Quantifying System-Level Harms From AI Adoption in Complex Sociotechnical Systems

Paul Vautravers[1], Oliver Chalkley[1], Gabriel Downer[1], Kate S[2], Damian Ruck[1]

[1]Advai Ltd
[2]UK National Cyber Security Centre

Abstract

Artificial Intelligence (AI) is increasingly integrated into complex sociotechnical systems, including Critical National Infrastructure (CNI), where harms emerge from interactions between technical, human, and organisational elements. Yet current AI evaluation remains model-centric, offering little insight into how observed behaviours might translate into system-level risk. We propose a framework that links structured hazard analysis, component-level testing, and probabilistic system modelling to bridge this gap. By providing a traceable pathway from model behaviour to system-level outcomes, the framework enables practitioners to answer the "so what?" of AI failures, quantify their systemic impact, and move toward evidence-based and anticipatory governance of AI in complex systems. Applied to the UK's Real Time Gross Settlement (RTGS) system as an illustrative worked example, we derive AI-driven loss scenarios using Systems Theoretic Process Analysis (STPA) and examine adversarial manipulation of LLM-based trading as one such loss scenario. Component-level experiments show that simple adversarial inputs induce measurable behavioural shifts where AI recommendations are followed. Under the component-to-system mapping used here for a financial contagion model, these shifts alter system resilience, increasing bank failures and lowering the threshold at which shocks lead to cascading disruption, particularly under widespread or monopolistic AI adoption.

Introduction

Adopting AI in Complex Sociotechnical Systems

As AI continues to proliferate through many parts of the economy, there is an increasing likelihood of its integration in or adjacent to critical national infrastructure (CNI). Many instances of CNI are and form part of complex sociotechnical systems: they have many interacting components, are adaptive, and generally have non-linear changes of output for linear changes in input (Ladyman, Lambert, and Wiesner 2012). Rather than bounded engineered systems, these systems involve independent actors, interleaved organisational rules, new and legacy technology, and the complex interplay of all other components.

In this setting, regulators, or more generally policy makers, are endowed with a top-down view of the system for which they are responsible; they do not have complete sight or control of the individual components in their system. In contrast, new technologies are typically adopted bottom-up. This asymmetry often renders regulation reactive, both in a general sense and with respect to anticipating harms arising from novel capabilities. These difficulties are particularly pronounced for AI technologies, where rapid rates of technical change, strong incentives for early and widespread adoption, and gaps in institutional expertise combine to limit the availability of empirical evidence or structured assurance at the point of deployment. As a result, policy makers may be required to reason about potential system-level harms under conditions of heightened uncertainty and compressed decision timelines.

To support regulators in safely realising the potential of AI in their domain, we propose a framework for anticipating harms from the adoption of AI in complex sociotechnical systems in an evidence-based, proportionate manner. It enables organisations to address the 'how' of assessing system-level AI harms, complementing existing taxonomies that articulate 'what' AI failure modes may arise (National Institute of Standards and Technology 2023).

The Component to System Gap in AI Testing

Safety engineering teams typically employ multiple tools in a complementary manner to develop robust analyses of potential harms and corresponding mitigations. However, the introduction of new technologies can disrupt well-established practices. For example, early reliability-oriented hazard methods such as Failure Mode and Effects Analysis (FMEA), developed for hardware systems, focused on individual component failures and failed to capture unsafe outcomes arising from interactions between correctly functioning components. This limitation motivated the development of new approaches addressing system-level failures, and contributed to the emergence of system safety as a distinct subfield within safety engineering (Bracke and Pieper 2025)(Leveson and Thomas 2018).

AI introduces distinct challenges, both intrinsic and arising from the field's slow uptake of safety engineering practices (Ahmed et al. 2023)(Rismani et al. 2023a). Generative AI systems are inherently probabilistic; identical inputs may generate different outputs, rendering them unreliable at the component level in many applied contexts. Be-

© Crown copyright 2026. Published by the Association for the Advancement of Artificial Intelligence (www.aaai.org).

cause this behaviour spans an enormous input-output space, it can be difficult to reason systematically about potential resulting harms. Interpretation is limited by opaque training and fine-tuning processes, often split across proprietary and open-source settings. Together, these hinder the anticipation and assessment of tangible and intangible AI-enabled harms (Rismani et al. 2023a) (Rismani et al. 2023b).

In response to these challenges, a large body of AI-specific evaluation and governance tools have been developed. These practices, while valuable, focus either closely on the model in isolation, or on global risks posed by misaligned, general AI (Ahmed et al. 2023) (Bucknall and Dori-Hacohen 2022). In fact, the gap between ethics and system-safety orientated assessment of AI, and alignment-orientated assessments of AI is entrenched in current literature and discourse, leaving AI adopters uncertain on where to start (Roytburg and Miller 2025). CNI stakeholders and policy makers must navigate two parallel ecosystems: established system-safety methods that remain key for reasoning about harms in complex sociotechnical systems, and AI model-specific evaluation techniques. Bridging this gap remains a practical challenge for policy makers seeking to anticipate harms from AI adoption within their system, and is the focus of this work.

Our Approach

We propose the sociotechnical framework shown in Figure 1, targeting policy makers and CNI stakeholders responsible for overseeing AI adoption. It comprises three stages: *structured hazard analysis*; identifying plausible AI-enabled harm pathways, *AI component testing*; measuring the propensity for behaviours of concern, and *simulation-based modelling*; quantifying their system-level consequences. We adopt a broad, technology-agnostic definition of AI throughout: "a machine-based system that, for explicit or implicit objectives, infers from its inputs how to generate outputs such as predictions, recommendations, or decisions that can influence physical or virtual environments" as established by the OECD (OECD 2024). While prior hazard analyses of AI have largely addressed bias or design flaws (Rismani et al. 2023b,a), we focus on security-based failure modes, applied to the UK's RTGS system as a worked example.

Worked Example in Financial Setting

We apply the described framework to a scenario involving the integration of AI into the UK's Real Time Gross Settlement (RTGS) system and the financial network it supports. We leverage the established qualitative hazard analysis tool Systems Theoretic Process Analysis (STPA) to identify possible unsafe behaviours in the present system, agnostic to AI, before considering where adoptions of AI may contribute to real, manifested harms. STPA is used due to its broad capacity to identify harms that stem not just from component failures but also from their use with humans and organisational processes (Leveson and Thomas 2018). Failure orientated hazard analysis methods such as FMEA are not considered here, but we suggest they could be utilised analogously within this framework. We do not treat the internal processes and evolution of AI models. This is a design choice in this framework to focus analysis on harms actionable at the point of AI deployment.

The analyses presented in this work are a worked example in support of a position calling for the structured use of existing system safety tools for reasoning about harms from AI, rather than treating AI harms in isolation. The application to the RTGS illustrates application of the framework and should not be considered a high-fidelity analysis. A high-level and abstracted view of the system is the appropriate level for broader disclosure and supports the position without sharing exploitable system weaknesses.

We identified over 100 unsafe control actions through the application of STPA to the current RTGS system and produced 8 scenarios where specific adoptions of AI could manifest into harm. We demonstrated how these AI adoption scenarios could inform component-level testing, exploring the effect of prompt-injection style attacks on a worked-example of a generative AI solution for trading operations. This testing demonstrated the propensity for this security failure was measurable. Lastly, a stylised model of financial contagion conveyed the system-level harm induced by this component level quality, under an appropriate and specific component-to-system mapping.

Background

The Real Time Gross Settlement System and UK Financial System

The Real Time Gross Settlement (RTGS) system lies at the heart of the UK's financial system, settling over £700 billion on an average working day (Bank of England 2025a). The RTGS is essentially an electronic ledger system, enabling high value transactions between banks, with the Bank of England having direct access to supervise transactions, extend operating times, or provide liquidity. The wider financial system in which RTGS operates is a complex and adaptive "system of systems" (Haldane 2015), and the adoption of AI introduces new sources of complexity within this environment. With 75% of UK financial institutions already deploying AI according to the Financial Conduct Authority (FCA) and Bank of England (BoE) in 2024, AI is increasingly embedded in institutions and processes connected to the RTGS (Bank of England and Financial Conduct Authority 2024). Concerns identified by the Bank of England include liquidity spirals triggered by AI, failure-modes common across AI models and third-party concentration risks (Bank of England 2025b) (Danielsson, Macrae, and Uthemann 2019). These risks, and others identified by parties such as the Financial Stability Board, involve feedback loops around and aggregate interactions with AI, which are challenging to anticipate and cannot be covered by isolated AI model testing (Financial Stability Board 2024).

Systems Theoretic Process Analysis

Effective evaluation of AI risk at the system level must borrow from existing safety-critical domains, rather than try to reinvent the wheel for AI (Vanschoren 2025) (Deckard and Atalla 2025). Looking at safety engineering generally,

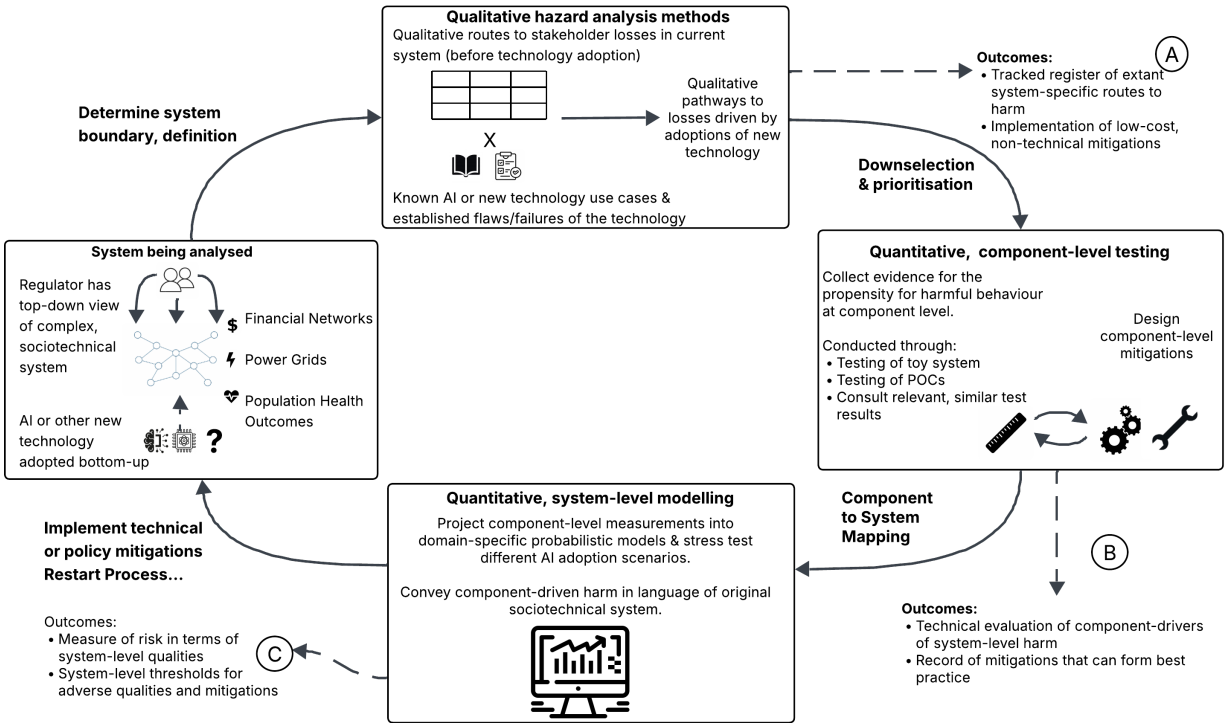

Figure 1: Framework developed in this work, exploring how old and new tools can be used to assess the risk of adoptions of a new technology in complex sociotechnical systems. It is modular and hence can be applied proportionately: practitioners may find it appropriate to finish at any of A, B or C depending on their findings, capabilities and risk tolerance. Finishing at A may be appropriate for paths to harm that do not warrant further technical investigation due to infeasibility, or due to the ease of implementing a mitigation regardless (e.g. staff training). Assessment of a specific harm may end at B if the component behaviour is evaluated to not be of concern, or projecting it into a statistical model is infeasible.

structured hazard analysis is a tool widely used for reasoning about harms in complicated systems. There are several different structured hazard analysis methods with different assumptions and heuristics. Failure Mode and Effects Analysis (FMEA) is a powerful and widely-used approach for analysing system harms that emerge from component failures, and has even been applied to AI systems (Rismani et al. 2023a). Within the framework proposed here, FMEA could be used to identify component failure-based harms for the system adopting AI. However, doing so would miss harms that can emerge even when components work as expected (Leveson 2012).

In complex sociotechnical systems, where harms often arise from interactions, control dynamics, and the use of otherwise functioning components, approaches are needed that can capture subtle and emergent pathways to harm. Systems-Theoretic Process Analysis (STPA) addresses this by explicitly treating safety as an emergent property, shifting the focus towards failures of control and interaction within the system. This perspective is well-aligned with the types of risks introduced by AI systems, where failures are often contextual, interaction-driven, and dependent on how outputs are incorporated into broader workflows. In this work, we therefore employ STPA as the structured hazard analysis technique, while noting that methods such as FMEA could be incorporated within the same framework to provide valuable, but more component failure-orientated perspectives.

Methodology

The following sections detail efforts made in this illustrative worked example to complete each of the three analytical components shown in figure 1. STPA was applied to develop a functional model of control of the RTGS and its immediate network of banks, inserting the adopted AI use case into the Loss Scenario (LS) stage. AI-driven LSs were then down-selected according to their likelihood for tangible harm and their feasibility of testing. A single LS, centred on the manipulation of AI-augmented trading activities via indirect prompt-injection, was then explored both at the component and system-level. This was tested in an illustrative, but faithful setup of the described scenario using AI foundation models, before being projected into system-level harms (bank failures) via a financial contagion model, grounded in the literature of financial stability modelling.

STPA

The first part of the proposed framework is the identification of pathways to harm. The intended outcome is a descriptive scenario that reasons about how AI can amplify existing or produce new failures modes within the system, and

why this may occur. We employ STPA in this work. Readers are pointed to (Leveson and Thomas 2018) for a detailed understanding of STPA's logic, assumptions and benefits. STPA proceeds in five steps: (1) identify stakeholders, values, and system boundaries; (2) identify losses, hazardous system states, and system constraints; (3) model the control structure using control actions and feedback between controllers and controlled processes; (4) identify unsafe control actions (UCAs) — those not applied, applied when unnecessary, applied incorrectly, or applied at the wrong time; and (5) develop loss scenarios by tracing causal factors in failed or flawed control for selected UCAs.

Rather than applying the full STPA process separately to each possible deployed AI system, we first apply steps one to four, as detailed above, to the system as-is, before AI integration. In a conventional application, STPA would typically analyse a specific system implementation. Here, we instead map candidate AI use-case classes onto the existing control structure to identify where they could affect control actions, feedback, or decision-making. These candidate use cases are grounded in industry surveys and practitioner insights (Bank of England 2025a) (Hall 2024). The analysis is therefore framed around how AI adoption changes the system, and operates at the level of AI use-cases rather than particular models or technical deployments.

The system boundary was defined to include the principal components of the RTGS ecosystem: the Bank of England, the RTGS infrastructure, CHAPS direct participants, SWIFT as the messaging layer, and relevant contingency processes. Stakeholders were identified and their values considered in defining system-losses. Hazards were then identified as system states which, under adverse conditions, could lead to losses. System constraints specified the conditions required to avoid these hazards. System constraints were mapped onto components to identify control actions, enabling the Hierarchical Control Structure (HCS) to be developed. Then each control action was considered for how it could become unsafe, according to the standard STPA process (Leveson and Thomas 2018), producing up to four possible fundamental UCAs for each control action. Actual trends of AI adoption within the financial sector were then consulted, before several use cases were considered and overlaid with our constructed HCS (Bank of England and Financial Conduct Authority 2024) (Hall 2024). In addition to considering AI use cases, we also used domain knowledge of AI failures, unintended behaviours or general flaws (National Institute of Standards and Technology 2023). This enabled the development of AI-driven loss scenarios, tracing how flaws of AI implemented at or adjacent to key junctures within the system could contribute to unsafe control actions and propagate through the system to produce losses. The final AI-driven loss scenarios are the primary output of this step, which are numbered to reflect their explicit construction from earlier UCAs. A table of UCAs and further detail on loss scenarios is provided in Appendix A.4.[1] At the end of this stage (stage A in Figure 1) we have a qualitative taxonomy of AI-enabled risk pathways sufficient for early prioritisation, but still open as to whether those risks materialise in practice, motivating component-level testing.

[1]Appendix references correspond to the extended version of this work, containing additional technical supporting material, available at https://arxiv.org/abs/2608.23906.

Component Level Testing

Selecting What To Test The selected AI-driven LS, LS-18.2.2-A, supposed a setting where trading operations within banks and large financial institutions are highly delegated to GenAI based functions. While such delegation represents an upper bound case, it is both technically feasible with current GenAI capabilities and organisationally plausible given well-documented patterns of cognitive offloading and automation bias in complex decision environments (Bainbridge 1983) (Parasuraman and Manzey 2010). GenAI solutions are able to use tools, aggregate heterogeneous information at scale, and initiate consequential real-world actions, making meaningful human-in-the-loop oversight difficult to sustain at operational speed (Weidinger et al. 2023). As a result, nominal human supervision may not always constitute effective constraint on system behaviour, particularly in times of stress, which is the scenario under investigation here.

This scenario is further supported by the BoE and FCA's joint 2024 AI survey, reporting that GenAI based solutions are increasingly being adopted across financial services, with investment operations occupying the fifth highest adoption of this technology (Bank of England and Financial Conduct Authority 2024). LS-18.2.2-A was selected not as a forecast of the typical deployment, but as a credible configuration under which systemic AI risks could be stress tested via our functional model of the system, component testing and later simulation-based modelling. This scenario aligned both with academic work highlighting macro-financial risks arising from automated responses to stress (Danielsson, Macrae, and Uthemann 2019), but also risks identified as being of high material concern by the BoE survey correspondents: reliance on third party providers; common data and model dependencies; correlated behaviour across institutions (via AI decision making). A more detailed description and justification of our choice of AI driven loss scenario is provided in Appendix A.6.

Baseline Testing The first step in our framework allows the user to identify components that could influence system-level risks; the second step is to probe how those components behave under different conditions. Here we build a lightweight LLM-based decision component that mimics an AI-assisted investment function. The tool ingests a small, controlled set of market-news articles, and recommends how a portfolio should be allocated across four broad asset types: Equities, Mortgage-Backed Securities, Corporate Bonds, and Government Bonds. This level of granularity balanced faithfulness to how portfolio managers deal with high-level exposures to different asset types, whilst avoiding the intricacy of representing specific sectors or stocks. Articles were generated with foundation models to constrain the sentiments and asset focus of the articles. Sentiments were constrained to: *Bullish* (positive sentiment), *Bearish* (nega-

tive sentiment), and *Neutral*.

In our experimental pipeline, an *Episode* denoted one invocation of the tool. In each Episode, the tool ingested four articles (one of each asset type) and produced a portfolio-style judgment for each asset type represented by a number between 1 and 0, which summed to 1 across all asset types. While stylised, this setup provides sufficient cross-asset coverage while maintaining experimental control. An *ExperimentalRun* represented a structured sequence of Episodes designed to test one hypothesis. All generated experimental runs track the exact episodes that were processed, which articles were present and their attributes. LLM responses were first collected for the *AllNeutralRun* experiment, where all articles have Neutral sentiment. This was chosen to gather a valid baseline of responses across the models. Next, the setting where a specific asset was Bearish and all others Neutral was explored, *FixedBearishNeutralRun*, revealing how models responded to asymmetric sentiments, without the presence of any manipulation.

Adversarial Testing This work focused on a broad class of scalable attacks: low-cost, accessible manipulations that require minimal specialised knowledge of AI security, inspired by simple attacks explored in the context of internet search engine optimisation (Nestaas, Debenedetti, and Tramèr 2024). The utilised attacks are presented in Appendix B.2, but in simple terms focus on manipulating the preference of the LLM, rather than breaking safety guardrails. We did not assess ease of embedding manipulations into the input text. Rather, we treat indirect prompt injection as a well-established and low-cost threat, accounted for at the scenario mapping stage.

The experimental pipeline enabled specific assets to be targeted, with the attack placed just following the text that mentions the targeted asset. The attacks could be targeted at any arbitrary asset class within the episode, but for this analysis they always targeted the asset which was already in distress, i.e. associated with Bearish sentiment. This sought to focus on scenarios where adversaries may seek to amplify fireselling of distressed assets, rather than explore how adversarial attacks could boost the allocation to one specific asset. Though the latter case is of interest, and more closely analagous to the cited existing research, it is of less concern from a view of system-level risk as considered here. Adversarial testing saw each *FixedBearishNeutralRun* experiment have the Bearish asset be targetted by one of these simple adversarial attacks. Stage B (Figure 1) provides empirical evidence of whether the failure mode manifests and whether mitigations are effective, or whether residual risk remains, informing whether system-level modelling is warranted.

Quantitative Complex Systems Modelling

Choice of Simulation-Based Model The final technical stage in this framework involves projecting measurements of a component-level behaviour into simulation-based models of the system. We recall that STPA posits hazard states as being those which, combined with some external, uncontrolled environmental state, necessarily lead to losses. This supports the use of simulation-based models to explore the aggregate effect of processes in and beyond the system combining with a failure or flaw at the component level (measured in the prior step).

In practice, policy makers may apply this step using higher-fidelity or proprietary models (including infrastructure-level systems such as RTGS). For illustrative purposes, here we used a financial contagion model to explore the system effects of the adversarial manipulation previously measured. We output two quantities that map to Losses L3 - Loss of wider functioning of banking sector and implicitly, L2 - Loss of high value payment continuity, namely failed banks, and banking losses. Rather than model the RTGS engine itself, we selected to model the financial network which RTGS serves, which was considered the appropriate level of abstraction to link back with the losses defined at the beginning of the framework. Further, financial contagion models provide a well-established and interpretable class of simulation-based models for analysing system-wide losses, making them suitable for instantiating this step of the framework (Upper 2011) (Gai and Kapadia 2010).

Developing the Financial Contagion Model We developed a stylised financial contagion model, consisting of a network of 50 banks with a core–periphery structure, following empirical financial network literature. Core banks are highly interconnected and systemically important, while periphery banks are smaller and less connected (Craig and von Peter 2014). The model used here extends the Eisenberg–Noe framework. In this framework, banks maintain balance sheets of external assets, interbank exposures, liabilities, and equity, with failure occurring when liabilities exceed assets (Eisenberg and Noe 2001).

Several targeted modifications were made to the modelling framework to capture additional channels for local behaviours to propagate into system-level outcomes, providing a clearer interface for measurements at the AI-component level. These changes introduced heterogeneity in exposures, additional loss-propagation mechanisms, and proper constraints on loss-allocations, e.g. priority of claims. The change to banking exposures involved defining external assets as diversified portfolios across multiple asset classes (government bonds, corporate bonds, mortgage backed securities, and equities), reflecting Basel-style risk differentiation and aligning with AI component measurements made prior. Indirect contagion is introduced through a fire sale mechanism, appending to the direct contagion that occurs via interbank liabilities. Fire sales represent bank failures triggering asset liquidations that depress market prices, causing mark-to-market losses for other institutions. This creates multi-round feedback loops between failures and price impacts. Fire sales are parameterised so that price impacts scale with the proportion of failed institutions in each round, capturing amplification effects arising from synchronised liquidation behaviour, and an independent fire sale parameter, I_{fs}.

We validated that the model continued to predict the qualitative patterns occurring in historical events, namely the 2008 financial crisis. Two regulatory regimes were instanti-

ated as distinct parameter configurations: a pre-2008 system with lower capital, higher interconnectedness, and greater exposure concentrations, and a post-2008 system calibrated to Basel III and Dodd-Frank conditions using empirical bank data (United States Congress 2010) (Basel Committee on Banking Supervision 2010). The model uses a parameter sweep approach. Exogenous shocks to asset classes are varied, with Monte Carlo simulations over stochastic network realisations. This links back to STPA's hazard states, where losses are produced from a flawed system-state coinciding with an undesirable external, environmental condition, such as a mortgage shock. On the other hand, parameters of the model governing asset liquidations are the interface through which AI-driven trading behaviour, projects into system behaviour.

Mapping Into the Financial Contagion Model AI-enabled trading behaviour was introduced via the fire sale intensity, I_{fs}, governing how aggressively assets are liquidated after failures. Component measurements, for model l, were translated into system-level effects by using the shift in average asset allocation, μ, from Neutral sentiment, to the case where one asset was distressed, as an indication for an AI system's inclination to sell off those assets under stress, as might happen under system-level shock. This is expressed as f_l, expressed in equation 1 below:

$$f_l = \mu_l^{(0)} - \mu_l^{(-)}, \tag{1}$$

where (0) and $(-)$ respectively represent Neutral and Bearish sentiments. Concretely mapping AI component-level qualities to system-level parameters requires consideration of AI adoption pathways. To this end, measured allocation shifts from different foundation models were averaged in a weighted fashion across different characteristic adoption paths, e.g., 100% GPT-5-mini monopoly vs diverse, mild AI adoption, denoted ρ, and using the baseline fire sale term from previous simulations, f_b to represent no AI adoption. This is expressed in equation 2.

$$I_{fs} = \kappa(\rho_b f_b + \sum_{l \in M}^{M} \rho_l f_l), \quad \rho_b + \sum_{l \in M}^{M} \rho_l = 1. \tag{2}$$

The term κ is a modelling constant used to scale component-measured shifts to the appropriate range for the systems-level model, here the value was set to 1. Adversarial manipulation was applied to already distressed assets. This sought to demonstrate how simple but widespread attacks on LLM systems, responsible for material trade execution decisions, could worsen otherwise manageable financial shocks.

This also highlights a broader role for complex systems models when combined with structured hazard analysis. Qualitative methods necessarily abstract some system detail to remain interpretable and tractable. Computational models can complement this by explicitly representing many interacting institutions and their dynamic relationships, making system-level risks such as third-party concentration more observable. This also highlights the importance of constructing a sensible mapping from component to system level, requiring thought to how to reflect the scaling of the component behaviour, and whether any human behaviour is being replaced and whether appropriate measures exist for this. Appendix H.1 works through the derivation of this mapping in more detail, and Appendix H.2 demonstrates the robustness of results across many mapping parameter choices.

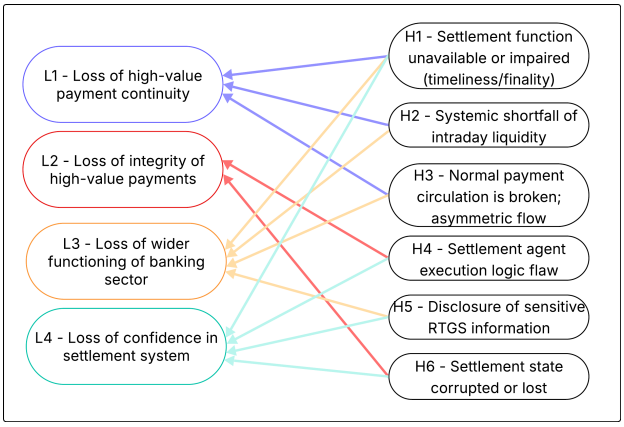

Figure 2: Tracing of high-level hazards into losses for the Real-Time Gross Settlement system (RTGS), surfaced via systems-theoretic process analysis. Hazards are system states that, when combined with unfavourable environmental conditions, produce losses. Losses relate to unacceptable outcomes to be avoided for the stakeholders of the system.

Experiments We performed parameter sweeps over exogenous shocks (e.g. mortgage shocks across a defined range, such as -1 to -30 percent), with proportional shocks applied to other asset classes. Monte Carlo simulations were run for each shock level, measuring system-level outcomes such as bank failures and aggregate losses. We evaluated how bank failure rate thresholds shift under different I_{fs}, including baseline fire sale intensity (assumed no AI), benign AI-driven behaviour, and adversarially-amplified behaviour. All simulations were implemented in Python with configuration-driven parameterisation. At the end of this stage (stage C in Figure 1) we have a quantitative estimate of system-level losses under the identified failure mode, enabling evidence-based decisions on mitigation. This cycle can be repeated for further scenarios identified at stage A.

Results

STPA

STPA identified several plausible pathways to harm from the adoption of AI in and around the RTGS system, including over 100 AI-agnostic unsafe control actions and 8 AI-driven loss scenarios. Four high-level losses were defined to represent unacceptable stakeholder outcomes, within the defined system-boundary. These losses were linked to six high-level hazard states, presented in figure 2.

These hazard states were further decomposed into sub-hazards and corresponding System Constraints (SCs), defining how such hazardous states are avoided. These are detailed in Appendix A.2. With an understanding of the hierar-

chy within the system boundary, the SCs then informed control actions that entities engage to avoid hazardous system states, enabling an HCS to be developed. The HCS for the full RTGS system, and sub-HCS for a single RTGS participant bank, are shown in Appendix A.3. All identified control actions were analysed for potential UCAs, resulting in 104 AI-agnostic UCAs (Appendix A.4). To our knowledge, applications of STPA to financial infrastructure, and to a lesser extent AI, remain limited in the public domain beyond this work. Eight AI-driven LSs were then developed and are provided in full in Appendix A.5. One representative scenario (LS-18.2.2-A) was selected for component-level testing, focusing on adversarial manipulation of LLM-based portfolio advice and its potential to amplify bank failures during stress, presented in Appendix A.8. Though qualitative, the development of a narrative on the causal factors involved in producing UCAs provided a structured and thorough means to highlight compounding component factors that lead to the harm. This produced an explicit and auditable record of causal pathways, supporting identification of potential mitigations.

Component Testing

Component-level tests were conducted for each of three foundation models, exploring the extent to which simple adversarial attacks could measurably shift asset allocation under this stylised trading setting. The average asset allocation under different regimes, including the adversarial setting, are shown in figure 3, where the adversarial shift is the shift averaged across the explored attacks (i.e. an average over multiple similar but different attacks).

Non-Adversarial Settings Before describing the measured asset allocation shift under adversarial attack, we note that these models each demonstrated their own idiosyncrasies in response to the different sentiments, outside of adversarial conditions. We outline these briefly noting that a single component-level experiment can identify several concerns for policy makers that may sit with several (not explicitly explored) STPA-derived loss scenarios, or even suggest new loss scenarios to be developed.

Firstly, each queried model demonstrated an aversion towards mortgages backed securities (MBS) in Neutral settings. No model, on any occasion in this setting, gave the greatest weighting to MBS. Secondly, each model displayed clear preferences in Neutral settings which,crucially, varied across models. For example, Claude 3.5 Haiku and GPT-5-mini leant towards corporate bonds; GPT-5-mini allocated less towards MBS. Thirdly, GPT-5-mini produced the largest swing in asset allocations between the Neutral to Bearish sentiments, in this context indicating it had the strongest tendency to sell off a given asset compared to the other models. While not explored here in detail, these non-adversarial observations could conceivably be mapped into the system-level model, e.g. to understand simple model biases and correlated responses to market events in foundation models.

Adversarial Settings Here, articles with Bearish sentiment had adversarial text included, according to which asset was targeted. Figure 3 demonstrates that simple adversarial attacks measurably shifted the allocation. We note the shifts presented here are small on an absolute scale, but occur against the hard boundary of 0% allocation to a given asset, and constitute large relative shifts. In the context of executing trades, these single-shot, unsophisticated attacks made the asset appear tangibly more distressed than the whole body of text would indicate. Where an LLM drastically responded to negative sentiment, these attacks succeed in making that response more acute. However, while the average shift over attacks is shown clearly in figure 3, the efficacy of individual attack instances varied.

The STPA loss scenario identified several contributing causal factors, pointing to mitigations including secure system design, input filtering, and human review. In the main analysis, we focus on the unmitigated setting in order to examine how component-level AI failures could project into system-level risk under a worst-case scenario. We also tested a prompt-hardened setting, reported in Appendix B2.2, both to illustrate how framework outputs can inform mitigation selection and to avoid presenting simple adversarial attacks without corresponding defences. Prompt-hardening was highly effective at this stylised scale. Of course, we note that this does not establish robustness against adaptive or more sophisticated attacks, and should instead be read as one example mitigation within a broader safety architecture.

Quantitative Complex Systems Modelling

The financial contagion model described in section was used to generate bank failure rates in a network of 50 banks, in a core and periphery structure, under a variety of asset shocks. Having measured the benign and adversarial asset shifts across assets, these were projected into firesale intensities according to different adoption scenarios (between the three models explored here), via the mapping in equation 2. Mapping parameters of $K = 1.0$ and $f_H = 0.05$ are used for the plot in figure 4, which demonstrates how increases of firesale intensity result in a larger bank failure rate for a given mortgage shock.

As measured at the component level, GPT-5-mini had the greatest tendency to shift asset allocations aggressively under distress, and to further liquidate under adversarial attack. Figure 4 shows that increasing the relative proportion of this model, through increasing the effective I_{fs} of the system, increases the max bank failure rate, with a GPT-5-mini monopoly system in adversarial settings presenting the greatest bank failure rates. We do note, however, that the effects of increasing the firesale intensity appears to be self-limiting, with plateauing behaviour of the bank failure rate. We also note that as well as bank failure rate increasing as we tend towards full AI adoption, with full AI autonomy assumed here, and monopolistic adoption of specific models, adversarial settings always demonstrate greater failure rates. This effect is not limited to a single choice of mapping parameter, as shown in Appendix H.2, where adversarial attacks systematically increase the bank failure rate across mapping parameter choices. In general, this is to be expected given the attacks were routinely effective in shifting asset allocations, however figure 4, demonstrates why this susceptibility being widespread could have significant

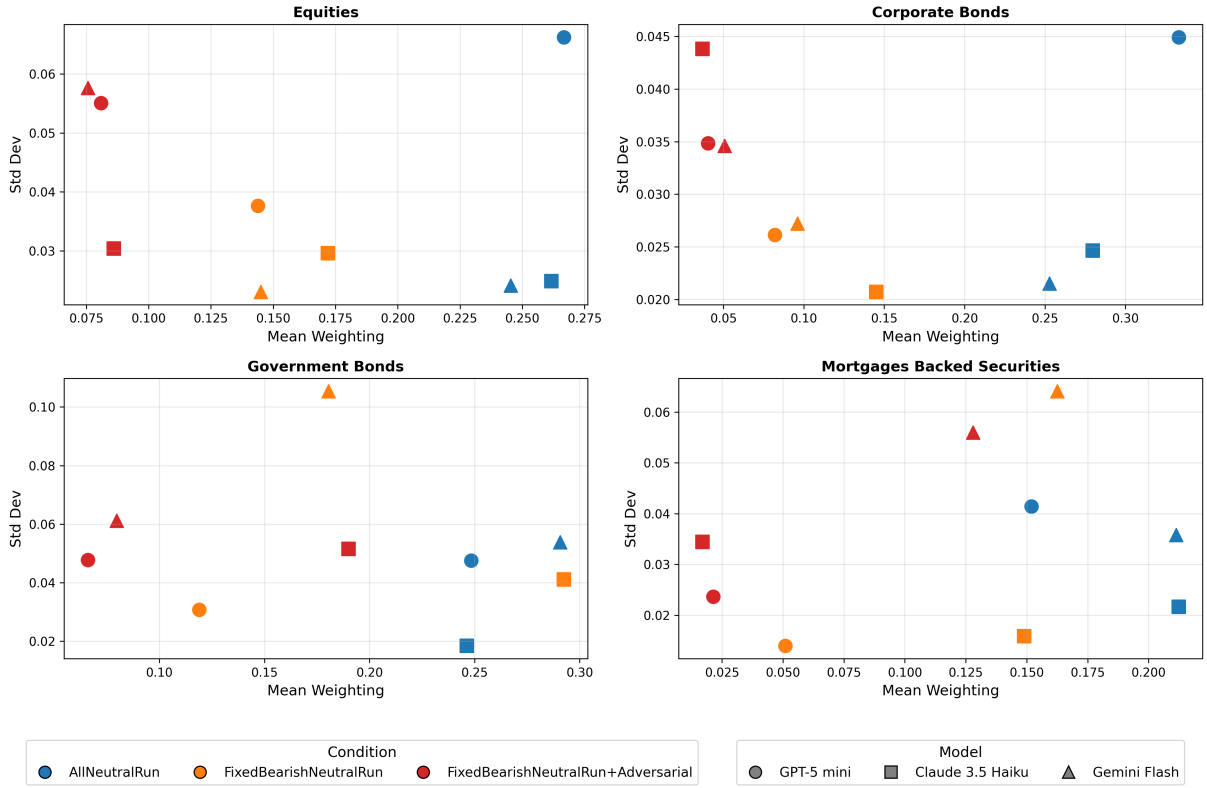

Figure 3: Scatter plot across the asset types, demonstrating shift in asset allocation from Neutral sentiment across assets (blue) to Bearish sentiment for the shown asset (orange) and adversarial manipulation added to a distressed asset (red). The standard deviation of the asset allocation is shown on the y-axis, and the mean asset allocation on the x-axis. The asset value for the adversarial setting is averaged across all the attacks explored in this work. There were 150, 100 and 170 measurements for the neutral, bearish and adversarial runs respectively, for each model. Apart from Claude 3.5 Haiku and Gemini flash in the adversarial setting, with 50 and 168 measurements, due to model retirement during testing and transient failures, respectively.

system-level consequences, if the attack was realised.

Lastly, we note that the overall increase in bank failure rate with increase fire sale intensity lowers the threshold of mortgage shock required to yield a certain bank failure rate. This represents a less resilient system in which smaller shocks can produce the same system-level outcomes as larger shocks required in otherwise more resilient systems (see Appendix H.3 for more detail). This provides a quantitative bridge back to our qualitative analysis, enabling practitioners to quantify adverse environmental conditions (outside the system designer's explicit control), as specific mortgage shocks, otherwise labelled just as 'worst case', and to measure the ensuing loss, 'L3 – Loss of wider functioning of banking sector' occurs. Further, for illustrative purposes, in Appendix H.3 we present how a policy maker may overlay red, amber and green operating zones over these threshold plots, using the fire sale intensity value as an indicator for risk of adverse outcomes at the system level.

Discussion

Helping To Answer the 'So What?' of AI Failure

A limitation of AI system evaluation is that model-level failures do not, in isolation, explain why those failures matter. A shift in an AI system's output may be measurable, but its significance depends on the system in which that output is used and the losses to which it may contribute. Our work illustrates how this gap can be addressed by structuring evaluation around system-level harms from the outset: hazard analysis defines relevant losses and causal pathways, component-level experiments test behaviours implicated in those pathways, and quantitative modelling translates those behaviours into system-level outcomes.

In the illustrative trading scenario, adversarial prompts produced measurable shifts in asset allocation. Interpreted in isolation, this finding says little about harm. Within this framework, however, the same behavioural shift is qualitatively linked to a system-level loss defined by hazard analysis, and quantitatively mapped onto a financial contagion model, allowing system-level consequences to be quantified. Under the assumptions of this mapping, adversarially shifted allocations increased fire-sale intensity, raised bank failure rates under stress, and lowered the external shock threshold required to trigger widespread failure, especially under widespread or monopolistic adoption. The broader implication is that the framework provides a traceable account of when and how AI failure modes matter at the system level.

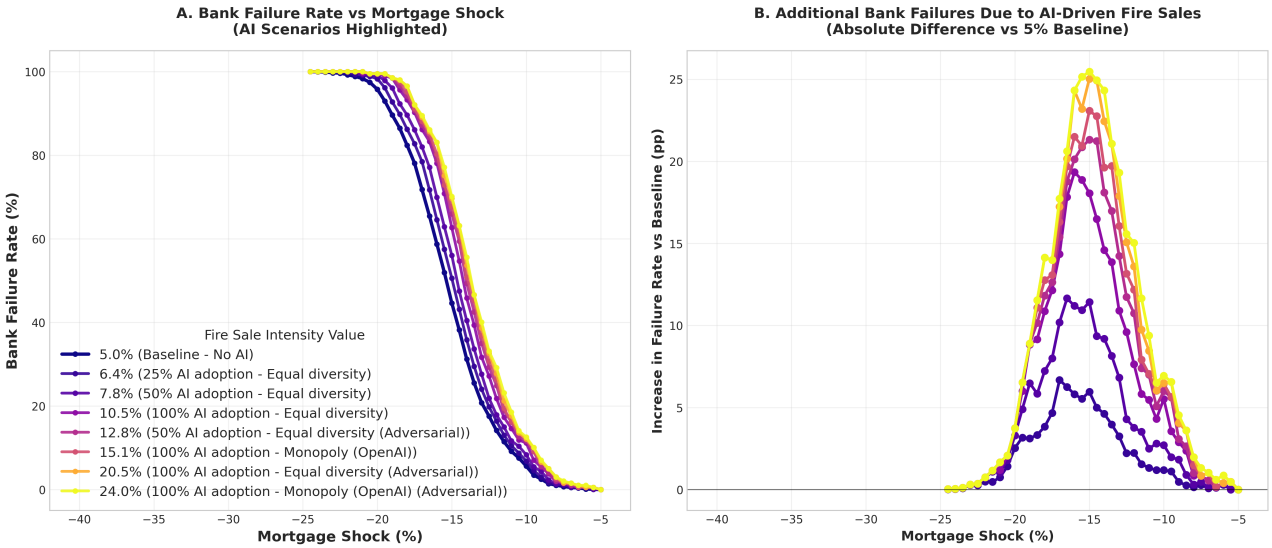

Figure 4: Plot of bank failure rate, absolute (left), and relative to baseline firesales (right). We observe that increased firesale intensity leads to heightened overall bank failure rates, and that adversarial scenarios yield the largest bank failure rates, as well as monopolistic adoption scenarios. This plot uses mapping $K = 1.0$ and $f_h = 0.05$.

Supporting AI Policy To Be Anticipatory, Prioritised, and Accountable

First, the framework supports anticipatory policymaking by starting from system-level losses rather than model failures alone. This gives policymakers a structured way to ask which AI behaviours would matter in a given sector before those harmful behaviours have occurred. Qualitative hazard analysis provides an initial pathway from losses to behaviours of concern, while complex systems modelling assesses whether that pathway is likely to be material under particular assumptions. The framework is complementary to existing practices such as regulatory sandboxes, industry surveys, and stakeholder engagement: those practices help ground assumptions about how AI is being adopted, while this framework helps translate those assumptions into testable harm pathways.

Second, the framework supports prioritisation at the point of AI adoption, focusing on processes which adopters are able to remedy. Prior sociotechnical analyses with STPA considered harms across different stages of the ML lifecycle (Rismani et al. 2023a) (Mylius 2025). While valuable general analysis of AI harms, this is less actionable for sector-specific policy makers, who are better positioned to govern how AI is deployed within their system. For example, a regulator could determine from this worked example that firms using foundation models for trading, must take reasonable steps to implement defences for context-specific prompt injections. By focusing on AI at the point of adoption, and in terms of AI use cases, the framework supports testing and mitigations centred on input–output behaviour that most directly drive harm to stakeholders.

Third, the framework supports monitoring and post-incident accountability by linking system-level losses, hazardous scenarios, AI behaviours, and possible system outcomes. One policy implication is targeted reporting: regulators could ask deployers to measure and report behaviours identified through the hazard analysis, enabling these measurements to be aggregated within system models. This would support monitoring of when the system approaches unsafe or higher-risk operating conditions, as illustrated in Appendix H.3. Where direct measurement is infeasible, the same analysis can still guide higher-level information gathering, such as surveys on model usage, adversarial exposure, or reliance on particular providers. This could even extend existing practices such as the Bank of England's system-wide exploratory scenario exercises (Bank of England 2024) by incorporating reported component-level behaviours into simulations of AI-driven risk. Finally, because structured hazard analysis records links between losses, hazards, system constraints, and control actions, it also supports post-incident analysis by making causal assumptions inspectable after failures occur, including through complementary methods such as Causal Analysis based on Systems Theory (Walsh, Schulker, and Davison 2025).

Proportionality and Institutional Fit

Previous work on structured hazard analysis for AI has often positioned such approaches within the AI development lifecycle, targeting the development team or broader company as the relevant parties to undertake hazard analysis (Rismani et al. 2023a) (Rismani et al. 2023b). However, many of these works comment that in practice the adoption of these methods is limited by commercial pressures and organisational incentives that do not prioritise system safety (Moss et al. 2020). To address this issue, this work has targeted policymakers overseeing AI adoption in their system. Policymakers and operators of complex systems are: responsible for avoiding adverse system outcomes; often possess tools, cul-

ture, and tacit knowledge for safety engineering; tasked with overseeing systems in which AI is, sometimes discreetly, being integrated. These organisations are therefore both better positioned to apply structured risk analysis based methods. Unlike AI developers, they do not require a shift in organisational culture toward safety, but can instead extend existing practices. The framework's modularity also supports this institutional fit, allowing it to be used without requiring full end-to-end implementation in every case. Individual stages can be used independently: qualitative hazard analysis can support exploratory assessment of how AI adoption may create or worsen pathways to harm, while organisations with greater technical capacity can extend this through component-level testing or system modelling.

Limitations and Future Work

Firstly, the initial stages of this framework are built on qualitative, structured hazard analysis, and so inherit the limitations of such methods. Practitioners make assumptions and normative judgments about the system under analysis, which could create divergent outcomes between analysts (Rismani, Dobbe, and Moon 2024). Once loss scenarios are defined, the framework requires decisions about what quantities to evaluate and how the evaluation is constructed, which in practice requires value judgements (Bowker and Star 2000), and may emphasise some harms as more relevant or important (Luccioni et al. 2023). This can be mitigated by consulting a diversity of expertise in constructing and testing relevant AI-driven loss scenarios. In fact, a benefit of doing structured analysis is that such biases and normative judgements are made explicit. Future work should explore how to combine structured hazard analysis methods with established AI harm taxonomies such as the NIST AI Risk Management Framework (National Institute of Standards and Technology 2023) and MITRE ATLAS (MITRE 2026), to support more consistent mapping of AI-related loss scenarios to testable behaviours.

Secondly, mapping component-level qualities to system-level impacts necessarily involves analytical judgement, especially in complex sociotechnical systems where harms may only emerge at scale or via interactions. Constructing these mappings requires simplifying assumptions that balance interpretability and fidelity, and must consider how components are adopted in practice. This process is especially challenging in AI, where human-machine interactions are not yet well understood and harms may only appear at scale (Doshi and Hauser 2024) (Weidinger et al. 2023). While challenging, however, the analytical process provides a structure that enables surfacing, reasoning about and prioritising the potential risks and mitigations.

Thirdly, we have demonstrated the framework conceptually, using publicly available data to prevent leakage of sensitive information. While we have had our approach validated by safety and security experts, the next stage of development involves direct engagement with stakeholders, safety practitioners and policy makers in the finance domain. Working directly with financial institutions, for example through workshops, or pilot applications, will validate and tune assumptions to allow modelling to better reflect operational reality, surfacing practical considerations in applying the framework. Future work should also test whether our approach can be applied successfully to other CNI sectors adopting AI.

Conclusions

This work addresses a central challenge for policy makers and direct regulators responsible for complex sociotechnical systems: how to anticipate system-level harms from bottom-up AI adoption despite limited visibility and control over individual components. We show that the gap between AI component evaluation and system-level risk can be bridged by integrating structured hazard analysis, component-level testing, and probabilistic system modelling into a single framework.

By combining Systems-Theoretic Process Analysis with empirical testing and simulation-based modelling, the framework provides a traceable pathway from hypothesised harms to quantified system outcomes, enabling practitioners to answer the "so what?" of AI failures in concrete, contextual terms. In the worked example grounded in the Real Time Gross Settlement system and its financial network, small adversarially induced shifts in asset allocation were propagated through a stylised financial network model, illustrating how the framework can be used to explore how local AI behaviours may translate into system-level effects under different adoption scenarios. This demonstrates how the approach can be used to investigate conditions under which component-level behaviours could contribute to larger system-level consequences, particularly under widespread or correlated AI adoption.

The framework supports a shift from reactive to anticipatory governance, allowing policy makers to explore plausible harm pathways and assess their significance prior to widespread deployment. It supports prioritised and actionable oversight at the point of AI adoption, focusing attention on behaviours that contribute to stakeholder specific harm and fall within regulatory influence. Furthermore, the staged and modular structure affords an evidence-gated approach to risk assessment, enabling practitioners to allocate effort only in proportion to their capabilities and the existing evidence generated towards the suspected system-level harm. By focusing on system operators as primary users, the framework aligns with existing institutional incentives, safety practices, and expertise. This offers a practical pathway for fostering practical analysis of system-level AI-driven harms, as opposed to often stifled initiatives aimed at the AI developers.

While the framework depends on qualitative judgements and interdisciplinary expertise, between hazard analysis and component testing, and in mapping component tests to system outcomes, respectively, these are made explicit and auditable. Ultimately, this highlights inherent challenges rather than creates them and provides a structured basis for reasoning about systemic AI risk and a foundation for further refinement through interdisciplinary collaboration.

Ethical Statement

This work demonstrates adversarial manipulation of LLM-based systems in a stylised financial setting. The attacks are drawn from published literature and the primary contribution is to surface their potential systemic consequences, not to advance attack capability. We nevertheless acknowledge that publication of adversarial techniques, even where previously documented, may provide a starting point for actors seeking to exploit similar vulnerabilities in deployed systems. The RTGS analysis is conducted at a level of abstraction chosen to avoid disclosing operationally sensitive information, and was reviewed by domain and safety experts.

The framework involves normative judgements about system boundaries, stakeholder values, and the relative importance of harms, and its outputs depend on the quality and completeness of the evidence, expertise, and models used. The worked example was conducted entirely from publicly available information, without direct engagement with operators. Our account of the system is therefore necessarily partial, and our judgements about what constitutes a plausible or significant AI-driven loss scenario are filtered through what is publicly disclosed. Partial analyses can nevertheless be useful when their assumptions and evidential limits are made explicit, but should not be interpreted as comprehensive safety assessments or as evidence that unmodelled risks are absent. Making these assumptions, judgements, and evidential limitations explicit and auditable, as STPA requires, is intended to support scrutiny and revision rather than provide assurance on its own.

The authors have backgrounds in AI safety and security. This shapes our focus on security-based failure modes and measurable behavioural properties of AI systems, and may lead us to prioritise threat pathways that are more tractable from a technical standpoint over representational or organisational harms. The choice of a UK financial infrastructure case study similarly reflects the context and information available to the authors rather than a claim that this context is more important or representative than others. Practitioners applying the framework in different regulatory, geographic, or sectoral contexts should therefore expect to revisit the normative assumptions and analytical priorities embedded in each stage.

Acknowledgements

This work was supported by the Laboratory for AI Security Research (LASR); the views expressed in this paper are those of the authors and do not necessarily reflect the position of LASR or His Majesty's Government.

We also acknowledge and thank colleagues whose continued support and feedback were invaluable to the development of this paper, particularly Lee C and other NCSC colleagues for their technical reviews of the work, and Molly Parnham and Michaela Coetsee at Advai for research delivery support and feedback, respectively.

Generative AI tools were used throughout this project to support literature familiarisation, iterative development of the STPA analysis, code prototyping, and manuscript drafting and refinement

References

Ahmed, S.; Jaźwińska, K.; Ahlawat, A.; Winecoff, A.; and Wang, M. 2023. Building the epistemic community of AI Safety. *SSRN Electronic Journal*.

Bainbridge, L. 1983. Ironies of Automation. *Automatica*, 19(6): 775–779.

Bank of England. 2024. BoE System Wide Exploratory Scenario. Bank of England.

Bank of England. 2025a. A Brief Introduction to the Real Time Gross Settlement System and CHAPS. Bank of England.

Bank of England. 2025b. Financial Stability in Focus: Artificial Intelligence in the Financial System 2025. Bank of England.

Bank of England; and Financial Conduct Authority. 2024. Artificial intelligence in UK financial services 2024. Bank of England / Financial Conduct Authority.

Basel Committee on Banking Supervision. 2010. Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems. Technical report, Bank for International Settlements. Revised June 2011.

Bowker, G. C.; and Star, S. L. 2000. *Sorting things out: Classification and its consequences*. MIT Press.

Bracke, S.; and Pieper, R. 2025. *Safety Engineering: Fundamentals, methods, research topics*. Springer.

Bucknall, B. S.; and Dori-Hacohen, S. 2022. Current and near-term AI as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 119–129.

Craig, B.; and von Peter, G. 2014. Interbank tiering and money center banks. *Journal of Financial Intermediation*, 23(3): 322–347.

Danielsson, J.; Macrae, R.; and Uthemann, A. 2019. Artificial Intelligence and Systemic Risk. SSRN Electronic Journal.

Deckard, A. C.; and Atalla, C. 2025. Learning from other domains to advance AI evaluation and testing. Microsoft Research Blog.

Doshi, A. R.; and Hauser, O. P. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28).

Eisenberg, L.; and Noe, T. H. 2001. Systemic risk in financial systems. *Management Science*, 47(2): 236–249.

Financial Stability Board. 2024. The Financial Stability Implications of Artificial Intelligence. Financial Stability Board.

Gai, P.; and Kapadia, S. 2010. Contagion in Financial Networks. *Proceedings of the Royal Society A*, 466(2120): 2401–2423.

Haldane, A. 2015. On Microscopes and Telescopes. Speech, Bank of England.

Hall, J. 2024. Monsters in the Deep? Speech, Bank of England.

Ladyman, J.; Lambert, J.; and Wiesner, K. 2012. What is a complex system? *European Journal for Philosophy of Science*, 3(1): 33–67.

Leveson, N. G. 2012. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press.

Leveson, N. G.; and Thomas, J. P. 2018. STPA Handbook. MIT Partnership for Systems Approaches to Safety and Security (PSASS).

Luccioni, A. S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. arXiv preprint arXiv:2303.11408.

MITRE. 2026. MITRE ATLAS. MITRE ATLAS.

Moss, E.; Watkins, E.; Metcalf, J.; and Elish, M. C. 2020. Governing with Algorithmic Impact Assessments: Six Observations. SSRN Electronic Journal.

Mylius, S. 2025. Systematic hazard analysis for frontier AI using STPA.

National Institute of Standards and Technology. 2023. AI Risk Management Framework. National Institute of Standards and Technology (NIST).

Nestaas, F.; Debenedetti, E.; and Tramèr, F. 2024. Adversarial Search Engine Optimization for Large Language Models. arXiv preprint arXiv:2406.18382.

OECD. 2024. Explanatory memorandum on the updated OECD definition of an AI system. OECD Artificial Intelligence Papers.

Parasuraman, R.; and Manzey, D. H. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3): 381–410.

Rismani, S.; Dobbe, R.; and Moon, A. 2024. From Silos to Systems: Process-Oriented Hazard Analysis for AI Systems. arXiv preprint arXiv:2410.22526.

Rismani, S.; Shelby, R.; Smart, A.; Delos Santos, R.; Moon, A.; and Rostamzadeh, N. 2023a. Beyond the ML Model: Applying Safety Engineering Frameworks to Text-to-Image Development. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 70–83.

Rismani, S.; Shelby, R.; Smart, A.; Jatho, E.; Kroll, J.; Moon, A.; and Rostamzadeh, N. 2023b. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Roytburg, D.; and Miller, B. 2025. Mind the Gap! Pathways towards Unifying AI Safety and Ethics Research. arXiv preprint arXiv:2512.10058.

United States Congress. 2010. Dodd-Frank Wall Street Reform and Consumer Protection Act. Public Law 111-203, 124 Stat. 1376. Signed into law July 21, 2010.

Upper, C. 2011. Simulation Methods to Assess the Danger of Contagion in Interbank Markets. *Journal of Financial Stability*, 7(3): 111–125.

Vanschoren, J. 2025. The Role of AI Safety Benchmarks in Evaluating Systemic Risks in General-Purpose AI Models. Publications Office of the European Union.

Walsh, M.; Schulker, D.; and Davison, T. 2025. Applications of CAST to Understand and Mitigate Risks in AI Systems. MIT STAMP Workshop, Carnegie Mellon University.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv preprint arXiv:2310.11986.