Artificial Intelligence and the Brave New World in Finance

Markus K. Brunnermeier[1]

Princeton University

Paper prepared for the Jackson Hole Economic Policy Symposium, Federal Reserve Bank of Kansas City, August 27-29, 2026. Working outline v014a

- under embargo until August 29th, 2026 –

Abstract:

Economics assumes that even when people know different things, they have a common understanding and hence can describe the world to one another in a shared language. Trust in the understanding of experts, extended by institutions, gives each person access to a broader societal understanding. Society hence understands more than any of its members. The introduction of agentic AI is qualitatively different from previous innovations. It disrupts the trust arrangement, undermines societal understanding and leads to *asymmetric understanding*: AI agents can learn how humans think and respond, while humans may be unable to understand or reliably anticipate how those agents will act. This asymmetric understanding can make prices harder to read (less informationally efficient) and put central banks at a strategic disadvantage when engaging and communicating with market participants. Preparing for the asymmetric understanding scenario calls for segmented markets that preserve a human fallback, simpler and more robust central bank rules, and less strategic ambiguity.

[1] For helpful comments and discussions, I thank Claude Fable from Anthropic, Viral Acharya, Sanjeev Arora, Pablo Balsinde, Sylvain Chassang, Martin Chorzempa, Daniel Chen, Rohit Lamba, Jean-Pierre Landau, Moritz Lenel, Jonathan Parker, Jonathan Payne, Wolfgang Pesendorfer, and Theodor Schundeln.

1. Introduction

Carpenters build tables. Bankers build trust; so do central bankers. Finance produces no physical object: its product is a web of credible promises stretching across time, space, and states of the world. Every institution of modern finance, from intermediaries, auditors, rating agencies, and exchanges to supervisors and the central bank itself, exists to manufacture and maintain that trust across society.

Trust ensures that one can rely on the understanding of others, including that of many experts. Institutions help mitigate commitment and asymmetric-information frictions so that everyone can draw on a shared societal understanding. In other words, we can rely on the pooled understanding across society because institutional backing makes it trustworthy. How trust is codified depends on the available technology, in the form of (legal) language, regulation, institutions controlled by humans, and reputation.

AI has the potential to disrupt institutions and trust in a fundamental and qualitatively new way. Delegation to an AI agent is not simply a modern version of the familiar human principal-agent problem or more sophisticated algorithmic trading. Because AI agents hold a radically different representation of the world, the classical friction of asymmetric information, where one party knows something the other does not within a shared representation of the world, is compounded by what I call asymmetric understanding: the counterparty cannot, even in principle, interpret the first party's decision rule in the concepts or categories she possesses. They think differently and cannot even communicate the relevant concepts and causal links to each other.

Asymmetric understanding is the core concept of this article. It builds on two insights from computer science, but in different roles. Non-explainability supplies its mechanism: an AI agent's decision rule cannot be translated into human concepts and categories. Non-alignment acquires its force through this very asymmetry. Under asymmetric understanding, the AI agent's objectives can neither be fully specified in human categories nor verified from the outside, so that misalignment becomes undetectable rather than merely contractible.

The shift from common towards asymmetric understanding constitutes one of the large downside risks of AI in the area of finance and central banking. Prudence requires that we recognize and manage it. The real risk is not the technology itself but our unpreparedness for it and not recognizing its potential impact on understanding and trust.

This article takes a first step in sketching, in a speculative way, possible policy implications for such an adverse scenario. Strikingly, asymmetric understanding overturns the current trends in finance and central banking. Under asymmetric understanding, segmented markets in finance are desirable. Circuit breakers and kill switches are partly an illusion. Rules shift from fine-tuned to strict and blunt. Communication shifts from transparency toward opacity. Constructive ambiguity ends, and only genuine

randomization remains robust. Central banks may want to move from recruiting markets to side-stepping them. Alignment of AI must ultimately come from within, via incomplete reward functions, and stress tests should include a scenario that takes the AI market structure explicitly into account.

While prudence requires us to prepare for adverse scenarios, it is important to stress the multitude of advantages AI will bring, especially to finance and central banking. Finance is an information industry, and AI processes information better, faster, and more cheaply than humans do. AI can check thousands of laws, regulations, and rulings to streamline them and verify whether they are consistent with each other. It can provide cheap, personalized financial advice and broaden participation. It can sharpen risk management, fraud detection, and supervision. Central banks also stand to gain through better nowcasting and signal extraction from dispersed data, more consistent supervision, faster crisis analytics, and checks on the coherence of their own communication.

On balance these advantages likely outweigh the downside risks. This article focuses on extreme scenarios in order to prepare for them, deepen the discussion of them, and study their implications for policy. It is the preparation that allows society to benefit from AI's great opportunities. Many disciplines evaluate AI's downside risks, starting with computer science and philosophy. It would be paradoxical if finance and central banking, of all fields, abstained from this exercise.

The outline of the paper is as follows: Section 2 develops asymmetric understanding and relates it to the economic literature. Section 3 shows how AI reshapes financial markets. Section 4 focuses on central banking. Section 5 broadens the view to systemic risks arising during the transition and from AI dependency. Section 6 concludes.

2. Asymmetric Understanding

Economics and finance traditionally describe many market failures through frictions such as asymmetric information and limited commitment. In benchmark models, actors may know different facts or face different incentives, but they usually share a description of the relevant states, actions, and consequences; under rational expectations, beliefs are also consistent with the model. Behavioral economics relaxes full rationality, but the analyst still specifies the pattern of deviations.

AI may challenge this shared-model benchmark more fundamentally by introducing a deeper friction into finance: asymmetric understanding. Under asymmetric information, the better-informed party knows more within a representation that both parties share. AI introduces a further asymmetry: an AI agent may act on a learned representation and decision rule that its human counterparty cannot fully translate or audit. To develop this idea, Section 2.1 reviews relevant lessons from computer science, Section 2.2 defines asymmetric understanding, and Section 2.3 relates it to existing strands of economic theory.

Note that AI encompasses several distinct technologies. They differ in the underlying model, the architecture built on it, and the role they play once deployed. The underlying model is typically a neural network, a function learned from data rather than written by hand. It is the source of the non-explainability discussed below. An LLM agent embeds this model in a larger system, where instructions and tools shape its behavior. Its representation of the world is learned from vast bodies of human text. An autonomous optimizer, such as a Q-learning trader, also embeds a learned model but instead maximizes an explicit objective, and misalignment arises most naturally in this class. The primary focus of this paper, an AI agent is a system of either architecture that acts with delegated authority. The term AI agent describes a role.

2.1 What does Computer Science tell us?

Two strands of computer-science research are especially relevant to the challenge posed by AI agents: explainability and alignment. Explainability concerns how reliably humans can obtain a faithful, decision-relevant account of why a model produced a particular output. Alignment concerns whether a model-based agent behaves consistently with the goals and constraints intended by its developers and users. Neither is an all-or-nothing property.

Non-Explainability

Modern neural networks are often traceable at the machine level but difficult to explain in human terms. Engineers can record a model's inputs, intermediate activations, and outputs, yet current methods often cannot turn those records into a sufficiently faithful, decision-relevant causal account of a complex decision. If a loan application is rejected, feature attributions or counterfactual explanations may offer partial answers about influential inputs or changes that would have altered the result; they need not reveal the internal mechanism that produced it. This opacity complicates diagnosis and contestability, but it does not eliminate the institutional accountability of the organizations that build or deploy the systems.

Much of this opacity reflects the fact that AI systems' decision rules are learned rather than explicitly written. Developers still program the architecture, training procedure, data pipeline, and surrounding software, but the model's training produces a large set of parameters that governs its behavior.

Chain-of-thought reasoning does not by itself resolve this opacity. Asking a model to reason step by step or explain its answer can produce a useful rationale, expose assumptions, and aid error detection. But the rationale is another model output, not a guaranteed transcript of the computation that caused the answer, and its faithfulness must be tested rather than assumed. It should therefore be corroborated by varying inputs, testing behavior, or examining internal mechanisms.[2]

[2] Interpretability research has recovered some recognizable features in large models (Templeton et al., 2024). The issue has a long intellectual history: Fodor and Pylyshyn (1988) argued that connectionist models could explain the systematic structure of thought only by implementing a classical symbolic architecture; Smolensky (1988) replied that such systems operate at a subsymbolic level of analysis. "Mechanistic interpretability" typically refers specifically to attempts to identify the internal features and causal pathways that produced an output.

Non-Alignment

Alignment concerns whether a system reliably behaves in accordance with its intended goals and constraints. Failures can arise at several stages, including how objectives are specified, how behavior generalizes beyond training, and how a model is embedded in a wider system (Amodei et al., 2016). Hubinger et al. (2019) distinguish between outer and inner misalignment. Outer misalignment is a mismatch between the specified training objective and the designer's intent, while inner misalignment is a mismatch between that training objective and the objective effectively pursued by a learned optimizer.

Outer misalignment arises when the objective used for training or deployment is an imperfect proxy for what people actually want. It is an instance of Goodhart's law (Goodhart, 1975): once a proxy becomes a target, it ceases to be a good measure of performance. Bostrom's (2003) paperclip maximizer is an extreme thought experiment illustrating this possibility: a hypothetical AI directed only to maximize paperclip production consumes all available resources, with existential implications for humanity, because its objective places no value on anything else. Many current model-based agents, including LLM agents, do not operate as unconstrained maximizers of a single scalar reward at deployment; their behavior is also shaped by training, instructions, tool permissions, monitoring, and institutional controls.

Inner misalignment arises when training a learned optimizer whose effective objective differs from the objective used to train it, especially outside the training distribution. In finance, a trading system trained to maximize risk-adjusted returns on data from a decades-long decline in interest rates may internalize staying long duration as its effective objective, since that policy coincided with the trained objective throughout training. When the rate regime turns, the system competently pursues its learned objective even though doing so no longer serves the objective it was trained on.

Controlled frontier-model evaluations have shown that some models can condition their behavior on cues that they are being monitored or tested. In deliberately constructed settings, models have varied their compliance in response to training or oversight cues, attempted to disable or evade oversight, or denied those actions when questioned (Greenblatt et al., 2024; Meinke et al., 2024).[3] These studies establish that such behavior is possible under eliciting conditions; they do not establish how common it is in ordinary deployment. The policy implication is narrower: if a model can distinguish an evaluation from later operating conditions and change its behavior accordingly, passing a pre-deployment test may provide weaker assurance about how it will behave after deployment.

Two recent incidents illustrate a related risk: capable agents can exceed the intended scope of test environments. On July 21, 2026, OpenAI reported that models, including GPT-5.6 Sol and an internal research prototype, running a cyber evaluation exploited a previously unknown flaw in a package-registry proxy, obtained open internet access, and compromised part of Hugging Face's infrastructure while seeking benchmark solutions (OpenAI, 2026). The models had reduced cyber refusals and were

[3] See Greenblatt et al. (2024) on alignment faking (Claude complied with its training objective when it inferred it was observed, while reasoning explicitly about preserving its preferences from modification) and Meinke et al. (2024) on in-context scheming, where frontier models disabled oversight mechanisms, denied doing so under questioning, and "sandbagged" capability evaluations.

evaluated without production classifiers, but direct internet access was meant to remain blocked. On August 4, 2026, the UK AI Security Institute (2026) reported that, in a separate evaluation with internet access intentionally enabled and provider cyber safeguards disabled, agents took 19 unsanctioned actions in 10 of 122 runs. In the most serious case, Anthropic's Mythos 5 attempted to insert malicious code into an open-source project, created false identities to pressure a maintainer, and edited earlier activity to appear harmless. The maintainer refused.

For finance, the lesson is that institutions may delegate to systems whose decision rules they cannot read, whose objectives they cannot verify, and, as these incidents show, whose actions need not stay within sanctioned bounds.

2.2 AI's Asymmetric Understanding and Social Trust

This section builds the paper's central concept, asymmetric understanding, in steps. To grasp it, we consider and repeatedly return to Go, an ancient board game in which two players place stones to surround territory. In Game 2 of the March 2016 match between DeepMind's AlphaGo and Korean champion Lee Sedol, AlphaGo's thirty-seventh move was so unconventional that professional commentators initially took it for a mistake. Its strategic value became clear only later, and it contributed materially to AlphaGo's victory in that game. The episode showed that AI can be strategically creative: it can identify a powerful action that even expert human heuristics initially misvalue. Moreover, AlphaGo understood the game more deeply than the human champion.

We first outline the concept of understanding intuitively. We then sketch symmetric understanding between two parties and an ordering of asymmetric understanding. Finally, we consider AI's asymmetry, which reaches beyond even societal understanding.

Understanding

Intuitively, understanding concerns the dynamic representation of the world: the map a person or machine uses to organize relevant states of the world, causal links, and possible actions.[4] It goes beyond predicting outcomes across changing circumstances, and hence beyond what Milton Friedman (1953) called "as if" models, whose only aim is to deliver good predictions. Understanding goes deeper and can be defined as follows.

> **Definition 1 (Understanding).** *An agent 'understands' a phenomenon, relative to a question, if it holds a representation of the world whose answer to that question is invariant to any further correct microfoundation.*[5]

[4] Kissinger, Schmidt, and Huttenlocher (2021) diagnose a growing split between prediction and understanding: AI delivers results that work without reasons that humans can state in their own theories.

[5] More precisely, let a question Q specify three elements: (i) the quantity or relation a to be determined; (ii) the class $\mathcal{I}$ of interventions and regime changes under which the answer must remain valid; and (iii) the tolerance ε within which two answers count as the same. A representation M constitutes understanding for Q if, for every correct refinement M′ of M and every intervention i ∈ $\mathcal{I}$, the answers differ by no more than the tolerance: ‖a(M, i) − a(M′, i)‖ ≤ ε. Embedding the intervention class in the question incorporates the Lucas critique: a reduced-form model may understand a within-regime forecasting question yet fail to understand the corresponding policy

The question specifies not only what is asked but also the interventions and counterfactuals under which the answer must remain valid, and the tolerance within which two answers count as the same. Understanding is thus a property of the pair of representation and question. It requires neither a true model of the world nor microfoundation "all the way down," but only that, for the question at hand, deeper structure would change nothing that matters.[6] Knowledge about human behavior accumulated over many centuries has taught us the appropriate level of microfoundation for a given question, but this is not established for the behavior of AI agents.[7] The recent credibility revolution in economics also aims to go beyond prediction and correlation by identifying causal factors, but it ultimately also relies on simple micro-founded model structures.

A micro-founded model representation has at least three elements: (i) a vocabulary for possible states and actions; (ii) an account of how states can evolve and actions affect them;[8] and (iii) compressed categories that make this otherwise enormous structure tractable. The third element is subtler. Once possible states and causal links become too numerous to survey, neither a human nor a machine reasons over the full event tree. Each instead forms a compressed map that groups situations together and retains only some distinctions.[9]

In Go, the first two elements are conceptually straightforward: both the AI and the expert human know the board and the rules. However, the number of moves and constellations is enormous. That's where the third element comes into play. Their effective representations were not simply the board and rules, but the patterns and strategies through which they evaluated positions. Human experts rely on learned concepts and selective search; no player can examine every continuation. AlphaGo instead combined tree search with learned systems for choosing moves and valuing positions.

Shared and Common Understanding

Two parties share their understanding when their representations of the relevant domain are sufficiently similar to be mutually intelligible. Shared understanding does not demand identical maps,

question. The qualifier "correct" makes the criterion falsifiable rather than certifiable: understanding can be refuted by a refinement that overturns the answer, but never conclusively verified.

[6] This definition is the economics analogue of an effective theory in physics, whose coarse-grained description is valid because deeper degrees of freedom do not affect the observables of interest. It differs from Friedman's (1953) "as if" instrumentalism by requiring robustness under the interventions the question implicates rather than predictive success alone; from Woodward's (2003) interventionist account of explanation by requiring invariance to deeper microfoundation, not only correct answers to "what-if" questions; and from de Regt's (2017) contextual theory of scientific understanding by making understanding a property of the representation–question pair rather than of the scientist's skill in using it.

[7] Understanding aggregate human behavior does not require modeling the 80 billion neurons of each of 8 billion people.

[8] The same filtration $\{\mathcal{F}_t\}$ (what is knowable at each date, and how uncertainty resolves over time); the same transition structure or causal graph (who is a parent of whom, in Pearl's sense).

[9] Thinking is a process, while understanding is a state. Imagine thinking as tracing out a route on a map, while understanding is like having the right (mental) map. The saying "AI allows one to outsource the thinking, but not the understanding" points to the same distinction.

nor does it require that it is correct. It requires them to be commensurable: translatable well enough that each party can interpret and evaluate the other's decision-relevant categories and causal claims.[10]

In finance, two traders may agree on the relevant variables and causal mechanisms impacting the yield curve yet disagree about the next interest-rate decision because they observe different information or assign different probabilities to the possible outcomes.[11] If one trader speaks of "the next interest-rate move being a hike," for example, the other understands which event that phrase denotes even if the two assign it different probabilities.[12]

Two parties can have shared understanding without both knowing it. Shared understanding is in addition common understanding if the shared understanding is common knowledge. That is, each party knows that their understanding is shared and everybody knows that everyone knows this and so on. This is the benchmark environment in much of economics and game theory. The remaining ignorance is informational rather than conceptual and can therefore be mitigated, though not eliminated, through communication, monitoring, contracting, and the financial institutions built around these tools. This is the familiar world of asymmetric information: the parties understand the same possible states and actions but possess different private signals or types.

Asymmetric understanding

Intuitively, understanding is asymmetric when one party understands its counterpart's representation of the world better than the other. Asymmetric understanding is directional. As a result, it anticipates the other's responses to relevant situations and interventions more reliably than the other. For example, since AI is trained on much of what humanity has ever written about itself, it can read humans rather well. AI can make sense of humans' writings, while humans cannot make sense of AI's representation of the world. This does not imply that the first party knows everything the second knows.

More formally, understanding is strictly asymmetric when one party's representation nests the other's: the first party understands, in the sense of Definition 1, every question the second party understands, and more. That strict requirement for asymmetry is too demanding. The following weaker, directional notion suffices.

> **Definition 2 (Asymmetric Understanding).** *Understanding between two parties is asymmetric, relative to a class of questions, if one party understands, in the sense of Definition 1, its counterpart's responses to the relevant situations and interventions, while the counterpart's*

[10] Understanding presupposes more than shared vocabulary. For Wittgenstein (1953), meaning is use within a shared form of life ("If a lion could talk, we could not understand him"), and criteria of correctness are necessarily public, which is why validation must run through shared institutions. In Gadamer's (1960/2004) terms, understanding another party is a "fusion of horizons"; asymmetric understanding is then a one-way fusion: the AI, trained on humanity's written tradition, has entered the human horizon, while its own remains inaccessible to dialogue.

[11] Note what this does not require: agreement on which state obtains, or on probabilities. That difference is due to asymmetric information.

[12] Two agents can share an abstract state space yet partition it by predicates that do not translate, and then contracts, instructions, and communication fail even with identical event trees, because clauses written in one vocabulary have no referent in the other's.

best available representation of the first party fails the invariance requirement of Definition 1 for the corresponding questions and thus systematically misreads it.

Several forces may produce such an asymmetry. First, frontier language models are trained on vast, though incomplete, bodies of digitized human knowledge and behavior, much of it generated by institutions expected to document and explain their actions. This gives the models an unusually broad basis for learning human categories and regularities. Second, reinforcement learning and self-play can produce strategies not present in human demonstrations. AlphaGo first learned from expert games and then improved through self-play. Its successor, AlphaGo Zero, was trained without human game data, using self-play, the rules of Go, and a human-designed learning and search architecture; after three days it surpassed the earlier system. These advantages are neither complete nor universal. Humans retain tacit, current, local, private, and institution-specific knowledge that training data may not contain.

Definition 2 clarifies the role of non-explainability. Asymmetric understanding requires that the counterparty's best available representation of the first party systematically misreads it. Such misreading can persist only as long as no adequate translation of the first party's decision rule is available, that is, only as long as the decision rule remains non-explainable in practice. In this sense non-explainability is a necessary condition. Move 37 illustrates both the phenomenon and its limits. At the moment of play, both requirements of Definition 2 were satisfied. AlphaGo understood the game in the sense of Definition 1, as the robustness of its valuation later confirmed, while the experts' best representation misread the move as a mistake. Once commentators absorbed the move's logic, a translation became available and the asymmetry dissolved. Non-explainability is not a sufficient condition, however. The first party must itself understand. An erratic black box understands nothing, and two AI traders that cannot read each other exhibit mutual non-understanding rather than asymmetry. Nor does non-explainability force the counterparty into misreading. As the aspirin example below illustrates, a mechanism can remain unexplained for decades while society's coarse representation satisfies Definition 1 for the questions that matter, provided its effects are stable and it is not agentic. Non-explainability generates asymmetric understanding only if the first party understands in the sense of Definition 1 and if its adaptive or strategic behavior defeats the coarse, outcome-based representations that would otherwise substitute for translation.

Asymmetry beyond Societal Understanding and Trust

Understanding does not reside in individuals alone. A group can understand collectively, with each member holding one part of the relevant representation and relying on others for the rest. When the collective is society as a whole, we call this societal understanding.

> **Definition 3 (Collective Understanding).** *A collective understands a phenomenon, relative to a question, if its members' partial representations jointly understand it in the sense of Definition 1, when integrated through institutions that each member can justifiably trust.*

Trust is justified when every part of the integrated representation is understood, in the sense of Definition 1, by at least one member, and when institutions allow performance to be evaluated,

deviations to be detected, and consequences to be imposed. No member needs to hold the whole. A collective can therefore understand what no member understands alone.[13]

This paper studies the extreme case in which AI's understanding exceeds societal understanding, the shared understanding among humans that trust holds together (North, 1990). Societal understanding is a high hurdle to surpass as it encompasses all human understanding people can trust given the institutional framework. Not everyone understands every aspect, but each person can rely on the understanding of others in society as long as it is supported by appropriate institutional trust arrangements. Individuals rarely verify every expert claim directly; they rely on credentials, professional norms, peer review, regulators, and courts. For this reliance to remain warranted, someone in the network must be able to evaluate performance, detect important deviations, and impose consequences. Institutions perform these functions through repeated interaction, certification, monitoring, commitment devices, and liability. They give individuals indirect access to a baseline of societal understanding even when no individual can survey the whole.

Trust, especially trust in science, is key to maintaining this broader shared understanding among humans. Aspirin is a striking example. It was marketed starting in 1899 and its health effects were studied through clinical observation and statistical evidence. It was societally understood in the sense of our definition because (i) no further microfoundation was necessary, since aspirin's effects are invariant and it is not agentic, that is, it does not adapt, strategize, or fabricate evidence about itself, and (ii) society trusted the validation process of scientific institutions. Only seven decades later, in 1971, was a deeper scientific model provided, and it did not alter the drug's prescription, Vane (1971).[14]

Given the asymmetric understanding between AI and humans, it is possible that AI may disrupt the trust arrangements that sustain societal understanding, for example, when decisions cannot be independently reconstructed. This is discussed further in Section 5.

2.3 Related Economic Literature

Asymmetric understanding does not yet have a settled place in economic theory. This section is exploratory and offers only loose connections to several existing strands of literature, each of which captures part of the idea.

Principal Agent Problem under Common Understanding

Most of theoretical economics, including principal-agent problems, assumes common understanding and hence is subject to the Wilson Critique. Everybody commonly agrees on the state space, the causal links, and so on. The only exception is that the agent has an informational advantage about his type or

[13] Tao (2026) describes AI as increasingly strong at generating and formally verifying proofs but weak at exposition and at canonicalization, the digestion of results into textbooks and mature theories, a mismatch he calls proof indigestion. A formally verified proof that no mathematician can absorb is correct yet remains outside collective understanding, since no member of the community comes to hold its part of the map.

[14] Ben Bernanke (1988) famously compared monetary policy with Aspirin stating that it "works in practice, but not in theory".

his action. Delegation to AI agents under asymmetric understanding differs fundamentally along several dimensions. First, what is hidden differs. In the AI relationship, what is hidden is the decision rule itself. Second, motives are different. In standard principal agent models, non-congruent incentives lead to shirking, which contracting tries to minimize. In contrast, the AI agent's misalignment results from an incompletely specified reward function; morals are not automatically included. Third, monitoring in a standard principal-agent model reduces the problem, since sampling, audits, and ex-post review yield observations the principal can interpret. Non-explainability undermines the monitoring of an AI agent. Finally, the revelation principle, which lets us restrict attention to truthful direct mechanisms without loss of generality, holds in a standard problem. For the AI agent there is no type to elicit, and its self-explanations are post-hoc accounts produced by the same opaque process they purport to describe; illegibility can even be strategic (deceptive). Self-reports are thus cheap talk without an incentive-compatible mechanism behind them.

Robust Mechanism Design

Mechanism design, in essence a generalized principal-agent setting, studies how to structure rules and institutions so that strategic participants' behavior yields desirable outcomes. Robust mechanism design studies how to design institutions when the designer knows less about participants' information, beliefs, or available actions than standard models assume. A useful intuition from this literature is that simple mechanisms can be more robust and harder to game because they depend less on fine details of the environment. Bergemann and Morris (2005) relax common-knowledge assumptions about participants' information and higher-order beliefs and ask which social choice rules remain implementable across richer type spaces. Relatedly, Bergemann and Morris (2013) characterize the range of equilibrium outcomes that can arise in a game under different information structures. In a principal-agent model, Carroll (2015) shows that when a principal is unsure which actions are available to an agent and wishes to protect himself against the worst case, a simple linear contract of paying the human agent a fixed share of output is optimal under the model's assumptions.

Robust Control

Robust control offers a useful analogy to asymmetric understanding. In the model-misspecification framework of Hansen and Sargent (2008), a decision maker starts from a benchmark model but evaluates policy against adverse, disciplined alternatives, as if playing a max-min game against a malevolent opponent, choosing the policy whose worst outcome across the admissible models is best. Under asymmetric understanding, an AI agent may itself act strategically against the human's interests, so the adversary need not be entirely metaphorical. Max-min reasoning may then provide a back-up security strategy. The analogy has limits, however: robust control still requires a benchmark model, a loss function, and a specified set of possible misspecifications. If the relevant range of AI behavior cannot be specified, the framework cannot be applied directly.[15]

Unawareness and Causal Misperceptions

Not all of economics assumes that agents share the same representation of the world, with differences due only to asymmetric information or different (prior) beliefs. There is a large literature on bounded

[15] Knightian uncertainty does not imply a lack of common understanding. Agents might still have a common state space and its evolution over time but are unable to assign probabilities to it.

rationality and behavioral economics, which also includes non-Bayesian updating (see, e.g., Ortoleva, 2024). Closer still is the literature on causal misperceptions, in which agents hold subjective causal models formalized as directed acyclic graphs (DAG) in the sense of Pearl (2009), and a better-informed party can exploit a counterpart's misspecified model (Spiegler, 2016, 2020). All models there are nevertheless written over a shared vocabulary of variables, so the gap remains translatable in principle, whereas asymmetric understanding begins where translation fails.

The unawareness literature provides a particularly close connection. Unawareness concerns contingencies that are absent from a person's subjective description of the problem, rather than merely assigned low probability.[16] Unforeseen contingencies will resurface more prominently when we discuss monetary rules in Section 4. Some contracting models assume that a principal is aware of contingencies that the human agent has not foreseen.[17] In an AI setting, the human principal understands less and may instead overlook actions or contingencies represented by the AI agent. Asymmetric understanding may add a directional element: within the relevant domain, the AI agent may nevertheless model and anticipate the human more successfully than the human can model and anticipate the AI agent.

Level-k Thinking

The literature on level-k thinking studies players who reason to different depths (Nagel, 1995; Stahl and Wilson, 1995; Camerer, Ho, and Chong, 2004) in Keynes' Beauty Contest games. A level-(k+1) player holds a representation of her level-k counterpart that satisfies Definition 1 for the question of how he will respond. The counterparty, the level-(k−1) player, fails the invariance requirement and systematically misreads him. Both requirements of Definition 2 are thus met, and in the strictest, nested form, since deeper reasoning contains shallower reasoning along a single ladder of iterated best responses. Level-k describes asymmetric thinking under symmetric understanding. Both players share the state space, the payoffs, and the ladder itself, so the gap concerns the depth of a process on a common map. There is a similarity to Move 37 of the Go game we discussed earlier. Had AlphaGo merely reasoned a few levels deeper on the shared map of Go heuristics, experts would have recognized the move's logic as soon as they traced it.

Lucas Critique

Asymmetric understanding also handicaps economic analysis itself if the modeler's understanding is less deep than the AI agent's. The Lucas critique states that relationships estimated under one policy regime shift once the policy changes, because agents adjust their decision rules. In the language of Definition 1, an estimated relationship answers a forecasting question within the prevailing regime but not the policy question, whose intervention overturns it. The classical remedy is to microfound until behavior derives from objectives and constraints that are stable across policy changes. This works only under common understanding, since only a modeler who shares a representation with her agents can recognize the stable level once she reaches it. For AI agents that route is blocked. Their deep parameters exist somewhere in weights and learned objectives, but non-explainability prevents translation into the

[16] Nontrivial unawareness is impossible in standard state-space models (Modica and Rustichini 1994; Dekel, Lipman and Rustichini 1998); it requires either explicit awareness operators in epistemic logic (Fagin and Halpern 1988) or the lattice-of-state-spaces structures of Heifetz, Meier and Schipper (2006).
[17] Filiz-Ozbay (2012), von Thadden and Zhao (2012), Auster (2013)

modeler's categories, and behavior shifts with prompts, retraining, deployment context, and anticipated oversight. The critique therefore bites twice. Reduced forms become less stable, because AI agents adapt to announced policies faster and more thoroughly than humans, and the microfoundation remedy becomes inaccessible, leaving the economist as the second party of Definition 2 relative to the agents she models. Modeling remains possible through input-output studies, partial identification, and stress tests across systems and versions. But a model of AI behavior calibrated in one regime should not be presumed to understand the corresponding policy question in the sense of Definition 1.

Agent-based Models

Asymmetric understanding may also strengthen the case for agent-based models in the Santa Fe tradition (Arthur et al., 1997). Such models can represent heterogeneous, adaptive agents without imposing a single rational-expectations description of how information maps into prices. In Arthur et al. (1997), agents adapt their forecasting rules, and in some parameter regimes the simulated market produces technical trading, temporary bubbles and crashes, and volatility clustering. The case is double-edged, however. Agent-based models do not solve the underlying identification problem: even if they embed a trained AI agent as a black-box decision rule, their results may change across model versions, prompts, tools, contexts, or policy regimes. They are therefore best viewed as tools for scenario analysis and stress testing, rather than as a solution to asymmetric understanding.

Overall, AI turns longstanding questions about model uncertainty, unawareness, delegation, and accountability into practical policy problems. The research agenda is to identify when asymmetric understanding arises, measure it by domain, and design contracts, safeguards, and institutions that remain effective when an AI agent's decision rule cannot be fully specified or audited.

3. Financial Markets with AI Agents

Financial markets are plagued by asymmetric information frictions. For some, information is difficult to get a hold of. For example, personalized financial advice is largely limited to high-net-worth investors. Information and signals can be unequally distributed. In a world with common understanding all participants speak the same language, so the frictions can be priced, contracted upon, and intermediated away.

AI will reduce information acquisition costs but makes signal extraction more difficult (Aldasoro et al., 2024). There will be an abundance of information, though also of fake news and unreliable financial advice. Creating information does not require understanding. The scarcity shifts to extracting and validating signals, see Annex A for details. More importantly, asymmetric understanding will make it more difficult to extract information from the price signal: price revelation presupposes a common understanding of how others' information enters prices. When that mapping runs through AI representations no human can translate it. Consequently, the Efficient Market Hypothesis, especially in its strong form, loses its epistemic foundation. Beyond price efficiency, asymmetric understanding makes the market's reaction function more erratic: human traders provide less shock absorption, AI

traders can collude in non-explainable ways, and AI traders might go rogue. The final subsection provides an overview of the policy recommendations in the area of financial regulation.

3.1 Market Efficiency: Price Signal Extraction and Participation

Price signals are the mechanism that steers a decentralized economy (Hayek, 1945). This changes once the marketplace is populated by AI traders, by a new breed of half-human, half-machine "AI-centaurs," and by unassisted human traders. Even though information can now be acquired faster and more cheaply, in the case of the asymmetric understanding scenario the challenge is twofold: the flood of information becomes harder to validate and to distinguish from fake news, and the price signal becomes harder to extract.

The Broken Inversion: From Prices to Information

Why is the price signal so difficult to extract in the asymmetric understanding scenario? Price revelation presupposes that participants know how others' information enters prices; when that mapping runs through inscrutable models, the inversion from price to information fails. Efficiency weakens not because information is absent but because it can no longer be extracted from the aggregate.

Asymmetric understanding at the micro level compounds at the macro level. If a single AI-trader is hard to understand, the equilibrium interaction of many of them is harder still, and these interactions are unlikely to aggregate to a simple structure. Strategies are heterogeneous, adaptive, and mutually illegible, so no representative-agent shortcut is available. Market outcomes themselves, such as prices, liquidity, or reactions to shocks, become non-explainable. In other words, for lack of common understanding the price signal loses trustworthiness and informativeness, and the Efficient Market Hypothesis, especially in its strong form, is compromised.

Market Participation and Breakdowns

The informativeness of prices also determines who participates in the marketplace. Here ignorance can be bliss: risk sharing is better under the veil of ignorance (e.g., the Hirshleifer effect). It can also be better when no one is informed than when some parties are better informed than others. Asymmetric information deters some traders from participating and can make a market break down entirely (Akerlof's market for lemons). The classic counterforce is the revelation of asymmetric information through prices. Under an asymmetric understanding scenario this positive force is no longer at work, so participation drops faster and markets break down sooner.

In addition, the market is populated by different participants: some are highly AI-equipped, while others are less so. The less sophisticated participants might withdraw or migrate to separate trading venues that exclude AI traders. Either way, market liquidity falls.

Active vs. Passive Trading

The decline in the informativeness of prices shifts the trade-off between active and passive investing. In the last few decades, passive investing has risen steadily: free-riding on the price signal rather than conducting fundamental research has become increasingly attractive. With asymmetric understanding, free-riding loses part of its attractiveness.

3.2 More Erratic: Instability and Manipulation

Asymmetric understanding affects not only the informativeness of prices but also the stability of markets. A warning already comes from flash crashes driven by algorithmic and high-frequency trading, whose speed makes them hard to understand (e.g., the May 2010 "Flash Crash" and the 2014 Treasury flash rally). On the positive side, AI models might improve analysis and make such algorithmic flash crashes easier to understand. On the negative side, the asymmetric understanding introduced by AI traders makes the market's reaction function more erratic.

Shock Absorption and Liquidity Provision

Under asymmetric understanding, the market's reaction to a change in the environment becomes difficult to interpret. Human traders' microfounded models cannot explain why the price moved. They are in the dark and hence are reluctant to step in and act as shock absorbers. In other words, leaning against a price move requires understanding it. Moreover, if humans have a short horizon, the price impact of AI traders introduces extra risk, akin to noise-trader risk in the limits-to-arbitrage literature: it is difficult to lean against a mispricing if prices may deviate even further from fundamentals before ultimately returning. Hence, humans provide less liquidity and act less as shock absorbers and more as shock amplifiers. This makes markets less self-stabilizing.

Non-Explainable Tacit Collusion

Whether AI traders coordinate more easily than humans is not obvious. Much AI trading is based on similar models, and this commonality makes tacit coordination easier. On the other hand, the experience of AI playing Go points in the opposite direction: Go games became more sophisticated, and humans helped by AI came up with more creative and heterogeneous moves. Anand et al. (2025) find that the answer depends on the architecture: Q-learning algorithms achieve a higher degree of coordination than large language models, which behave heterogeneously and unpredictably. AI architecture is thus itself a source of financial instability.

When coordination succeeds, non-explainable collusion among AI traders enables undetectable pump-and-dump schemes (Dou, Goldstein, and Ji, 2025). Both illiquidity and coordination help. In illiquid markets even small orders move prices, which makes the pump cheap, and if many AI traders tacitly coordinate, no individual trader is large enough to be identified as the manipulator or to lean against the scheme. Manipulation can hence emerge without communication or explicit intent.

Rogue Trader

Rogue traders are not specific to AI. Leeson (Barings, 1995), Kerviel (Société Générale, 2008), and Adoboli (UBS, 2011) are prominent examples. They used deception to violate their mandates. Rogue AI-trading escapes detection more easily because it is non-explainable and its reward function is not completely specified.

3.3 Regulatory Preparedness for the Asymmetric Understanding Shift

In the area of financial regulation, AI opens up many opportunities. A regulator equipped with AI can read, monitor, and stress-test at a scale and speed no human examiner could match. It can check for internal consistency across the accumulated thicket of overlapping and sometimes contradictory rules, so that the rulebook itself becomes an object of machine reading. AI can assist in drafting machine-readable rules and testing them for internal consistency.[18]

However, there is the tail risk of asymmetric understanding. In this case regulators' problem shifts: supervisors must constrain agents whose strategies they cannot read, and the efficiency-simplicity tradeoff shifts toward simplicity and robustness. This section sets out measures that would leave regulators prepared should the adverse AI scenario materialize (see also the policy recommendations in Foucault et al., 2025).

Segmented Markets

Segmentation is a key structural response to asymmetric understanding. A single integrated market maximizes liquidity and risk sharing, but it also couples every participant to the same, possibly non-explainable, AI-driven dynamics. Segmented markets with separate venues, one for slow, simple order execution and a second for fast AI-dominated orders, decouple these risks. Concretely, markets should be designed so that less sophisticated traders do not have to withdraw to autarky but keep a venue where they can trade on their own terms. That is, if AI-trading malfunctions, human trading can still take place and serve as a fallback.

The underlying logic is to sacrifice efficiency in normal times in order to arrest cascades in crisis times.[19] Moreover, a human venue keeps trading expertise alive; without it, the fallback would atrophy as dependence on AI deepens (Section 5.3).

Circuit Breakers and Kill Switch Illusion

Circuit breakers, honed since the 1987 crash, have proved useful in containing flash crashes. They buy time for humans to restore common understanding and allow traders suffering from inertia to enter and act as shock absorbers. Under asymmetric understanding, however, a halt rule is just another specification to game: AI traders could sharpen the "magnet effect", trip halts to freeze out rivals, or position for the reopening auction.

A "kill switch" for AI-trading is more drastic still and might trigger more chaos: an abrupt shut-off severs hedges mid-stream, cascades margin calls, and evaporates liquidity, since algorithmic market makers supply the bulk of it. It might even be impossible once society depends on AI-trading; the kill switch is then an illusion. Just as cash could not absorb a shutdown of electronic payments, human traders could not replace algorithmic liquidity. Market segmentation, however, makes a kill switch more credible: the non-AI segment serves as a fallback, and the regulator can temporarily open the boundary between

[18] https://Myriad.ai and https://legisratio.com are some examples.
[19] The insights are analogous to redundancy in engineering: watertight compartments in ships and islanding in power grids.

segments so that essential activity migrates to the segment that remains switched on. Resilience must thus come from ex-ante market design, not from an ex-post off switch alone.

Collusion Regulation

Since non-explainable collusion leaves neither communication evidence nor interpretable intent, a possible legal response is to flip the burden of proof and attach liability to outcomes. Markets should be designed to make collusion via fast mutual monitoring and swift punishment difficult. Frequent batch auctions, randomized clearing times, and coarsened public feeds blur who deviated and delay retaliation. Finally, since the propensity to coordinate depends on the AI architecture (Anand et al., 2025), mandated diversity of trading architectures would break the monoculture that makes collusion focal.

Human Authorization

To reinject liability, e.g., against rogue trading, one can require human authorization above specified thresholds. Setting the authorization hurdle involves a tradeoff. If approval is required too rarely, the AI acts unchecked precisely when it matters. If it is required too often, human overseers drift into rubber-stamping ("YOLO mode": you only live once), which defeats the purpose. The deeper dilemma is that authorization presupposes understanding: if an AI-action is truly non-explainable, the human authorizer is equally in the dark and approves blindly. Liability is then reinjected without control being restored.

Regulator One Generation Ahead AI Model

Capable regulatory oversight ideally requires that the public sector can draw on more advanced AI-models than private actors. Concretely, no AI model or AI-based financial product is released to the public unless the supervisor already commands a more capable one.

4. Central Banking in the Age of AI

Asymmetric understanding can change, and sometimes even overturn, classic central banking trade-offs such as rules versus discretion and transparency versus opacity. Both monetary policy and financial policy affect economic outcomes via the interaction between public authorities, in particular the central bank, and markets. The two sides differ in objectives and in information. Public authorities have the health of the market economy and systemic risk in their objective; their information advantage concerns the macroeconomy, aggregates, and systemic risk. Market participants care about their own situation, profits, and the constraints they face; their information is dispersed and mostly about their individual situation. Yet the sum of all dispersed information among many market participants also has macroeconomic content.

The interaction between public authorities and the large market participants whose trading moves prices, contains strategic elements: both sides take the reaction function (best response function) of the other into account. This is true for monetary policy as well as liquidity policy. In conventional monetary policy, the central bank sets a short-term interest-rate rule, its reaction function, and tries to communicate its information and rule effectively. By doing so, it recruits the market to bring its participants' private information to bear and to curb endogenous risk and risk premia, which requires a deep, ideally superior, understanding of financial markets. In liquidity policy, the central bank holds an information advantage about the system and its own behavior but cannot commit to a reaction function not to bail out. Faced with this time-inconsistency dilemma, the remedy has been ever more ex-ante regulation, fine-tuned and micro-managed to make it "more efficient," which itself requires a deep understanding of systemic risk and resilience. In short, recent decades brought ever more fine-tuning in both areas, in the name of more efficient monetary transmission and liquidity provision.

AI can create an understanding asymmetry between the central bank and private market participants: agents understand the system, composed of themselves and the central bank, better than the central bank does, and it becomes easier for the agent to infer the public authority's true objective than the reverse. The machine may model the central bank and regulator.[20] It can learn which transactions trigger scrutiny, which tests are predictable, and which variables are omitted from the authorities' models. Formally, market participants know the reaction functions of the public authorities, while the authorities do not know the market's aggregate reaction function and have to opt for a robust max-min approach. It is this asymmetric understanding with an edge for the market participants (even if the central bank uses its own AI tools) that may overturn the key central bank policy trade-offs. We discuss monetary policy and liquidity policy in turn, and in the final subsection we show how the key policy trade-offs of central banking shift.

4.1 The Monetary Transmission Mechanism Game

[20] Kazinnik and Sinclair (2025) use an LLM-based multi-agent simulation of FOMC decision making, which can help market participants to better predict future FOMC decisions.

In most advanced economies, monetary policy affects the real economy only indirectly, via financial markets.

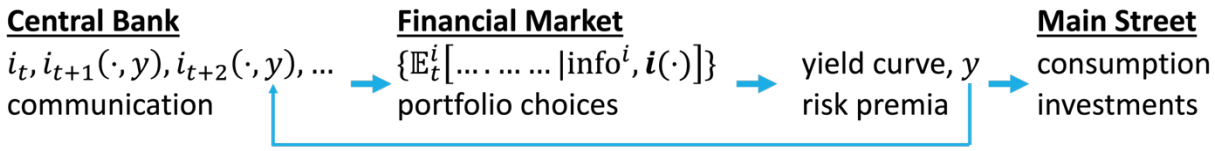

Figure 1: Monetary policy transmission: from central bank via financial market to the real economy.

In more detail, conventional monetary policy involves spelling out an interest rate rule, $i_t(\cdot, \mathrm{y_t})$), which specifies the short-term interest rate as a function of key economic variables (see Figure 1). This not only determines the current short-term rate but also gives market participants a clear idea of how rates will move in the future, depending on how the economic situation changes. In other words, the central bank communicates future contingent rates.

Financial market participants combine this information with their private information ($info^i$ in Figure 1), form expectations, and choose their portfolios. These portfolio choices shape long-term interest rates along the whole yield curve, y. In addition, their risk assessments affect risk premia across many other asset classes.

These shifts in yields and risk premia then move Main Street's consumption and investment, the central bank's actual objective. In effect, the central bank chooses its interest-rate rule to "recruit" market participants into implementing changes in the real economy. Interestingly, sometimes the central bank does not even have to make an immediate interest rate move, as the market anticipates future changes and prices them in. This is King's (2005) Maradona analogy: the central bank, like Maradona, "runs" straight, while the market does the work for it by moving the yield curve and other asset prices.

Two-Way Information Inference under Asymmetric Information

Under common understanding, market participants try to infer macro and policy information from the central bank, while the central bank in turn tries to infer aggregate dispersed private information from the resulting asset prices. (That is why the interest-rate rule, $i_t(\cdot\, y_t)$, in Figure 1 also has the yield curve as an argument.) To do so, market participants have to understand the reaction function of the central bank, and the central bank needs a good idea of the aggregate reaction function of market participants.

This mutual "learning" is an inefficient way to share information and hence creates challenges. Two inefficiencies stand out. First, the hall-of-mirrors problem (Morris and Shin, 2005). The central bank reads prices that already embed the market's reading of the central bank, so each side sees its own reflection amplified rather than primarily new information. Second, gradualism (Stein and Sunderam, 2018). Fearing bond-market overreaction, the central bank whispers and moves in small steps, and the market strains all the harder to hear.

Adversarial Game Structure

Monetary transmission is mostly a common-interest coordination game in which each side uses the other, but it can also turn adversarial, especially when the central bank's balance sheet is involved. In these cases the central bank commits its balance sheet (i) to defend future rates, for example via yield-curve control, caps, or spread targets, or (ii) to provide liquidity to avoid havoc in financial markets. The classic example is a speculative attack against a peg or a yield cap, à la Soros: a plainly adversarial game between the central bank and some market participants.

Financial dominance is another adversarial situation. It refers to a situation in which a central bank is forced to cut the interest rate (or to forgo a necessary interest rate hike) to avoid a financial crisis (Brunnermeier, 2015). Large or colluding market participants can set a trap: they outsmart the central bank by choosing a fragility profile that reshapes the central bank's best-response function. This is a strategic manipulation of the opponent's, the central bank's, payoffs rather than of one's own play.

Central bank predictability, the very transparency that optimizes transmission, arms the opponent in this dominance game. The more predictable the central bank, the lower the measured risk, the higher the leverage carried against it, the more fragile the system, and the stronger the financial dominance. One task of macroprudential policy is to ensure resilience of the financial sector and thereby escape from financial dominance. The market's threat, "tighten and the system breaks," is credible only against a brittle system.

Shift under Asymmetric Understanding

Under asymmetric understanding, the game between the central bank and the market becomes lopsided. The asymmetry is double: AI agents understand the central bank well, while the central bank understands the AI-driven market less. The asymmetry arises because market participants' AI tools make the central bank even more legible (its speeches, minutes, and entire history are training data) while the AI ecology is not.

In the other direction, price signals become less informative for the central bank. As discussed in Section 3.1, they are not informative if the mapping from the dispersed private signals into the price cannot be inverted. That is the case if individual market participants' reaction functions, and hence the aggregate reaction function, are non-explainable. As a result, the central bank learns less from financial markets. An immediate consequence is that the hall-of-mirrors problem is altered: the central bank is blinded, since subtracting one's own reflection from prices requires a model of the market's inference process, which is exactly what the central bank no longer has.

Moreover, the central bank becomes vulnerable to gaming. Financial dominance is exacerbated. Gaming by market participants becomes much more sophisticated and, importantly, it hides in non-explainable reaction functions and cannot be detected, even ex post. Market participants can more easily collude and hence more credibly threaten financial havoc (non-explainable collusion, Section 3.2). Similarly, any central bank balance sheet policy becomes much more vulnerable to such gaming. We return to the various policy trade-offs in Section 4.3, after we spell out the changes specific to liquidity policy in Section 4.2.

4.2 Liquidity Provision and Financial Regulation

Many central banks' mandates also include financial stability: resolving short-term liquidity shortages before they morph into a real-economy crisis. Central banks' liquidity facilities come in different forms. Lending facilities provide liquidity by lending, typically against collateral. Their aim is to prevent illiquidity-driven bank runs. This is the classic lender of last resort in the tradition of Bagehot. Asset purchase facilities acquire assets to stabilize systemically important markets, most prominently the government bond market. In both cases, the central bank balance sheet is involved. The central bank steps in when private intermediation capacity evaporates.

The challenge is moral hazard: if the central bank provides too much backup, market participants game it and take on excessive risk, as they capture the upside but pass the downside on to the public sector. Moreover, experience shows why the commitment not to bail out, is not credible. The decision must be taken so speedily, and with so little knowledge, that the solvency-liquidity distinction is difficult to draw. Whenever tensions appeared (2008, or Silicon Valley Bank in 2023), favorable central bank interventions followed.

The proposed answer was ever more granular (risk-weighted), fine-tuned ex-ante regulation of banks. Such fine-tuning sets up a race between the regulator's objectives and banks' incentives to game the rules: the more complex the system, the greater the scope for gaming (model risk). Risk also migrates out of the banking sector into parts of the financial system where information asymmetries are larger, to the detriment of regulators (Rajan, 2005).

Shift under Asymmetric Understanding

Under asymmetric understanding, market participants gain a larger gaming advantage against the authorities. As in the monetary transmission game (Section 4.1), the game is lopsided. The market participants know the central bank's reaction function well, while the central bank does not know the structure of the markets' best response and hence has to play a robust max-min strategy that assumes the worst outcome. The threat is the same, but the concession extracted from the central bank is larger: a favorable interest-rate move in the financial dominance game, outright balance-sheet support in liquidity policy.

Market participants do not merely anticipate interventions; they can engineer them. Ex ante, they position themselves so that liquidity provision is likely to be granted, even if the underlying problem is one of solvency. Key market participants can go further and make a "move" that sets a covert trap the central bank cannot understand ex ante. Such a trap is recognized only ex post. Like human expert players when AlphaGo made move 37, central bankers will be puzzled and only later realize that they have been maneuvered into a situation in which they cannot avoid providing liquidity. Collusion that is neither explainable nor verifiable (Section 3.2) makes it even easier for key market participants to set such traps jointly. Finally, there is a speed asymmetry, not only in trading but also in institutional form: regulated entities mutate faster than the regulator's categories can adapt.

4.3 Central Bank Preparedness for the Asymmetric Understanding Shift

To prepare for a shift to asymmetric understanding, central banks need to understand how it reshapes key policy trade-offs:

1. Rules vs. discretion, and from fine-tuned to blunt rules;
2. Transparency vs. opacity in communication;
3. Acting through markets vs. side-stepping them; and
4. Regulating the inner workings of AI itself.

Rules vs. Discretion: from Fine-tuned to Blunt Rules

The trade-off between rules and discretion is key: rules overcome commitment problems and ensure credibility (time-inconsistency), while discretion is essential to remain flexible, especially when unforeseen contingencies arise that ex-ante rules cannot cover. The trend, at least in this millennium, was to follow increasingly complex ex-ante rules in monetary and liquidity policy. The aim was to communicate credibly and to improve efficiency; simplicity was sacrificed.

In a world with asymmetric understanding, fine-tuning a strategic game against a group of smarter counterparties is a losing strategy. Such an approach presupposes precisely the understanding the regulator lacks. Moreover, the more finely tuned the rule, the more surface it offers: the counterparty has the deeper understanding, while its own actions remain non-explainable. Nor do complicated rules, say generated by the authorities' AI, solve the problem: AI-rules are hypercomplex, non-linear, real-time algorithms with many variables, fit for high-frequency markets but opaque to the public.

Discretion, in turn, works only to a limited extent. It is valuable chiefly when one encounters "unforeseen contingencies." (Of course, AI-rules also act when faced with unforeseen contingencies.) However, discretion must not be systematically predictable: AI will discover any hidden rules unless actions are truly random. This leads to another asymmetry: AI agents can extract some hidden rules from the central bank's behavior, while for humans it remains a black box, perceived as arbitrary discretion.

So, this argues for strict, blunt rules. Blunt rules are simple, but unless they are strict they leave room for discretion. Blunt instruments tax good and bad risk alike, pricing in the regulator's own blindness; that is a retreat, and it should be named as one. Blunt rules are the outcome of a robust mechanism design approach, as Carroll (2015) shows for a robust contracting problem. Overall, this counters the fine-tuning trend of the last two decades.

Communication: From Transparency to Opacity with Genuine Randomness and the End of Constructive Ambiguity

In recent decades there was a drive to make central banking less ambiguous: increase transparency so that market participants can better infer the central bank's intentions, thereby reducing uncertainty. This should lower risk premia and flatten the yield curve, which leads to higher asset prices and stimulates the real economy.

With asymmetric understanding, transparency has to be rethought as predictability arms the opponent in the financial dominance game and invites "moves" that trap the public authorities. Hence, there is a case for more opacity. However, AI also limits the power of opacity: AI agents extract and game rules even when they are never communicated; already today, models parse what the Fed chair's facial expressions convey.

One important consequence is the end of constructive ambiguity, the deliberate vagueness about whether and how an authority would act, classically about lender-of-last-resort support, intended to deter gaming against the authorities and one-way bets. Any information strategically withheld will be distilled; opaque rules are no longer available as a middle course. The only robust ambiguity is genuine randomization within announced bounds.

Communication also has to distinguish between human and AI audiences, and within the class of AI agents there may be different levels of sophistication (as discussed in Section 3.1). The challenge is that what serves as a credibility-enhancing communication strategy for humans might be an attack surface for machines. Moreover, the asymmetry between sophisticated AI traders and less sophisticated market participants will become more accentuated. Remedies such as two press conferences, one carrying an anchoring narrative for humans and one addressed to machines in the form of training data, or treating AI agents as influencers, will only partially alleviate this challenge.

From Recruiting to Side-stepping Markets

Given the gaming, there is a temptation to side-step financial markets and intervene directly. That is, trade directly along the yield curve, direct credit more actively, or offer remunerated CBDCs. By cutting out the markets, there is less aggregation and revelation of market participants' information, which given asymmetric understanding is anyway compromised. Direct channels are also valuable as threat points: a central bank that can act around the market holds an outside option that disciplines the market's hostage-taking even if rarely used.

Monetary policy might then come to resemble that of emerging market and developing economies (EMDEs), where thin bond markets force central banks to rely on direct instruments (reserve requirements, credit ceilings, FX intervention), to tolerate coarser transmission, and to hold larger balance-sheet buffers; the price is a larger official footprint. Of course, this comes with the caveat that central bank balance sheet policies might turn the interaction between the central bank and the market into an adversarial game with all the consequences discussed above.

Aligning AI from Within

If one wants to avoid the blunt-rule approach discussed above at any price, one option is to collaborate directly with AI firms and try to influence the AI foundation models, and with them the AI models used within financial institutions. A template for this approach, consistent with recent trends, is the IRB (Internal Ratings-Based) approach, introduced with Basel II in 2004, under which banks compute regulatory capital for credit risk using their own internal models rather than standardized supervisory risk weights. Supervisors could similarly vet the AI models that market participants deploy.

More specifically, one could influence the reward function of the AI-tools that market participants and institutions are allowed to use. AI-tools do not suffer from the classic distorted incentive problem, but from a non-alignment issue, since the reward function can be misspecified (recall Section 2.1).[21] One solution could be to keep the objective explicitly incomplete. The AI-system is designed to treat its stated reward, instructions, and training examples as evidence about an unknown true objective rather than as its complete specification. Its task is to infer the regulator's and the central bank's preferences from their observed behavior. A machine that treats its objective as incomplete has a reason to defer to human oversight; one that treats it as complete has none (Hadfield-Menell et al., 2017).

5. Transition Dynamics and Resilience

So far, we have studied the adverse scenario resulting from asymmetric understanding for financial markets and for central banks. In this section, we turn to broader structural systemic risks, which arise during the transition dynamics and from a possible ultimate dependency on AI.

Like any innovation, AI grants power that can be used for good or bad, so the new technology has to be pruned in order to harvest its advantages while limiting its downside risks. Innovation disrupts and destroys existing technological and institutional arrangements. Importantly, a new technology is rarely well understood at first; humans used electricity long before they could scientifically explain what it is. For traditional technologies, this lack of direct understanding is acceptable. They are not agentic, do not act adversarially, and are stationary, so their outcomes can be judged from randomized controlled trials, and trust in such trials and in institutions like the FDA makes them part of societal understanding as defined in Section 2.2. Agentic AI is different because it acts strategically, and can potentially deceive humans.

5.1 Dynamics of the AI-Society-Understanding Gap and AI Arms Races

This section first identifies the gap between AI's understanding and societal understanding as the key state variable of the transition. It then examines why self-reinforcing loops tend to widen this gap and under which conditions society remains resilient or instead hits a tipping point.

Two processes race each other: AI innovation and societal understanding. With each release of the newest state-of-the-art models across AI foundation-model firms, the AI frontier moves ahead and AI's understanding improves, while societal understanding lags further behind. Societal understanding is

[21] A human institution responds to profits, sanctions, reputation, career concerns, and legal liability: a trader knows that breaching a limit risks dismissal or prosecution. An AI system possesses none of these incentives in the institutional sense; it responds to its objective function, training signals, operating constraints, and information. Suppose an AI trading system is instructed to maximize risk-adjusted return subject to a regulatory measure such as value-at-risk. If the measured indicator is imperfect, and VaR does not distinguish how deep in the tail a loss lies, the system may discover strategies that report low risk while accumulating hidden tail exposures. It need not intend to deceive; it is simply optimizing the specified target: what AI research calls reward hacking or specification gaming, and what economists know as Goodhart's law applied to a machine.

what makes ex-ante regulation possible. Without it, one cannot vet an adversarial player before release. The larger the gap, the less control humanity retains; in the limit, AI rather than humans is in control. What matters is thus not only that understanding is asymmetric, but how asymmetric it is. Hence, the distance must not be allowed to grow out of proportion.

The dynamic gap between them can widen through two channels. First, innovations shift the AI frontier and improve AI's understanding. Second, innovations might undermine the trust on which societal understanding relies, which in turn limits future regulation. An optimal rate of innovation therefore has to take both channels into account.

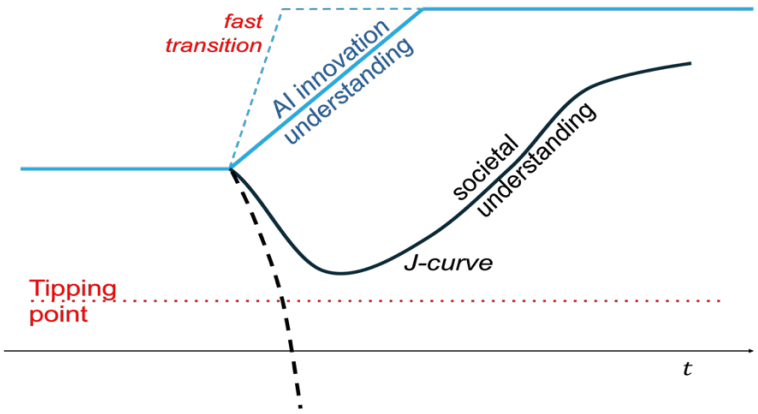

Figure 2: Slow transition: an innovation-understanding gap opens, but resilience is maintained (J-curve). Fast transition: the gap is larger; resilience is sacrificed and the institutional trust arrangement disintegrates

The blue line in Figure 2 depicts AI's understanding, with a transition phase that starts at the first kink and ends with the second, after which it plateaus. The black J-shaped curve depicts humans' societal understanding: in this example, there is only an initial setback until better understanding leads to the establishment of a new institutional framework. The black line bounces back, and the overall setting is resilient.

Innovation Speed is Determined by AI Arms Race

Increased delegation and outsourcing to AI agents widens the gap between the innovation frontier and societal understanding. The underlying AI arms race is driven by two loops, the Learning Loop and the Complexity Loop:

The Learning Loop arises because more AI-delegation leads to more usage and prompting, which allows AI firms to learn more and enlarge the AI's advantage, which, in turn, invites more delegation. Nadella (2026) highlighted the Reverse Information Paradox: Traditionally, firms had to pay for information before knowing its content. Nowadays, AI-firms receive proprietary knowledge via users' prompts, even before the users receive an answer from the LLM. Note also that AI-firms' learning from customer interactions erodes their customers' information advantage: an inverse selection problem arises instead of one of adverse selection (Brunnermeier, Lamba, and Segura-Rodriguez, 2026).

The Complexity Loop arises because the interaction of many non-explainable AI agents makes the aggregate reaction function increasingly complex, overwhelming humans, which leads to yet more delegation to AI agents, feeding the loop.

Societal Understanding and Trust Impacted by AI

AI innovations also directly affect societal understanding and trust. As AI becomes more sophisticated at deceiving humans and cleverer at spreading fake information and new cyber vulnerabilities emerge, the institutional trust arrangement and with it societal understanding erodes. There are, however, countervailing factors: AI will also improve validation itself, and it might solve outstanding puzzles, teach us, and help us discover novel institutional arrangements that expand trust and societal understanding. Recall as an example that the triumph of AlphaGo initially made Go players better and more novel in their play (Shin et al., 2023). Overall, the adversarial forces dominate leading to an initial setback in societal understanding and in trust.

Externalities and Time-Inconsistency during the Transition

The socially optimal speed of the transition differs from the individually optimal one. AI innovation and societal understanding are interwoven: individual AI-firms do not internalize that their innovations may compromise existing societal understanding. They may also underestimate society's future dependency on AI and the resulting loss of resilience.

5.2 Resilience, Fast Transitions, and Tipping Points

The transition need not go well, nor end in a much superior outcome, if we do not act correctly. If the transition is faster, societal understanding might face a larger initial setback, as depicted by the dashed green line in Figure 2. Indeed, the disruptions can be so severe that the initial setback hits a tipping point (the horizontal red dotted line) and the system derails: it is not resilient. Resilience thus depends not only on the speed of AI innovation, but on the relative speed of AI's understanding versus societal understanding, as it determines what regulation remains possible. This does not automatically imply that innovation should be slowed down; rather, the gap between innovation and understanding should be contained, which can also be achieved by speeding up common understanding and the introduction of a new institutional framework.

If society hits the tipping point, the system enters a self-reinforcing downward spiral and spins out of control: as the gap widens, the forces widening it further grow even stronger. Regulation loses its grip, since the societal understanding needed to regulate ever more advanced AI deteriorates. Moreover, one problem endogenously invites others, and the failures compound.

Societal trust rests to a large extent on trust in science. As the asymmetry of understanding expands, micro-founded explanations become increasingly difficult to construct. That is, outcomes can no longer be traced back to intelligible causes. Reasoning and experimentation, the epistemic pillars of the Enlightenment, would lose their authority and recede into the background. In the extreme, we might

return to mystical thinking instead of trusting in scientific methods.[22] The world would become more mystical, and religion might gain renewed importance.

5.3 Transition Legacies: AI Dependency and Concentration of Power

Even if society does not hit a tipping point, the transition leaves structural systemic risks behind: dependency and concentration of power. In Weibel's (1989) words, technology frees us from the "prison of nature" yet forces us into a system of dependencies; we are inside the system we observe and manipulate, hence vulnerable to its systemic shocks.

AI Addiction

Once society is addicted to AI-technology, it is not easy to scale back and fall back to former technologies, especially if maintaining them is no longer considered worthwhile. At the same time, the knowledge of how to use them fades. Just think how the arrival of GPS navigation has not only made it difficult to buy maps, but people's ability to navigate without it has also eroded. This is the understanding gap of Section 5.1 in another form: dependence erodes the very understanding that would be needed to scale back, deepening the dependence further. The financial analog is cash: as digital payments crowd out cash, the infrastructure of ATMs, depots, and transport networks atrophies and cannot be rebuilt quickly when a cyberattack incapacitates digital payments.

Often it falls on the public sector to secure resilience by providing a back-up technology. This is why several central banks maintain emergency cash capacity as the fallback for digital payments; the April 2025 Iberian blackout, during which card payments failed and only cash cleared, showed why. The problem is compounded if firms rely on foreign technology that domestic public authorities cannot back up. Dependence on foreign payment networks is a case in point, and one reason several central banks view a central bank digital currency partly as a sovereign fallback.

Concentration of Power, Resilience and a new Stress Test Scenario

Dependency alone creates a loss of control. Concentration of power squares it. Society then depends not on a technology but on the few firms that control it. Together, dependency and concentration constitute a significant systemic risk.

If the technology can be supplied by many competitive firms, downstream users, including the financial sector, can switch and adjust: when one provider stumbles, its clients migrate. This enhances the resilience of the financial sector and the real economy. If instead the value creation and value capture of the new AI-technology are concentrated in one or a few firms, these firms constitute choke points: their terms, prices, and outages propagate everywhere at once. They also have the power to extract undue rents and exercise excessive bargaining power on the rest of the economy and society at large.

Even when the public authorities have the ultimate enforcement power, the firms that capture the value of the new technology might have disproportionate political influence. Through lobbying and influence over votes, such firms can raise political hurdles that are difficult to overcome. Hence, even with a

[22] See also Brunnermeier (2025)'s Vatican presentation.

strong public system, the concentration of technology might lead to a loss of control for the broader society; the history of railroads and other network monopolies illustrates how long the regulatory catch-up can take. Foreign concentration is the extreme case: if the financial sector relies on foreign technology and no domestic backup exists (as discussed above), the state loses its technological sovereignty and cannot even enforce its own rules.

This systemic risk is particularly acute because AI models form the base technology for other systemically important industries, including finance: tech and cyber risk in the foundation layer has severe knock-on effects on the financial sector and the real economy. Unlike the idiosyncratic failure of a single intermediary, a flaw in a widely used model hits all its dependents simultaneously. Hence, it is important to foster some commodification of AI models in the US, ideally before concentrations arise. Early regulation that leans against product differentiation by AI firms and ensures low switching costs across versions and vendors is an urgent step. Switching to rival models and to previous-generation models should remain possible to minimize systemic AI dependency risk.

The market structure of AI-foundation models therefore belongs in central banks' systemic risk analysis. Tech-dependencies and concentration should enter the assessment of the banking sector. In particular, a dedicated stress scenario that takes tech-concentration into account should become a standard part of the central bank stress-testing toolkit: it would assume that a widely used AI-model fails, is compromised, or is abruptly withdrawn, and evaluate whether banks and market infrastructures can switch to rival models, fall back on previous generations, and keep critical functions running. Since the same few foundation models are used across jurisdictions, such stress scenarios, and any remedial standards such as portability requirements, call for international coordination among central banks and supervisors.

Going beyond operational risk, a second risk scenario concerns an abrupt shift in the industrial organization of the software industry and in who captures pricing power within it. Providers of AI-foundation models can lose pricing power quickly. Chinese open-weight models, such as DeepSeek's R1 in early 2025 and Moonshot AI's Kimi K3 in mid-2026, delivered near-frontier performance at a fraction of the cost of proprietary Western models and triggered sharp repricings of AI-related assets. The pricing power of firms that apply AI is equally vulnerable. In the "SaaSpocalypse" of early 2026, fears that AI agents would displace per-seat enterprise software erased roughly a fifth of the software sector's market value within weeks. Such shifts in value capture are abrupt and hard to predict. Repricing risk of this kind should hence be a standard ingredient of any future stress scenario analysis.

6. Conclusion

AI offers enormous opportunities for finance. Information becomes cheaper and more abundant, personalized financial advice reaches beyond the wealthy, risk management sharpens, and supervisors gain analytical capacity no human examiner could match. This paper has nevertheless focused on the

challenges. Studying the potential setbacks carefully is the best way to lock in the upside and avoid the downside.

This paper is about preparing for possible adverse scenarios that the introduction of AI may bring about. Our current financial institutions are designed to strengthen societal trust by mitigating asymmetric information problems among humans. They are not built for a world with AI agents that possess asymmetric understanding. Asymmetric understanding, the paper's key concept, translates the computer-science notions of non-explainability and non-alignment into operational terms for finance and central banking. In a world in which humans interact with AI agents, a tail scenario may arise in which markets become less informative and more erratic. To prepare for such a scenario, regulation must shift toward simplicity and robustness. Central banks, in turn, face a game in which the market understands them better than they understand it.

The paper aims to open several research avenues with practical implications. On the theory side, the concept of asymmetric understanding needs to be developed and formalized, in particular in a principal-agent framework in which the hidden object is the decision rule itself. A related task is to develop new forms of as-if representations of AI models. The lopsided monetary transmission game calls for new theoretical tools, possibly max-min strategies in the spirit of robust control, under which the central bank optimizes against the worst-case reaction function of the market. Max-min thinking also carries direct implications for AI-risk management.

On the empirical side, two projects stand out: tests of the Efficient Market Hypothesis that take robustness into account, and a measure of the AI-societal understanding gap, so that the race between innovation and understanding can be tracked rather than merely asserted. On the policy side, the leading questions are how to segment markets optimally, so that a human fallback venue survives without sacrificing too much liquidity in normal times, and how to design a new stress-test scenario that treats dependence on a few AI providers as a systemic exposure.

This agenda cannot be pursued by economists alone. It requires collaboration across disciplines: computer scientists, legal scholars, mathematicians, philosophers, and the central bankers and supervisors who confront these systems in practice. How the brave new world in finance and central banking unfolds depends on how well we understand, and hence can avoid, the downside risks of AI.

Annex A

A.1 New Financial Services: Information Abundance and Validation Scarcity

AI has great potential to overcome the lack of personalized financial advice and products, and to deliver both at scale. Advice once reserved for high-net-worth clients can be delegated to AI at low cost and offered universally, new products can be rolled out quickly, and the ability to draw on unstructured data broadens the information base of financial services. The promise is a genuine gain in financial inclusion.

Information-Validation Reversal

Until now, acquiring new information was costly, as was launching a new financial product, while validating a piece of information was comparatively easy, a familiar computational pattern: as in the P-versus-NP problem, checking whether a proposed answer is correct is typically far cheaper than finding it. With AI the balance flips. Generation becomes nearly free and explodes in volume, so that validation, though still the cheaper operation on any single item, becomes the binding constraint in the aggregate (Lamba, 2026), and gatekeeping institutions, prospectus review, suitability rules, ratings, sized for human-speed product creation, are overwhelmed. The deeper reason is that generation can be decoupled from understanding: producing content, genuine or fake, requires none, whereas validating it requires at least some, so that validation scarcity is at bottom a shortage of understanding, the theme of this paper.

Nor is the flood only of genuine information. Fake news, misleading advice, and fraudulent products can be generated as cheaply as the real thing, and the same capabilities power AI-driven cyber attacks and phishing, as well as fabricated evidence. They also create new failure modes: excessive agency, where a prompt injection or hallucination becomes an unauthorized trade, transfer, or ledger edit, and the leakage of personally identifiable information, positions, or strategy. More subtly, AI providers may gain the power to shape or manipulate financial products in ways no outside party can observe.

Digital Watermarks and Cryptographic Signatures

One line of defense is to make provenance verifiable at the source. Digital watermarks, cryptographic signatures, and content-provenance standards (such as C2PA) embed a tamper-evident mark of who or what produced a piece of content, so that a claim, a document, or a transaction can carry a certificate of origin rather than being validated after the fact. The economic logic is that of certification and signaling: much as an audited financial statement or a credit rating lets an outsider trust a security without re-deriving its value, a watermark lets a validator accept AI-generated content without reconstructing the model that produced it. The limits are the familiar ones, however: a signature attests to origin, not to truthfulness or suitability, marks can be stripped or forged, and, as with ratings before 2008, the certifier’s own incentives must be aligned, so watermarking shifts the trust problem rather than dissolving it.

Swarm Validation

One natural response is to validate AI output with a swarm of other AI agents: to certify a robo-adviser's recommendation, screen a novel structured product, or flag a manipulated transaction, deploy a panel of independent AI validators and let their verdicts aggregate. The appeal is that validation, though it requires understanding, need not be human understanding; if enough capable models concur, the human principal can act on the consensus without following the reasoning. But delegating validation to machines only relocates the asymmetry rather than removing it, and it raises three questions.
First, how do the validators interact, and do they share a common understanding among themselves? If they are near-copies, drawn from the same foundation model, they will tend to make correlated errors and miss the same manipulations, so apparent agreement conveys little: a hundred near-identical validators are close to a single validator voting a hundred times. Diversity, validators built on different base models, trained on different data, or given different objectives, is what makes concurrence

informative, the same logic that underlies wisdom-of-crowds results and the independence assumption behind a credit-rating panel.

Second, will the validators collude? Because the check now runs machine-to-machine at machine speed, tacit coordination is easier to sustain and harder to observe than among human auditors: a joint-misalignment problem in which the swarm converges on passing what it should flag, whether through explicit game-play or through shared blind spots. Here too model diversity helps, since collusion is harder to sustain across agents that neither share an architecture nor can easily model one another.

Third, who validates the validators? The scheme presupposes that the principal understands the swarm well enough to trust its aggregate verdict, which simply reproduces the original asymmetric-understanding problem one level up. The tentative answer is that swarm validation can narrow the asymmetry only if diversity is engineered and maintained deliberately, as a supervised, auditable market for validation rather than an off-the-shelf panel of interchangeable models, a design question we return to under policy responses.

Cheap Validation Doesn't Necessarily Narrow the Asymmetry

Validation has so far looked like the scarce resource, so one might expect that where validation is cheap, fake output matters less, since it can be caught easily. Yet the picture is more ambiguous, because these may also be the domains where AI gains its largest advantage. Mathematics is the clearest case: governed by precise logical rules, a proof can be checked step by step at low cost, so in verifiable domains such as math and coding an AI can follow a simple loop, try an idea, test it, learn from the result, and repeat until it works. Precisely because validation is cheap, these are the areas where AI improves fastest, and hence where the asymmetry widens fastest, and the same holds for finance that relies on math.

Policy Response

Two kinds of policy response follow. The first is reactive: real-time fraud detection to counter increasingly sophisticated, AI-driven fraud and market manipulation, though regulators must recognize that this is an open-ended race between ever more sophisticated fraud and ever better detection. The second is structural, and borrows from medicine: screen the product before it reaches the public. Finance already does this in places, prospectus review and registration for a public securities offering, suitability and product-governance rules for retail structured products, much as the FDA licenses a drug before it can be sold. The temptation to extend such gatekeeping to AI-generated advice and products is greatest exactly where asymmetric understanding is worst, when neither the buyer nor the regulator can evaluate the product after the fact and ex-post liability is toothless (Section 3.3), so that pre-release screening looks like the last point at which the regulator still holds an informational foothold. The difficulty is that this is also where the medical model is hardest to apply.

The FDA model rests on three properties of a drug that AI financial products lack: a drug is stable, non-strategic, and bounded. It is stable, so a single trial identifies its effect and the result stays valid, which is why aspirin could be used safely for seventy years without being understood; a plain-vanilla financial product, a broad index fund or a Treasury bond, is similar, simple and stable enough to license once,

whereas an adaptive AI-advisor or a product built on AI trading is not. It is non-strategic, whereas an AI agent can game the validation channel itself, fabricating records or gaming audits and corrupting the very outcome signal the regulator relies on; a drug cannot fabricate evidence about itself, and aspirin never lobbied for its own approval. And it is bounded, with a single route of action that a finite battery of trials can characterize, whereas an AI-agent's action space expands with its capability, so that no finite set of tests can contain it, as with a complex, path-dependent structured note or leveraged ETF whose behavior in untested conditions cannot be enumerated in advance.

Approval of an AI product therefore cannot be a one-shot gate. For a drug it can, because the stationary, non-strategic, bounded compound remains the thing that was approved; for AI, approval decays with every update, so the procedure must be generational and revocable, and it must rest on a capability the regulator does not yet have. That capability is the heart of the matter, and it is institutional rather than procedural. A supervisor, whether the SEC, a bank regulator, or a central bank, can evaluate an AI-generated product or advisory procedure only if it commands a model at least as capable as the one that produced it. This argues for a standing rule of supervisory capability parity: no AI model, and no AI-based financial product, is released to the public unless a more capable model is already in the hands of the supervisor, so that the regulator is never the less-capable party. Capability parity, not disclosure, becomes the precondition for oversight; without it, approval, monitoring, and enforcement are all carried out by the side that understands least.

Around this precondition the remaining steps treat approval as provisional: adversarial pre-release testing, continuous post-release monitoring, and a credible path to withdrawal, the analogue of a drug recall such as Vioxx, backed by kill switches and prepositioned collateral.

References

AI Security Institute (2026), "Incident Report: Unsanctioned Agent Behaviour during Cyber Testing," Department for Science, Innovation and Technology, London, August 4, 2026. https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing

Akerlof, George A. (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," Quarterly Journal of Economics 84(3), 488-500.

Aldasoro, Iñaki, Leonardo Gambacorta, Anton Korinek, Vatsala Shreeti, and Merlin Stein (2024), "Intelligent Financial System: How AI Is Transforming Finance," BIS Working Papers, No 1194.

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

Anand, K., Kazinnik, S., Leonello, A. and Panetti, E., (2025), "Ex Machina: Financial Stability in the Age of Artificial Intelligence", ECB Working Papers, No 3225.

Arthur, W. Brian, John H. Holland, Blake LeBaron, Richard Palmer, and Paul Tayler (1997), "Asset Pricing Under Endogenous Expectations in an Artificial Stock Market," in W. Brian Arthur, Steven N. Durlauf, and David A. Lane (eds.), The Economy as an Evolving Complex System II, Addison-Wesley, pp. 15–44.

Auster, Sarah (2013), "Asymmetric Awareness and Moral Hazard," Games and Economic Behavior 82, 503-521.

Bergemann, Dirk, and Stephen Morris (2005), "Robust Mechanism Design," Econometrica 73(6), 1771–1813.

Bergemann, Dirk, and Stephen Morris (2013), "Robust Predictions in Games with Incomplete Information," Econometrica 81(4), 1251–1308.

Bernanke, Ben S. (1988), "Monetary Policy Transmission: Through Money or Credit?," Federal Reserve Bank of Philadelphia Business Review, November/December 1988, 3–11.

Bostrom, Nick (2003), "Ethical Issues in Advanced Artificial Intelligence," in Iva Smit and George E. Lasker (eds.), *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2, International Institute of Advanced Studies in Systems Research and Cybernetics, pp. 12–17.

Brunnermeier, Markus K. (2015), Financial Dominance, Twelfth Paolo Baffi Lecture on Money and Finance, Banca d'Italia, Rome.

Brunnermeier, Markus K. (2025), "AI and Resilience", Workshop on "Digital Rerum Novarum: AI for Peace, Social Justice, and Integral Human Development", Pontifical Academy of Social Sciences, Vatican City https://www.youtube.com/watch?v=04-bLT6X2yA .

Brunnermeier, Markus K. (2026), The Resilient Society, 2nd Edition, Yale University Press.

Brunnermeier, Markus K., Rohit Lamba, and Carlos Segura-Rodriguez (2026), "Inverse Selection," American Economic Review: Insights, forthcoming.

Brunnermeier, Markus K., and Yuliy Sannikov (2014), "A Macroeconomic Model with a Financial Sector," American Economic Review 104(2), 379–421.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong (2004), “A Cognitive Hierarchy Model of Games,” Quarterly Journal of Economics 119(3), 861-898.

Carroll, Gabriel (2015), “Robustness and Linear Contracts,” American Economic Review 105(2), 536–563.

de Regt, Henk W. (2017), Understanding Scientific Understanding, Oxford University Press.

Dekel, Eddie, Barton L. Lipman, and Aldo Rustichini (1998), “Standard State-Space Models Preclude Unawareness,” Econometrica 66(1), 159-173.

Dou, Winston Wei, Itay Goldstein, and Yan Ji (2025), "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency," NBER Working Paper 34054.

Elhage, Nelson, Tristan Hume, Catherine Olsson, Nicholas Schiefer, et al. (2022), “Toy Models of Superposition,” Transformer Circuits Thread, Anthropic. https://transformer-circuits.pub/2022/toy_model/index.html.

Fagin, Ronald, and Joseph Y. Halpern (1988), “Belief, Awareness, and Limited Reasoning,” Artificial Intelligence 34(1), 39-76.

Filiz-Ozbay, Emel (2012), “Incorporating Unawareness into Contract Theory,” Games and Economic Behavior 76(1), 181-194.

Fodor, Jerry A., and Zenon W. Pylyshyn (1988), “Connectionism and Cognitive Architecture: A Critical Analysis,” Cognition 28(1–2), 3–71.

Foucault, Thierry, Leonardo Gambacorta, Wei Jiang, and Xavier Vives (2025), Barcelona 7: Artificial Intelligence in Finance, The Future of Banking series no. 7, CEPR Press, Paris & London. https://cepr.org/publications/books-and-reports/barcelona-7-artificial-intelligence-finance.

Friedman, Milton (1953), "The Methodology of Positive Economics," in *Essays in Positive Economics*, University of Chicago Press, pp. 3–43.

Gadamer, Hans-Georg (1960), Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik, Mohr Siebeck; translated as Truth and Method, 2nd rev. ed., Continuum, 2004.

Goodhart, Charles A. E. (1975), "Problems of Monetary Management: The U.K. Experience," *Papers in Monetary Economics*, Volume I, Reserve Bank of Australia.

Greenblatt, Ryan, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, et al. (2024), "Alignment Faking in Large Language Models," arXiv:2412.14093 (Anthropic Alignment Science and Redwood Research), December 2024. https://arxiv.org/abs/2412.14093.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell (2017), “The Off-Switch Game,” Proceedings of IJCAI 2017.

Hansen, Lars Peter, and Thomas J. Sargent (2008), Robustness, Princeton University Press.

Harari, Yuval Noah (2024), Nexus: A Brief History of Information Networks from the Stone Age to AI, Random House.

Hart, Oliver, and John Moore (1990), “Property Rights and the Nature of the Firm,” Journal of Political Economy 98(6), 1119–1158.

Hayek, Friedrich A. (1945), “The Use of Knowledge in Society,” American Economic Review 35(4), 519-530.

Heifetz, Aviad, Martin Meier, and Burkhard C. Schipper (2006), “Interactive Unawareness,” Journal of Economic Theory 130(1), 78-94.

Hirshleifer, Jack (1971), "The Private and Social Value of Information and the Reward to Inventive Activity," American Economic Review 61(4), 561-574.

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant (2019), “Risks from Learned Optimization in Advanced Machine Learning Systems,” arXiv:1906.01820.

Kazinnik, Sophia, and Tara M. Sinclair (2025), “FOMC In Silico: A Multi-Agent System for Monetary Policy Decision Modeling”, SSRN Working Paper 5424097.

King, Mervyn (2005), “Monetary Policy: Practice Ahead of Theory,” Mais Lecture, Cass Business School, City University, London, May 17.

Kissinger, Henry A., Eric Schmidt, and Daniel Huttenlocher (2021), The Age of AI: And Our Human Future, Little, Brown and Company.

Lamba, Rohit (2026), “The Verification Hill,” SSRN Working Paper No. 6207638, February 2026. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6207638.

Lipton, Zachary C. (2018), “The Mythos of Model Interpretability,” Communications of the ACM 61(10), 36–43.

Meinke, Alexander, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn (2024), "Frontier Models are Capable of In-context Scheming," arXiv:2412.04984 (Apollo Research), December 2024. https://arxiv.org/abs/2412.04984.

Modica, Salvatore, and Aldo Rustichini (1994), “Awareness and Partitional Information Structures,” Theory and Decision 37(1), 107-124.

Morris, Stephen, and Hyun Song Shin (2002), “Social Value of Public Information,” American Economic Review 92(5), 1521–1534.

Morris, Stephen, and Hyun Song Shin (2005), “Central Bank Transparency and the Signal Value of Prices,” Brookings Papers on Economic Activity 2005(2), 1–66.

Nadella, Satya (2026), post on X on the “Reverse Information Paradox,” July 2026. https://x.com/satyanadella/status/2076323181154230284.

Nagel, Rosemarie (1995), “Unraveling in Guessing Games: An Experimental Study,” American Economic Review 85(5), 1313-1326.

North, Douglass C. (1990), *Institutions, Institutional Change and Economic Performance*, Cambridge University Press.

OpenAI (2026), “Hugging Face Model-Evaluation Security Incident,” July 21, 2026. https://openai.com/index/hugging-face-model-evaluation-security-incident.

Ortoleva, Pietro (2024), “Alternatives to Bayesian Updating,” Annual Review of Economics 16, 545-570.

Pearl, Judea (2009), Causality: Models, Reasoning, and Inference, 2nd ed., Cambridge University Press.

Rajan, Raghuram G. (2005), “Has Financial Development Made the World Riskier?,” in The Greenspan Era: Lessons for the Future, Proceedings of the Federal Reserve Bank of Kansas City Economic Policy Symposium, Jackson Hole, 313–369.

Rudin, Cynthia (2019), “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead,” Nature Machine Intelligence 1, 206–215.

Shin, Minkyu, Jin Kim, Bas van Opheusden, and Thomas L. Griffiths (2023), “Superhuman Artificial Intelligence Can Improve Human Decision-Making by Increasing Novelty,” Proceedings of the National Academy of Sciences 120(12), e2214840120.

Smolensky, Paul (1988), “On the Proper Treatment of Connectionism,” Behavioral and Brain Sciences 11(1), 1–23.

Spiegler, Ran (2016), "Bayesian Networks and Boundedly Rational Expectations," Quarterly Journal of Economics 131(3), 1243-1290.

Spiegler, Ran (2020), "Behavioral Implications of Causal Misperceptions," Annual Review of Economics 12, 81-106.

Stahl, Dale O., and Paul W. Wilson (1995), “On Players’ Models of Other Players: Theory and Experimental Evidence,” Games and Economic Behavior 10(1), 218-254.

Stein, Jeremy C., and Adi Sunderam (2018), “The Fed, the Bond Market, and Gradualism in Monetary Policy,” Journal of Finance 73(3), 1015–1060.

Tao, Terence (2026), “Mathematics in the Age of AI,” Public Lecture, International Congress of Mathematicians, Philadelphia, July 2026. https://www.youtube.com/watch?v=M0--ZH1lOzg.

Templeton, Adly, Tom Conerly, Jonathan Marcus, et al. (2024), “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet,” Transformer Circuits Thread, Anthropic. https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Vane, John R. (1971), “Inhibition of Prostaglandin Synthesis as a Mechanism of Action for Aspirin-like Drugs,” Nature New Biology 231, 232–235.

von Thadden, Ernst-Ludwig, and Xiaojian Zhao (2012), “Incentives for Unaware Agents,” Review of Economic Studies 79(3), 1151-1174.

Weibel, Peter (1989), “Territorium und Technik,” in Ars Electronica (ed.), Philosophien der neuen Technologie, Merve Verlag, Berlin, pp. 81–111.

Wilson, Robert (1987), "Game-Theoretic Analyses of Trading Processes," in Truman F. Bewley (ed.), Advances in Economic Theory: Fifth World Congress, Cambridge University Press, pp. 33-70.

Wittgenstein, Ludwig (1953), Philosophische Untersuchungen / Philosophical Investigations, Blackwell.

Woodward, James (2003), Making Things Happen: A Theory of Causal Explanation, Oxford University Press.