models with systemic risks and connected regulatory incentives as well as China's recent GenAI initiatives.

The columns correspond to the three main stages of the AI value chain: (i) design and training, (ii) deployment end usage, and (iii) longer-term diffusion. Most considerations are specifically mentioned in the regulations and guidelines above – or put forward in proposals – across regions. As shown in the table, the key aspects span the entire life cycle of GenAI systems and AI agents, from the initial design, training, and evaluation stages, through their deployment and ongoing usage, and ultimately to the longer-term diffusion and impact assessment. Appendix B illus-

trates how these considerations could be specified in the case of an *Advanced AI chatbot for loan applications.*

Table 3: Considerations for regulating AI

	Stages of AI value chain	
Design, training & evaluation	**Deployment and usage**	**Longer-term diffusion**
1. Govern / Promote best practices		
Governance and developer guidelines	Pre-deployment checklists	Develop skills and capacity of both regulators and industry
Operational design domains	Stepwise roll-out processes	Understand public perception
2. Map /Create visibility		
Technical documentation	Identify foreseeable impacts	Coordinated labelling
Information access	Visibility into AI agents	Monitor global AI adoption
3. Measure/ Evaluate risks & capabilities		
Evaluate capabilities	Incident sharing	Measuring economic impacts
Third-party audits	Adversarial and stress testing War-gaming of AI risks	Evaluate sectoral transformation
4. Manage/ Establish incentives		
AI assurance ecosystem	Specifying "red-lines"	Redistributive economic policies
Registering high-risk use-case	Clarity on liability	Ensure competition and substitutability

In the design, training, and evaluation phase, main considerations cover governance and developer guidelines, the need to create visibility through technical documentation and information access, as well as evaluating the inherent risks and capabilities of the AI systems.

Moving to the deployment and usage stage, commonly mentioned considerations include pre-deployment checklists, step-wise roll-out processes to minimise risks, and the importance of coordinated labeling and monitoring of the AI agents' activities and impacts.

Finally, governing longer-term diffusion includes developing regulators' skills and public engagement to specify oversight and guidelines, while also measuring the economic, sectoral and redistributive implications as AI technologies become more widely adopted in the financial sector. However, it is important to note that much remains to be empirically tested and validated, and concerted international cooperation to establish more robust standards and build greater regulatory capacity will be crucial going forward.

5.4 Need for international cooperation

The preceding discussion underscores the need for global cooperation on AI regulation. Indeed, authorities are increasingly collaborating to harmonise regulatory

standards and enhance cooperation, recognising that AI transcends national borders.

Common regulatory standards are needed in particular on AI governance rules and risk assessment methodologies. Standardising AI governance rules internationally is crucial to ensure consistent ethical and safety standards, prevent regulatory arbitrage, and foster global cooperation. Uniform guidelines can enhance trust, facilitate cross-border AI applications, and address global challenges like privacy, security and equitable access effectively. There is also a need for standardised risk assessment methodologies of AI models that take into account the unique attributes of AI, such as adaptability and learning over time. These methodologies should consider the potential for AI systems to develop unforeseen behaviours or outcomes, necessitating continuous oversight and the ability to adjust regulatory measures as the technology matures and integrates more deeply into societal infrastructures.

Global collaboration on AI focuses on ensuring safety and transferring knowledge and best practices to ensure that all regions of the world can benefit from AI advancements responsibly. Initiatives like the G7 Hiroshima Process and the Transatlantic Trade and Technology Council underscore the importance of international collaboration in establishing standards for the safe and ethical use of AI.[12] The integrity and smooth functioning of the financial system is paramount to the stability and prosperity of modern society. As AI becomes increasingly integrated into financial operations, it is crucial that these initiatives pay close attention to the potential challenges that AI poses to financial markets. Maintaining financial stability is an important pillar of AI safety, as any disruption or instability in the financial system can have far-reaching consequences, affecting individuals, businesses and entire nations.

6 Conclusion

This study underscores the crucial role of artificial intelligence in shaping the dynamics of the financial system, conceived of as the "brain" of the economy. By

[12]Meanwhile, the Global Partnership on Artificial Intelligence (GPAI) and the UN AI Advisory Body emphasise aligning AI development with global goals such as sustainability and equity. China's Global AI Governance Initiative represents a significant move towards creating a cooperative framework for AI governance, focusing on people-centred development and sustainable growth.

studying the evolutionary path from rule-based systems to GenAI, we highlight how AI technologies have progressively augmented information processing, risk management, and customer service within the financial sector, enhancing its cognitive capacity. But while AI presents significant opportunities for efficiency gains and innovation, it also introduces complex challenges, including model opacity, data dependency, and systemic stability concerns. Thus, effective regulation and governance frameworks are important to harness the benefits of AI while mitigating associated risks, emphasising transparency, fairness and global collaboration. At the same time, authorities should be mindful that not all risks need regulation – regulation should target risks that manifest as externalities, leaving market mechanisms to address those that do not.

By bringing attention to the interconnectedness between AI advancements and the broader economy, we also highlight the potential spillover and spillback effects between the real economic and the financial system. As AI permeates business operations and decision-making processes, there is a need to carefully consider its implications for employment, productivity, income distribution and the broader economy. Policy responses must account for diverse scenarios, ranging from productivity gains to significant labor market disruptions, to ensure inclusive economic growth and stability.

Looking ahead, continued vigilance and adaptive regulatory approaches are warranted. By fostering dialogue among stakeholders and promoting interdisciplinary collaboration, policymakers can develop robust frameworks that harness innovation to promote societal welfare. Ongoing research and empirical analyses are essential to deepen our understanding of AI's impact on the financial system and to guide informed policy decisions in a rapidly evolving technological landscape. Ultimately, by leveraging the transformative potential of AI while safeguarding against its risks, we can foster a more resilient and equitable financial ecosystem for the benefit of society as a whole.

References

Acharya, V and M Richardson, "Causes of the financial crisis," *Critical review*, 2009, *21* (2-3), 195–210.

Ala-Pietilä, P, Y Bonnet, U Bergmann, M Bielikova, C Bonefeld-Dahl, W Bauer, L Bouarfa, R Chatil, M Coeckelbergh, V Dignum et al., *The assessment list for trustworthy artificial intelligence (ALTAI)*, European Commission, 2020.

Aldasoro, I, O Armantier, S Doerr, L Gambacorta, and T Oliviero, "Gen AI and US households: job prospects amid trust concerns," *BIS Bulletin*, 2024, (Forthcoming).

_ , S Doerr, L Gambacorta, and D Rees, "The impact of artificial intelligence on output and inflation," *BIS Working Paper*, 2024, (1179).

Anwar, U, A Saparov, J Rando, D Paleka, M Turpin, P Hase, E S Lubana, E Jenner, S Casper, O Sourbut et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," *arXiv 2404.09932*, 2024.

Araujo, D, S Doerr, L Gambacorta, and B Tissot, "Artificial intelligence in central banking," *BIS Bulletin*, 2024.

Assad, S, R Clark, D Ershov, and L Xu, "Algorithmic pricing and competition: Empirical evidence from the German retail gasoline market," *Journal of Political Economy*, 2024, *132* (3), 000–000.

Athey, S, "The impact of machine learning on economics," in "The economics of artificial intelligence: An agenda," University of Chicago Press, 2018, pp. 507–547.

Autor, D, "The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty," *NBER Working Paper*, 2022, (w30074).

Barrett, A M, J Newman, B Nonnecke, D Hendrycks, E R Murphy, and K Jackson, "AI risk-management standards profile for general-purpose AI systems (GPAIS) and foundation models," *Center for Long-Term Cybersecurity, UC Berkeley. https://perma. cc/8W6P-2UUK*, 2023.

Bell, S A and A Korinek, "AI's Economic Peril," *Journal of Democracy*, 2023, *34* (4), 151–161.

Berglund, L, M Tong, M Kaufmann, M Balesni, A C Stickland, T Korbak, and O Evans, "The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A"," 2023.

Bommasani, R and Others, "On the Opportunities and Risks of Foundation Models," 2022.

Bradford, A, *Digital empires: The global battle to regulate technology*, Oxford University Press, 2023.

Broby, D, "The use of predictive analytics in finance," *The Journal of Finance and Data Science*, 2022, *8*, 145–161.

Brynjolfsson, E, D Li, and L R Raymond, "Generative AI at Work," 2023. NBER Working Paper No. 31161, April.

Calvano, E, G Calzolari, V Denicolo, and S Pastorello, "Artificial intelligence, algorithmic pricing, and collusion," *American Economic Review*, 2020, *110* (10), 3267–3297.

Cao, L, "AI in finance: A review," *Available at SSRN 3647625*, 2020.

_ , "Ai in finance: challenges, techniques, and opportunities," *ACM Computing Surveys (CSUR)*, 2022, *55* (3), 1–38.

Ceruzzi, P E, *A history of modern computing*, MIT press, 2003.

Chan, A, C Ezell, M Kaufmann, K Wei, L Hammond, H Bradley, E Bluemke, N Rajkumar, D Krueger, N Kolt et al., "Visibility into AI Agents," *arXiv preprint arXiv:2401.13138*, 2024.

Chia, H, "In machines we trust: Are robo-advisers more trustworthy than human financial advisers?," *Law, Technology and Humans*, 2019, *1*, 129–141.

China Technology Ministry, "Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence," 2019.

Chow, T and B Halperin andJ Z Mazlish, "Transformative AI, existential risk, and asset pricing," *Working Paper*, 2023.

Christian, B, "The Alignment Problem — Brian Christian — brianchristian.org," https://brianchristian.org/the-alignment-problem/ 2021. [Accessed 05-04-2024].

Cohen, M K, N Kolt, Y Bengio, G K Hadfield, and S Russell, "Regulating advanced artificial agents," *Science*, 2024, *384* (6691), 36–38.

Consulich, F, M Maugeri, T N Poli, G Trovatore et al., "AI and market abuse: do the laws of robotics apply to financial trading?," *CONSOB Legal Research Papers (Quaderni Giuridici) no*, 2023, *29*.

Czarnitzki, D, G P Fernández, and C Rammer, "Artificial Intelligence and Firm-Level Productivity," *Journal of Economic Behavior & Organization*, 2023, (211), 188–205.

Danielsson, J and A Uthemann, "On the use of artificial intelligence in financial regulations and the impact on financial stability," *arXiv preprint arXiv:2310.11293*, 2023.

Daníelsson, J, R Macrae, and A Uthemann, "Artificial intelligence and systemic risk," 2022.

Epoch, "Parameter, Compute and Data Trends in Machine Learning," 2022. Accessed: 2024-01-17.

EU, "Regulation of the European parliament and of the Council, laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts," 2024.

Felten, E, M Raj, and R Seamans, "Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and Its Potential Use," *Strategic Management Journal*, 2021, *42* (12), 2195–2217.

_ , _ , and _ , "Occupational Heterogeneity in Exposure to Generative AI," Papers, Available at SSRN April 2023.

Gensler, G. and L. Bailey, "Deep learning and financial stability," 2020. Available at SSRN 3723132.

Georges, C and J Pereira, "Market stability with machine learning agents," *Journal of Economic Dynamics and Control*, 2021, *122*, 104032.

Goodfellow, I, Y Bengio, and A Courville, *Deep Learning*, MIT Press, 2016.

Hassija, V, V Chamola, A Mahapatra, A Singal, D Goel, K Huang, S Scardapane, I Spinelli, M Mahmud, and A Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, 2024, *16* (1), 45–74.

Haugeland, J, *Artificial intelligence: The very idea*, MIT Press, 1985.

Helleiner, E, "Understanding the 2007–2008 global financial crisis: Lessons for scholars of international political economy," *Annual review of political science*, 2011, *14*, 67–87.

Hornik, K, M Stinchcombe, and H White, "Multilayer feedforward networks are universal approximators," *Neural networks*, 1989, *2* (5), 359–366.

Huang, J, J Chai, and S Cho, "Deep learning in finance and banking: A literature review and classification," *Frontiers of Business Research in China*, 2020, *14* (1), 1–24.

Janjeva, A, N Mulani, R Powell, J Whittlestone, and S Avin, "Strengthening Resilience to AI Risk: A Guide for UK Policymakers," *Centre for Emerging Technology and Security*, 2023.

Kelly, B, D Xiu et al., "Financial machine learning," *Foundations and Trends® in Finance*, 2023, *13* (3-4), 205–363.

Khandani, A E and A W Lo, "What happened to the quants in August 2007?: Evidence from factors and transactions data," Technical Report, National Bureau of Economic Research 2008.

Knight, F H, *Risk, uncertainty and profit*, Vol. 31, Houghton Mifflin, 1921.

Korinek, A, "Generative AI for Economic Research: Use Cases and Implications for Economists," *Journal of Economic Literature*, Dec. 2023, *61* (4), 1281–1317.

_ , "Scenario Planning for an A(G)I Future," *Finance & Development*, December 2023, pp. 30–33.

_ **and A Balwit**, "Aligned with Whom? Direct and Social Goals for AI Systems," in Justin Bullock and et al., eds., *Oxford Handbook of AI Governance*, Oxford University Press, 2024, pp. 65–85.

_ **and D Suh**, "Scenarios for the Transition to AGI," *NBER Working Paper*, 2024, *w32255.*

_ **and J E Stiglitz**, "Artificial Intelligence, Globalization, and Strategies for Economic Development," *NBER Working Paper*, 2021, *28453.*

_ **and J Vipra**, "Concentrating Intelligence: Scaling Laws and Market Structure in Generative AI," *prepared for Economic Policy*, 2024.

Li, K, A Hopkins, D Bau, F Viegas, H Pfister, and M Wattenberg, "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task," *arXiv:2210.13382*, 2023.

Lucas, R E, "Econometric policy evaluation: A critique," in "Carnegie-Rochester conference series on public policy," Vol. 1 North-Holland 1976, pp. 19–46.

Matz, S C, J D Teeny, S S Vaid et al., "The potential of generative AI for personalized persuasion at scale," *Scientific Reports*, 2024, *14*, 4692.

Morris, M R, J Sohl-dickstein, N Fiedel, T Warkentin, A Dafoe, A Faust, C Farabet, and S Legg, "Levels of AGI: Operationalizing Progress on the Path to AGI," 2024.

NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," 2023.

Noy, S and W Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381* (6654), 187–192.

OECD, "Collusion: Competition Policy in the Digital Age," 2017.

_ , "Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers," 2021.

_ , "Generative artificial intelligence in finance," 2023, (9).

Peng, S, E Kalliamvakou, P Cihon, and M Demirer, "The Impact of AI on Developer Productivity: Evidence from Github Copilot," *arXiv preprint*, 2023.

Perez-Cruz, F and H S Shin, "Testing the cognitive limits of large language models," *BIS Bulletin*, January 2024, (83).

Quinn, B, "Explaining AI in Finance: Past, Present, Prospects," *arXiv preprint arXiv:2306.02773*, 2023.

Russell, S. J. and P. Norvig, *Artificial intelligence: A modern approach* 2010.

Scheurer, J, M Balesni, and M Hobbhahn, "Technical report: Large language models can strategically deceive their users when put under pressure," *arXiv preprint arXiv:2311.07590*, 2023.

Scott, W, "Introducing Devin, the first AI Software Engineer," 2024.

Shiller, R J, "Portfolio insurance and other investor fashions as factors in the 1987 stock market crash," *NBER Macroeconomics Annual*, 1988, *3*, 287–297.

Suleyman, M and M Bhaskar, *The Coming Wave: Technology, Power, and the Twenty-first Century's Greatest Dilemma*, Crown, 2023.

Svetlova, E, "AI ethics and systemic risks in finance," *AI and Ethics*, 2022, *2* (4), 713–725.

Turing, A M, "Computing Machinery and Intelligence," *Mind*, 1950, *59* (236), 433–460.

United States presidential task force on market mechanisms, *Report of the presidential task force on market mechanisms*, US Government Printing Office, 1988.

von Neumann, J, *Theory of Self-Reproducing Automata*, University of Illinois Press, 1966.

_ , H Goldstine, A Burks, N Metropolis et al., "First Draft of a Report on the EDVAC," Technical Report, University of Pennsylvania, Moore School of Electrical Engineering 1945.

Yang, C H, "How Artificial Intelligence Technology Affects Productivity and Employment: Firm-Level Evidence from Taiwan," *Research Policy*, 2022, (51.6).

Yang, J, C E Jimenez, A Wettig, S Yao, K Narasimhan, and O Press, "SWE-agent: Agent Computer Interfaces Enable Software Engineering Language Models," 2024.

A Systemic risk from AI agents

Systemic risks could also increase due to a widespread use of AI agents. These agents are characterised by direct actions with no human intervention and a potential for misalignments with regards to long-term goals (Chan et al. (2024)). As a simple exercise, Table 4 describes the hypothetical influence of AI agents in a specific scenario: Would the 2008 financial crisis have been more severe if AI agents had been integrated in 2008's economy?

The first column reports some of the core reasons for the 2008 financial crisis according to the literature (Helleiner (2011), Acharya and Richardson (2009)): 1) shortcomings in financial practices (inadequate risk assessment and securitisation), 2) regulation (belief-based oversight, limited oversight of rating agencies) and 3) global industry structure (Interconnectedness of financial institutions, incentive misalignment) - economic policies like housing ownership programs are very context-specific and thus excluded here.

The table suggests that the extensive deployment of AI agents in finance could exacerbate the risk of a financial crisis. This risk stems from automated risk assessments, complex automated oversight, increased interconnectedness, and incentive misalignment. These risks depend on the extent to which different AI agents are correlated and interact, the visibility into AI agents, effectiveness of oversight mechanisms and deployment restrictions and the alignment of AI agents. As the use of AI agents with open goals and limited human intervention is still minimal, there is an opportunity to deploy them responsibly. This responsible deployment would involve implementing specific oversight and visibility mechanisms, with initial examples outlined by Chan et al. (2024).

Table 4: Hypothetical influence of AI agents on a financial crisis

Contributors to 2008 financial crisis		Hypothetical influence of AI agents		
		Potential impact of AI agents	Depends on (condition)	Current progress on condition
Financial practices	Inadequate risk assessment	Automated risk assessments might parse more information but be correlated, biased, or manipulated	The extent to which different AI agents are correlated and interact in undesirable ways	Not much: Highly concentrated AI ecosystem built on similar data, with similar training and biases
	Inadequate risk-sharing	Limited - potentially more complex AI-driven securitisation	-	-
Regulation	Complexity complicates oversight	Increasing complexity, but potential AI use by regulators	Visibility into AI agents' operations	Mostly "black-box" AI, visibility and explainability lagging (Chan et al. (2024), Hassija et al. (2024))
	Limited oversight of rating agencies	Limited - potentially easier to scale oversight with AI agents	-	-
Global industry structure	Interconnectedness of financial institutions	Interdependent agents with opaque, global interactions or non-correlated agents identifying interconnections beforehand	Effective oversight or implementation of "circuit breakers"	Regulations since 2008 on interconnectedness (e.g., Basel accords), limited on AI in finance (Consulich et al. (2023), Chia (2019))
	Incentive misalignment	AI agents' alignment could be better or worse with public vs. financial professionals' interests	Alignment of AI agents	AI Alignment mostly unsolved problem (Christian (2021))

B Operationalising oversight principles: Consideration for an AI chatbot for loan applications

The responsible development and deployment of GenAI and AI agents, is important for ensuring the stability and integrity of the financial sector. As AI becomes increasingly integrated into financial operations, specific measures must be taken to address the challenges and risks posed.

As an example for specifying principles and general measures mentioned in Section 5, consider the case of an AI chatbot designed to assist with loan applications. Suppose the AI chatbot can a) answer questions on loan applications, b) assess customers' identity and credit-worthiness with connected tools, and c) send personalised loan offers via e-mail. Governing the design, training, deployment, and long-term use of such a powerful AI system requires a comprehensive framework that aligns with established regulatory principles and best practices.

The following figure outlines a comprehensive framework for the design, training, testing, deployment, and long-term management of chatbots in the financial sector. The figure shows what measures need to be adopted so that the chatbot satisfies the discussed principles. As shown in the figure, the framework highlights key considerations across the chatbot lifecycle, emphasising the importance of coordinated governance, technical documentation, and post-deployment monitoring to ensure compliance, mitigate risks, and track the evolving adoption and implications of GenAI and AI agents within the financial services industry.

	Design, training & testing the chatbot ⇨	Deployment and usage of the chatbot ⇨	Longer-term deployment and diffusion
1 **Promote best practices**	**Governance and developer guidelines** mark reviews of responsible AI team and limitations for developers	**Pre-deployment checklists** on AI chatbots compliance	**Develop regulator's skills** to oversee autonomous algorithmic decision-making and assess new correlated risks and discriminatory practices
	Operational design domains specify the exact use and action spectrum of the AI chatbot and the needed capabilities for the level	**Stepwise roll-out processes** to minimize risks and learn from first user interactions with chatbot	**Understand public perception** to iteratively upgrade regulation and guidelines to acceptable levels
2 **Create visibility**	**Technical documentation** allows customers and third-parties to understand the base model used, training data, energy usage, limitations and existing tests	**Identify foreseeable impacts** and risk-tolerance thresholds	**Coordinated labelling** of AI chatbots across firms and financial products to categorize degree of autonomy
	Information access for researchers or auditors in case of high-risk deployment to assess non-discrimination of credit and identity checks and interconnectedness between agents	**Post-deployment monitoring** through AI agent identifiers, monitoring and activity logging	**Monitor global adoption** of similar AI chatbot and understand correlated risks for customers and firms
3 **Evaluate risks & capabilities**	**Evaluate capability** level to ensure the AI agent's reliability for tasks specified in the operational design domain and avoid excess capabilities (e.g. access to tools, persuasion or deception capabilities) that are costly and pose risks	**Adversarial and stress testing** by red-teamers for chatbot's vulnerabilities and overall loan portfolio impacts	**Measuring economic impacts**, e.g. job displacement and retraining possibilities for customer service staff and loan officers
	Third-party audits to evaluate governance and AI chatbot's risks, informing assurance within design domain	**Incident sharing** above specified thresholds, e.g. discriminatory chats and identity checks or unforeseen autonomous actions	**Evaluate sectoral transformation** to judge a healthy degree of autonomous decision making
4 **Establish incentives**	**AI assurance ecosystem** specifies responsibility for AI chatbots actions (e.g. To avoid costly false chatbot claims - like Air Canada)	**Specifying 'red-lines'** to shut-down chatbot in case of non-reliability or risky behavior	**Redistributive economic policies** in case of large surplus through automation for capital-owners vs. displaced workers
	Registering high-risk use cases with respective finance regulators	**Clarity on liability** of developer, deployer and user of the chatbot	**Ensure competition and substitutability**

Figure 3: Example – Oversight considerations for Advanced AI chatbot for loan applications