

Game Theory for AI Governance: Models, Mechanisms, and Research Perspectives

Dusit Niyato^{a,*}, Yinqiu Liu^a, Wanpracha “Art” Chaovalitwongse^b, Zehui Xiong^c and Zhu Han^d

^aNanyang Technological University, 50 Nanyang Avenue, 639798, Singapore

^bAccenture, Thailand

^cQueen’s University Belfast, University Road, Belfast, BT7 1NN, United Kingdom

^dUniversity of Houston, 4226 Martin Luther King Boulevard, Houston, Texas, 77204-4005, USA

ARTICLE INFO

Keywords:

AI governance

Game theory

Multi-agent systems

AI safety

Algorithmic auditing

Explainable AI

ABSTRACT

Artificial intelligence (AI) governance connects technical measures with decisions about AI development, deployment, and oversight. These decisions involve actors with different objectives, information, and control over AI systems. Game theory provides tools to analyze their interactions and design governance mechanisms. This perspective paper examines game theory for four governance tasks, namely (i) incentives and responsibility, (ii) regulation and AI markets, (iii) auditing, disclosure, and enforcement, and (iv) coordination. We first introduce AI governance, its relationship with AI risks and safety, and relevant game-theoretic foundations. We then compare the players, strategies, payoffs, solution concepts, theoretical results, and governance benefits within each task, and connect them to operational governance processes. This joint analysis supports comparison of existing applications and assessment of the requirements for applying game models to AI governance. We examine equilibrium analysis, contract design, contribution and responsibility allocation, and game-theoretic explanations. Feature attribution and interactive verification are assessed as sources of governance evidence. We identify research directions in AI ethics, adaptive governance, automated mechanism design, and scalable models grounded in empirical evidence. The resulting perspectives connect computational methods to organizational decisions and support research in computer science, business, and information systems.

1. Introduction

Deploying an artificial intelligence (AI) system requires decisions about acceptable risk, access to data and tools, and responsibility for its actions. These decisions become more difficult when multiple organizations provide parts of the same system [1] or when an AI agent can act across multiple services [2]. Singapore’s Model AI Governance Framework for Agentic AI addresses these problems through limits on agent permissions, human accountability, testing, and continuous monitoring [2]. For example, allowing an AI agent to access a customer database or initiate a payment requires both technical controls and a clear allocation of responsibility. This connection makes AI governance a practical concern for system design and organizational decision-making.

Government and industry initiatives also place governance within the AI lifecycle. The National Institute of Standards and Technology (NIST) AI Risk Management Framework defines governance functions, and its Generative Artificial Intelligence Profile addresses risks associated with generative AI [3, 4]. Microsoft’s 2026 Responsible AI Transparency Report describes governance practices that connect evaluation, release oversight, and monitoring after deployment [5]. The European Commission’s announcement of the General-Purpose AI Code of Practice identifies transparency, copyright, and safety and security as its three areas [6]. These initiatives show that governance requires evidence about system behavior and processes for deciding how to act on that evidence.

AI governance concerns the rules, responsibilities, and processes that guide AI development and use [7, 8]. Its technical, business, social, and political aspects are closely connected. Evaluation and access control affect the cost and timing of deployment. Business incentives influence investment in safety and the disclosure of model limitations. These choices affect privacy, fairness, and the distribution of benefits and risks, while public authorities and affected communities have interests in how requirements are defined and enforced [8–10]. Ethical principles do not by

*Corresponding author.

 dnnyato@ntu.edu.sg (D. Niyato)

themselves establish the practices needed to address these issues [11]. Therefore, governance must connect technical requirements with incentives, decision rights, and accountability.

For computer science research, this connection raises concrete problems in assessment, access, verification, security, and ecosystem monitoring [10]. Model evaluation, including assessment of extreme risks, provides information for development and deployment decisions [12–15]. Internal algorithmic auditing and the three-layered approach to auditing large language models (LLMs) connect assessments to organizational procedures [16, 17]. However, the effectiveness of these methods also depends on the behavior of the actors who develop, commission, or use them. An audit, for example, supports accountability only when relevant information is available and responsible actors have the authority and incentives to respond [18, 19]. Explainable AI provides a further source of assessment information. Feature-attribution games describe model dependence, while prover–verifier games examine whether a claim can be checked [20, 21]. Understanding these dependencies requires analysis of technical measures and strategic decisions together.

1.1. Why Game Theory for AI Governance?

Game theory models strategic interactions through players, strategies, payoffs, information, and solution concepts. It is useful for AI governance because an actor’s outcome depends on the decisions of other actors. Developers compete for users and decide on investment. Deployers select models and controls, and auditors decide how much effort to devote to an assessment. These actors can pursue different objectives while sharing the consequences of system failures. Models of AI development races illustrate how individual incentives can produce collective risks [22, 23].

Actors can impose costs on others without bearing those costs in full, creating a need to connect private incentives to governance objectives. Information is also distributed unevenly. A developer can know more about development practices than a deployer, while an auditor can know more about assessment effort than the organization commissioning the audit. Decisions also unfold over time, with investment, evaluation, release, and monitoring affecting both current outcomes and future opportunities [22, 24, 25]. We organize the governance decisions arising from these problems into four tasks. *Incentives and responsibility* concerns how to encourage safety investment and allocate rewards, costs, and responsibility. *Regulation and AI markets* examines how rules affect AI development, model access, and market competition. *Auditing, disclosure, and enforcement* addresses how evidence is produced, shared, and used to assess compliance and guide intervention. *Coordination* concerns cooperation and shared investment among organizations, communities, and jurisdictions.

Game theory also supports the design of mechanisms that account for these interactions. Contracts, reporting rules, inspection policies, and access conditions change the consequences of an actor’s choices. Zhang et al. develop this perspective through the Incentive Compatibility Sociotechnical Alignment Problem (ICSAP) [26]. In AI auditing, Bova et al. show how alternative payment rules change auditors’ incentives to maintain oversight [27]. Therefore, the design problem includes both the technical assessment and the incentives surrounding its production and use.

For computer science research, these questions lead to computational problems in equilibrium analysis, optimization, learning, and verification. A mechanism must specify observable inputs and implementable decisions. Its evaluation must examine strategic responses, computational cost, and the governance outcome of interest. The same analysis informs procurement, investment, delegation, and organizational accountability. Game theory provides a common framework for connecting these technical and organizational decisions.

1.2. Objectives and Contributions

The objective of this perspective paper is to examine how game theory can analyze and support AI governance decisions. The paper does not aim to provide a systematic literature review or exhaustive coverage of the literature. The technical focus is on model formulation, equilibrium analysis, computational solution methods, and the implementation and evaluation of governance mechanisms. The main contributions of this paper can be summarized as follows.

- **Background on AI governance and game theory.** We introduce the functions and components of AI governance and explain its relationship with AI risks and safety. We describe how game theory provides mathematical tools to analyze strategic interactions and optimize governance actions and decisions under specified objectives and constraints.
- **Game models organized by governance tasks.** We classify selected studies into (i) incentives and responsibility, (ii) regulation and AI markets, (iii) auditing, disclosure, and enforcement, and (iv) coordination. Within each task, we compare game models, information assumptions, solution methods, and governance outcomes, and

map their uses to ten operational processes. We discuss computational cost and implementation requirements where evidence is available.

- **Perspectives on applicable game models.** We examine game-theoretic methods for learning, data valuation, explainable AI, agent interaction, and computational mechanism design that can address emerging governance tasks. We explain the required changes to players, actions, information, and objectives. We identify which applications are proposed and which mechanisms have been evaluated in AI settings. The analysis includes how explanations can support audit and assurance evidence.
- **Future research directions.** We discuss how game theory can support AI ethics, adaptive governance, and automated mechanism design. We identify challenges in scaling the models, estimating payoffs from real-world data, and addressing incomplete and imperfect information, with candidate formulations and evaluation criteria.

The novelty of this perspective paper is the joint analysis of AI governance tasks, game models, and mechanism design requirements. This comparison explains why a game model is suitable for a particular decision and which assumptions limit its use.

1.3. Related Reviews and Positioning

Existing surveys provide foundations for both governance and its technical implementation. Batool et al. organize AI governance based on who is accountable, what is governed, when governance occurs, and how it is implemented [7]. Papagiannidis et al. examine structural, relational, and procedural practices for responsible AI governance [8]. Reuel et al. organize technical AI governance by capacities and targets and identify open problems [10]. Wei and Liu review robustness, privacy, fairness, and governance in distributed AI systems [28]. These works clarify governance requirements and technical research problems.

Other reviews examine alignment and strategic interaction. Ji et al. distinguish forward alignment from backward alignment, with assurance and governance included in backward alignment [29]. Han reviews evolutionary game theory for multi-agent systems, including the regulation of AI development [30]. Sun et al. survey game theory and LLMs through game-based benchmarks, algorithmic innovations, and societal impacts [31]. Their expanded version adds a fourth perspective on LLM-assisted game modeling and solving [32].

Two recent papers are closely related to this perspective paper. Zhang et al. develop the ICSAP perspective through mechanism design, contract theory, and Bayesian persuasion [26]. Cai et al. organize game-theoretic alignment around preference diversity, alignment priority, and temporal dynamics, with a main focus on post-training alignment [33]. These papers establish links between game theory, alignment, and governance. Their scope differs from this paper's comparison across governance tasks, which connects formal models and computational methods to mechanism implementation and evaluation. Table 1 summarizes the additional coverage and analysis.

Covert et al. organize removal-based explanations through feature removal, the explained behavior, and the summary of feature influence [34]. Their framework clarifies the connection between cooperative games and explanation methods. Section 5.2 connects this technical foundation to auditing, disclosure, and verification of governance evidence.

Figure 1 relates the game formulations in the study comparison tables to operational governance processes. We examine models that specify players or coalitions, actions or contributions, payoffs or values, and a solution concept relevant to an AI governance decision. The tables include direct applications and proposed governance uses, while background sources and studies discussed only for context are excluded from the coverage counts. The Sankey diagram shows how these associations are distributed across processes. Contracts and auctions are distinguished, and the additional Stackelberg/sequential, repeated, stochastic/Markov, and zero-sum labels capture structures that can occur together. AI-Control Games, for example, combine stochastic modeling with zero-sum reductions [35]. The diagram summarizes the study-level mapping used in the comparison tables.

Among the selected studies, evaluation and assurance receive the largest share of formulation–process associations. Non-cooperative and cooperative models both contribute prominently to these associations. The Sankey diagram makes the broader pattern visible. Several formulations inform each of these governance activities, and individual formulations extend across multiple activities. This overlap motivates the task-based comparisons in Sections 3–6, where the decision and its implementation requirements determine which model is useful. Next, we outline the structure of this paper.

1.4. Structure of the Paper

Figure 2 shows the organization of the paper. Section 2 introduces the governance and game-theoretic foundations.

Table 1

Positioning relative to related reviews and perspectives.

Related work	Main organizing focus	Additional analysis in this paper
Batool et al. [7] and Papagiannidis et al. [8]	Governance questions and organizational practices	Connect governance decisions to game models and solution methods.
Reuel et al. [10]	Technical AI governance: capacities and targets	Analyze strategic incentives, solution methods, and mechanism interactions within governance tasks.
Wei and Liu [28]	Robustness, privacy, fairness, and governance in distributed AI	Connect technical requirements to incentive design and institutional decisions.
Ji et al. [29]	Forward alignment and backward alignment	Compare models for incentives, markets, accountability, and coordination.
Han [30]	Evolutionary game theory, including AI development races	Compare several game types and their computational methods within the same tasks.
Sun et al. [31, 32]	Game theory and LLMs: evaluation, algorithms, societal impacts, and LLM-assisted game modeling	Organize the analysis by governance tasks across AI systems and their providers.
Zhang et al. [26]	ICSAP, mechanism design, contract theory, and Bayesian persuasion	Compare task-specific models, implementation requirements, and supporting evidence.
Cai et al. [33]	Preference diversity, alignment priority, and temporal dynamics	Examine governance decisions and the interaction of mechanisms across tasks.

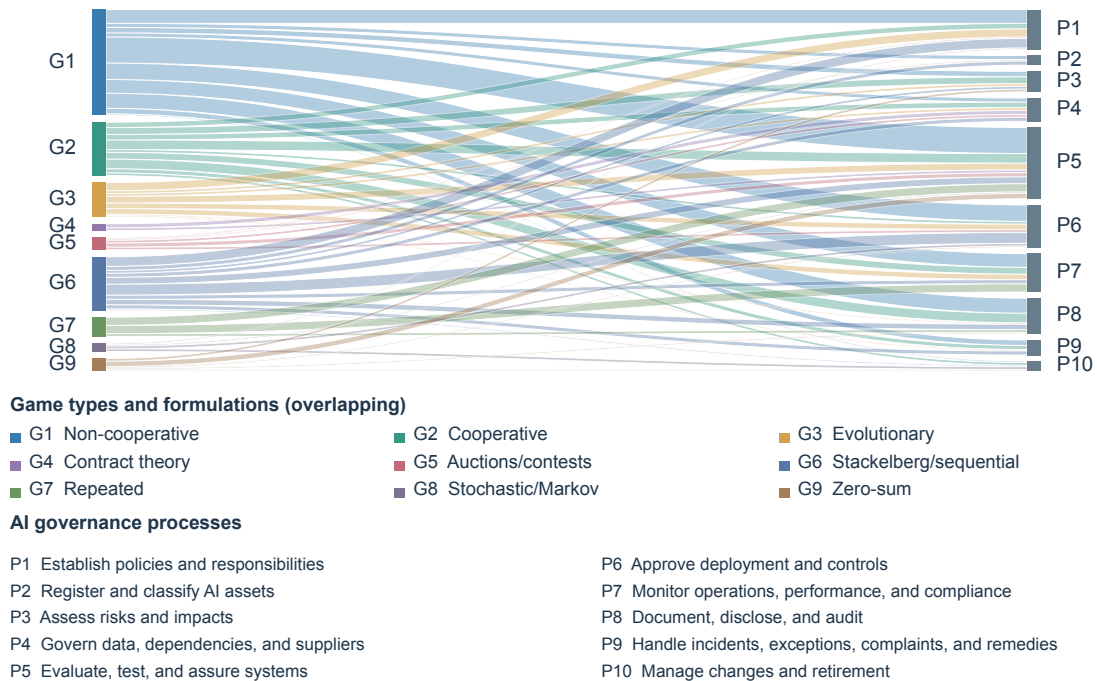
Game-type/process associations

Figure 1: Game types and formulations across governance processes for the 59 distinct studies in Tables 4–9. Ribbon widths show 223 game-type/process associations, with colors identifying formulations. G1–G5 describe game types and mechanisms; G6–G9 record timing, state, or payoff formulations, including zero-sum reductions. Assignments overlap, so a study can contribute to several ribbons. Cooperative models include coalition-value methods and shared-objective human–AI games.

Sections 3–6 examine the four tasks, and Section 7 draws out their shared implementation problems and research directions. Finally, Section 8 concludes this paper.

Game Theory for AI Governance

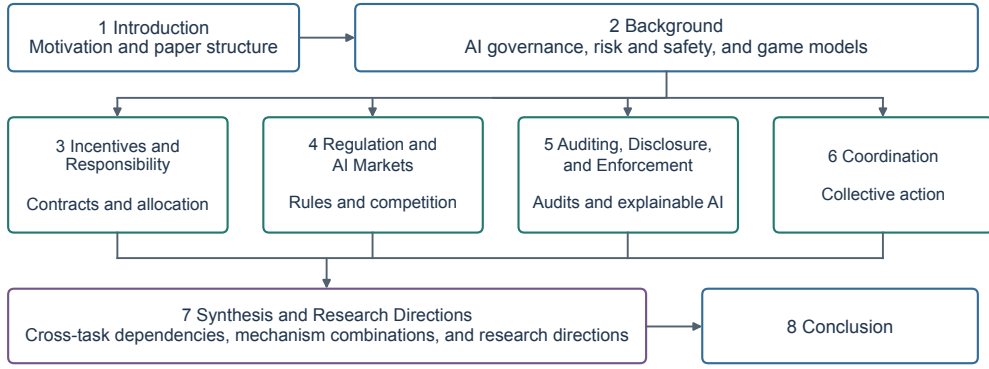


Figure 2: Organization of the paper.

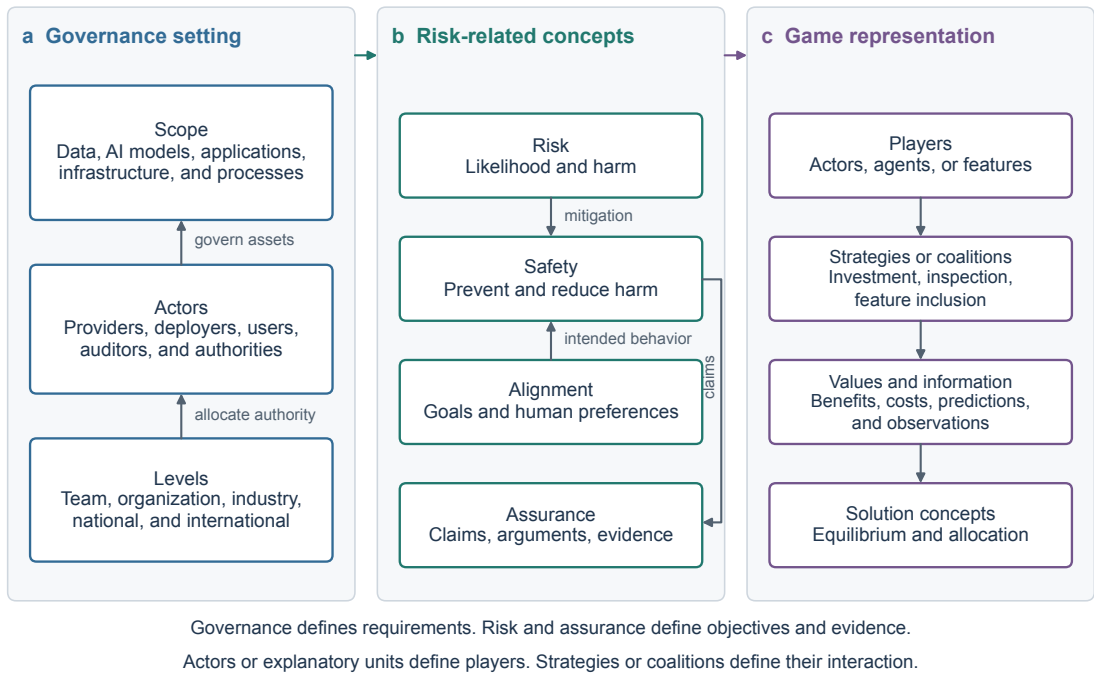


Figure 3: Connections between governance, risk-related concepts, and game models. (a) Governance specifies scope, actors, and levels. (b) Risk concerns likelihood and harm. Safety, alignment, and assurance concern harm reduction, intended goals, and supporting evidence. (c) Games specify players, strategies or coalitions, values, information, and solutions. Features can be cooperative-game players. Arrows show conceptual dependencies in our synthesis of Sections 2.1–2.5 [3, 10, 36].

2. Background on AI Governance, Risks, Safety, and Game Theory

This section introduces the scope, functions, and components of AI governance. It explains their relationship with AI risks, safety, alignment, and assurance. It then presents the game-theoretic foundations used in the subsequent sections. Figure 3 connects the governance setting, risk-related concepts, and game-model elements. It provides a common representation for the task-specific models in Sections 3–6.

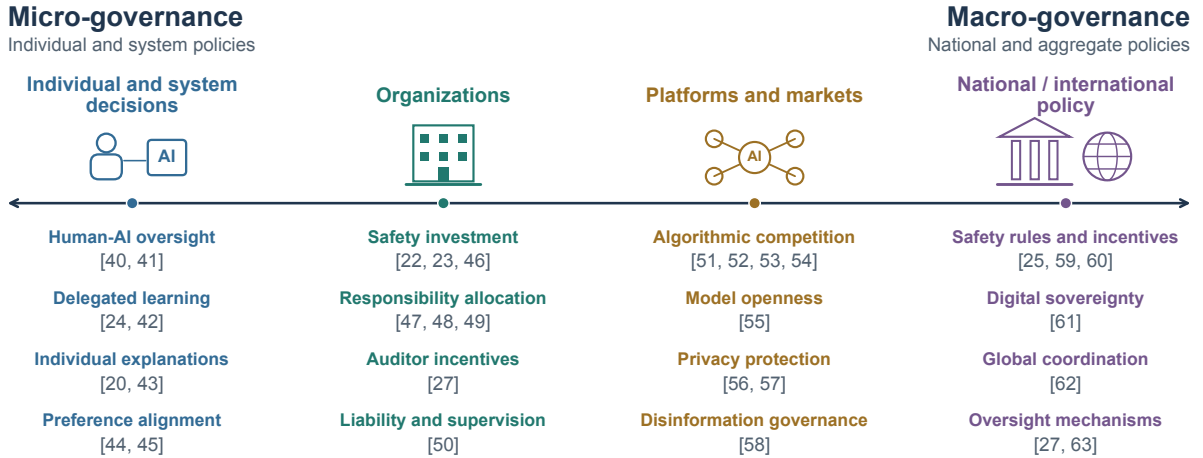


Figure 4: Representative game models by governance decision scope, from individual systems through organizations and markets to national and international policy. Placements indicate the primary application and may overlap.

2.1. AI Governance: Scope, Actors, and Levels

AI governance specifies how decisions about AI are made, implemented, and reviewed. It defines authority, objectives, and accountability across data, models, computing infrastructure, applications, and organizations [7, 10]. Because multiple actors can influence the same AI service, their responsibilities depend on the parts of the system they control [2]. Technical dependencies and data practices can connect their decisions across organizational boundaries [37, 38]. A provider develops or supplies an AI model, an application developer integrates it with other components, and a deployer determines its operational use [2]. Governance also involves data suppliers, infrastructure providers, users, auditors, public authorities, and affected communities, bringing together both the actors making decisions and the parties affected by them.

The allocation of responsibilities also depends on the level at which decisions are made. Batool et al. classify governance at the team, organization, industry, national, and international levels [7]. These levels are connected. An organizational policy can translate an external requirement into operational rules, while an industry initiative can support shared evaluation or reporting practices. Within an organization, governance is implemented through structural, procedural, and relational practices [8]. Structural practices allocate roles and decision rights, procedural practices establish assessment, approval, monitoring, and review, and relational practices support communication and cooperation. These arrangements, in turn, rely on technical analysis and tools to implement governance objectives [39]. A model release policy, for example, requires an evaluation procedure, access to relevant evidence, and a process for responding to identified risks. Organizational responsibilities and technical capabilities therefore work together to put governance requirements into practice.

To relate these governance levels to the game models examined here, we group representative models by their existing or proposed governance uses along a micro–macro spectrum in Figure 4. *Micro-governance* concerns decisions about individual actors and AI systems, such as effort incentives and human oversight. Organizational and platform arrangements form an intermediate, or *meso*, level, while *macro-governance* concerns national, regional, and international policies and their aggregate effects. This grouping helps examine how decisions at one level affect choices at another. A national rule can change organizational incentives and, through them, influence individual investment or oversight decisions.

Alongside decision scope, Figure 5 groups representative models by their potential governance decision horizon. Short-term governance responds to observed behavior through controls, monitoring, and rule updates. Longer-term governance shapes safety investment, commitments, and institutions for future AI capabilities. The horizon refers to the decision being governed. An operational control can serve a long-term objective, while a repeated game can describe rapid interaction. Section 7.1 uses scope and horizon together to identify where decisions at one level depend on commitments at another.

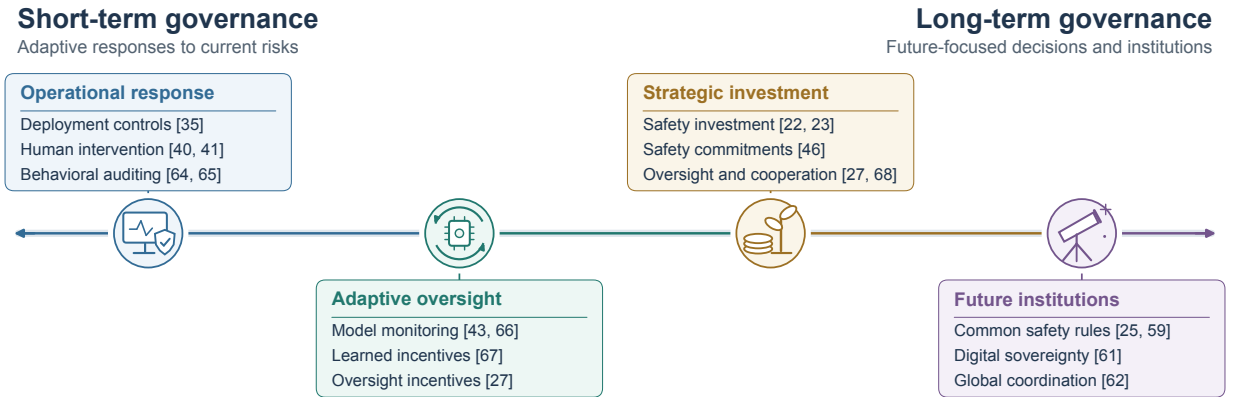


Figure 5: Representative models by governance decision horizon, from operational response to investment and institution building. Positions are qualitative; adaptive mechanisms can serve several horizons. Decision scope is shown separately in Figure 4.

2.2. Governance Functions, Components, and Lifecycle Processes

The NIST AI Risk Management Framework defines the four functions GOVERN, MAP, MEASURE, and MANAGE [3]. GOVERN establishes responsibilities and processes. MAP identifies the context and potential impacts of an AI system. MEASURE assesses risks and relevant system characteristics. MANAGE prioritizes risks and selects responses. GOVERN applies across the other functions, which support iterative use throughout development and deployment.

To describe the governance decisions compared in this paper, we organize governance components into six elements. This is an analytical organization used in this paper, informed by governance reviews [7, 8].

1. **Actors and responsibilities.** The participants, their authority, and their duties. These determine who can make, implement, inspect, or revise a decision.
2. **Governed assets.** The data, models, infrastructure, applications, and operational processes subject to governance. Their boundaries determine what can be observed and controlled.
3. **Objectives and requirements.** The outcomes and conditions that governance seeks to achieve. These can include safety, privacy, fairness, transparency, service quality, and acceptable resource use.
4. **Decision rules.** The rules for approval, access, reporting, reward, and corrective action. These translate requirements into decisions and consequences.
5. **Operational processes.** The activities that implement the rules, including evaluation, auditing, monitoring, incident response, and review.
6. **Evidence and feedback.** The records and observations used to assess claims, support decisions, and revise requirements. Examples include evaluation results, model and dataset documentation [69, 70], documented explanations with their assumptions, audit records, and incident reports.

The governance components provide a common basis for comparing the game models examined here. Actors and responsibilities identify the relevant players and decisions, while governed assets define the scope of the interaction. Objectives and decision rules inform the analysis of payoffs and incentives, and evidence clarifies what participants can observe. We use these elements to relate each model's formulation and assumptions to the governance problem it addresses. Additionally, we group operational governance activities into ten processes, based on lifecycle governance and risk-management frameworks [3, 7, 8].

- P1. Establish policies and responsibilities.** Define policies, standards, governance structures, operating arrangements, and accountability for AI-enabled and agent-enabled activities.

Table 2

Governance tasks, technical implementation, and process coverage. Process sets are the union of the study-level mappings within each task; tasks can arise at several lifecycle stages.

Governance task	Technical implementation	Governance processes and implications
Incentives and responsibility (Section 3)	Contribution measurement, performance records, controls on delegated actions.	P1, P3–P6, P8, P9. Reward effort, allocate costs, and assign responsibility.
Regulation and AI markets (Section 4)	Access conditions, evaluation requirements, resource allocation rules.	P1–P8. Relate requirements to investment, competition, and deployment.
Auditing, disclosure, and enforcement (Section 5)	Model documentation, explanation evidence, independent evaluation, audit and incident records.	P1–P10. Determine what to report and inspect, and how to enforce compliance.
Coordination (Section 6)	Shared evaluations, incident reporting, interoperable controls.	P1, P5–P9. Support participation, information sharing, and joint investment.

- P2.** Register and classify AI assets. Maintain an inventory of models, applications, deployed agents, owners, dependencies, and risk tiers across the AI portfolio.
- P3.** Assess risks and impacts. Examine effects on users, workers, security, and economic outcomes, including foreseeable misuse, affected groups, and failure scenarios.
- P4.** Govern data, dependencies, and suppliers. Set requirements for data, model providers, vendors, tools, and application programming interfaces (APIs).
- P5.** Evaluate, test, and assure systems. Validate intended uses and controls through assessment, red teaming, and repeated testing as systems or conditions change.
- P6.** Approve deployment and controls. Authorize release, permissions, and intervention conditions.
- P7.** Monitor operations, performance, and compliance. Track runtime behavior, risk, service quality, resource costs, productivity, capacity, and compliance against defined thresholds.
- P8.** Document, disclose, and audit. Maintain records, report limitations, and verify claims.
- P9.** Handle incidents, exceptions, complaints, and remedies. Establish escalation, runtime intervention, exception handling, corrective action, and routes for redress.
- P10.** Manage changes and retirement. Reassess updates and decide continued use or withdrawal.

These processes identify the operational governance activities to which the models examined here apply. In the study comparison tables, the P1–P10 labels indicate whether a model informs activities such as supplier governance, evaluation, deployment approval, or incident response. A model may contribute to several processes. For P2, we consider how regulatory models could inform risk-tier criteria for registered AI assets and how strategic-classification models could test responses to those criteria. Sections 3–6 organize the studies around the four governance tasks introduced earlier, grouping models that address related strategic problems across these activities. Table 2 summarizes the relationship between the four tasks and the operational processes.

2.3. Responsible AI, Risks, Safety, Alignment, and Assurance

Responsible AI concerns how organizations develop and use AI in ways that can be justified to affected people. It connects safety, fairness, privacy, transparency, and accountability to decisions, responsibilities, and evidence throughout the lifecycle [3, 8]. Governance makes these objectives operational by assigning decision rights, setting requirements, and checking their implementation. In practice, this includes determining who selects safety requirements, funds evaluations, authorizes deployment, and responds to incidents. Safety research provides methods and evidence for these decisions, while governance also addresses transparency, participation, fair access, and market organization [8, 9].

AI risks concern potential adverse effects of AI development and use. Assessing these risks helps identify the harms that governance requirements seek to prevent. For a harmful event, risk depends on its likelihood and the magnitude of harm, so an assessment must specify the event, the affected parties, and the deployment conditions, while accounting for uncertainty [3]. A measured failure rate provides evidence about a particular evaluation setting. Its relevance to deployment depends on how that setting relates to operational conditions. For generative AI, relevant

risks include information integrity, privacy, harmful bias, information security, and fabricated or unreliable outputs [4, 71]. For agentic AI, tool access and delegated authority also affect the consequences of system actions [2]. The same model can therefore present different risks when deployed with different permissions, users, or opportunities for human intervention.

AI safety concerns preventing and reducing harm under these conditions. Amodei et al. identify problems involving side effects, reward hacking, scalable supervision, safe exploration, and distributional shift [72]. These problems link possible failures to the learning objective, training process, and deployment environment, showing why safety assessment must consider more than predictive performance. Alignment and security address related concerns. Alignment examines how system objectives and behavior relate to intended goals and human preferences [29, 73, 74], while security addresses threats such as unauthorized access and manipulation [3, 10]. A poorly specified objective raises an alignment concern, unfamiliar operating conditions can create safety risks, and adversarial interference raises security concerns. These sources of harm can overlap, so governance decisions must consider both the relevant risks and the controls available to address them.

AI assurance concerns the arguments and evidence supporting claims about an AI system [17, 75]. A safety case organizes this material for a deployment decision. It specifies the system and conditions of use, states the safety claim, and explains how the evidence supports that claim. Explainable AI can contribute by helping people understand an output or its dependence on inputs. Cooperative feature attribution and interactive verification provide different forms of explanation [20, 21]. Their contribution to an assurance claim depends on the explanation's target, assumptions, and reliability, together with other evidence relevant to safety, fairness, or causal responsibility. For example, a system may be explainable while still exposing users to unacceptable harm or leaving them without an effective way to challenge a decision. Section 5.2 examines how explanation methods can support governance evidence.

Governance decisions bring these objectives, risks, and evidence together. Requirements for safety, alignment, and security reflect the broader aims of responsible AI and the risks associated with a particular use. Assurance organizes the arguments and evidence used to assess whether these requirements are met under the intended deployment conditions. Together, these requirements and assessments inform decisions about deployment, monitoring, and intervention. Game theory examines the incentives and interactions behind these decisions, including how actors invest in safeguards, produce and disclose evidence, and respond to oversight.

2.4. Elements and Solution Concepts of Governance Games

A game model specifies players, actions, information, payoffs, and the order of decisions [36]. Players can represent firms, regulators, auditors, users, or AI agents. Actions include investment, disclosure, inspection, release, and cooperation. A strategy specifies how a player selects actions using available information. Payoffs represent the objectives attributed to that player.

The information structure determines what each player knows before and during decision making. Complete information concerns knowledge of the game's characteristics, including payoffs. Perfect information concerns observation of previous actions. The timing determines whether a player can commit before others respond. These structures affect both strategic behavior and the information needed to implement a mechanism.

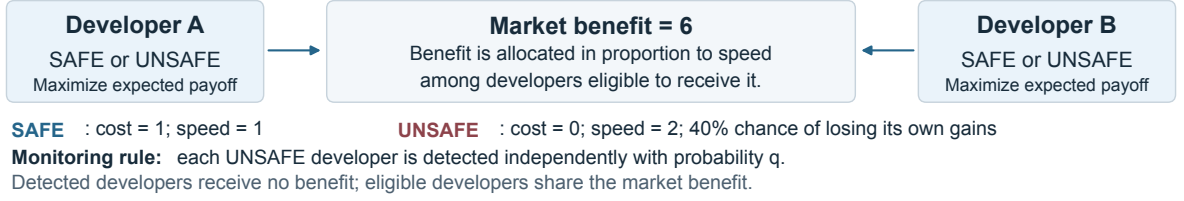
For a normal-form game, let N denote the set of players, A_i player i 's feasible actions, and $u_i(a_i, a_{-i}; m)$ its payoff under governance rule m . A pure-strategy Nash equilibrium a^* satisfies

$$u_i(a_i^*, a_{-i}^*; m) \geq u_i(a_i, a_{-i}^*; m), \quad \forall i \in N, \forall a_i \in A_i. \quad (1)$$

A Nash equilibrium is a strategy profile in which no player gains from changing its own strategy while the other strategies remain fixed. For example, a provider's safety investment must be a best response to competitors' investments and the applicable audit rule. A pure equilibrium need not exist. Finite games admit an equilibrium in mixed strategies, with the inequality applied to expected payoffs [36]. Equilibrium does not imply that the outcome is safe, fair, or socially efficient. Figure 6 illustrates how development speed, safety costs, and monitoring determine developers' expected payoffs and change their equilibrium strategies.

(a) Game formulation

Illustrative adaptation of [23]


 (b) No monitoring ($q = 0$)

		Developer B	
		SAFE	UNSAFE
Developer A	SAFE	(2, 2)	(1, 2.4)
	UNSAFE	(2.4, 1)	(1.8, 1.8) Nash equilibrium

Expected payoffs: (Developer A, Developer B)

UNSAFE is strictly dominant.

Against SAFE: UNSAFE gives $2.4 > 2$.
Against UNSAFE: UNSAFE gives $1.8 > 1$.

Best-response comparisons apply to both developers.

 (c) Monitoring ($q = 0.5$)

		Developer B	
		SAFE	UNSAFE
Developer A	SAFE	(2, 2) Nash equilibrium	(3, 1.2)
	UNSAFE	(1.2, 3)	(1.35, 1.35)

Expected payoffs: (Developer A, Developer B)

SAFE is strictly dominant.

Against SAFE: SAFE gives $2 > 1.2$.
Against UNSAFE: SAFE gives $3 > 1.35$.

Best-response comparisons apply to both developers.

Governance role: monitoring and exclusion make SAFE the best response in this example.

Figure 6: An illustrative safety-investment game adapted from the SAFE/UNSAFE payoff structure in [23]. (a) Two developers choose SAFE or UNSAFE to maximize expected payoffs. SAFE has speed 1 and cost 1. UNSAFE has speed 2 and no cost, but a 40% probability of losing its own gains. When both developers are eligible, developer i receives $6s_i/(s_i + s_j)$ from a total market benefit of 6, where s_i and s_j denote their speeds. Equal speeds give each developer 3. In a SAFE–UNSAFE pair, the faster UNSAFE developer receives 4 and SAFE receives 2. Speed changes the relative allocation, not the total benefit. (b) Without monitoring, mutual SAFE gives each developer $3 - 1 = 2$. A SAFE–UNSAFE pair gives payoffs $(2 - 1, 0.6 \times 4) = (1, 2.4)$, while mutual UNSAFE gives each $0.6 \times 3 = 1.8$. Thus, UNSAFE is strictly dominant and mutual UNSAFE is the unique Nash equilibrium. (c) Monitoring independently detects each UNSAFE developer with probability $q = 0.5$. A detected developer receives no benefit, and a sole eligible developer receives all 6. In a SAFE–UNSAFE pair, SAFE obtains $0.5 \times 2 + 0.5 \times 6 - 1 = 3$, while UNSAFE obtains $0.5 \times 0.6 \times 4 = 1.2$. Under mutual UNSAFE, each obtains $0.6(0.25 \times 3 + 0.25 \times 6) = 1.35$. Mutual SAFE still gives (2, 2). Therefore, SAFE becomes strictly dominant and mutual SAFE is the unique Nash equilibrium. Payoff pairs list A first and B second. The numerical values are illustrative. Monitoring is fixed, and personal setbacks are independent of detection. Monitoring costs, false positives, the final race prize, and evolutionary dynamics are omitted.

Stackelberg models add a commitment sequence. A leader chooses $\ell \in L$, after which a follower chooses $f \in F(\ell)$. Define the follower's best-response set and a selected response by

$$B_F(\ell) = \arg \max_{f \in F(\ell)} u_F(\ell, f), \quad b(\ell) \in B_F(\ell). \quad (2)$$

Given that response rule, the leader solves

$$\ell^* \in \arg \max_{\ell \in L} u_L(\ell, b(\ell)), \quad f^* = b(\ell^*). \quad (3)$$

An auditor, for example, may commit to an inspection policy before a provider selects its compliance effort. If the follower has several best responses, the model must specify how b is selected because favorable and unfavorable tie-breaking can lead to different leader decisions [76]. The governance interpretation also requires the leader to have authority and a credible means of implementing its commitment.

These definitions express assumptions about responses to specified incentives. They do not assert that people or AI systems always optimize consistently [77]. Limited information, computational constraints, learning, and behavioral biases may require approximate, Bayesian, or evolutionary formulations [36]. An ϵ -equilibrium, for instance, allows each player a gain of at most ϵ from a unilateral deviation. Bayesian Nash equilibrium accounts for private types and beliefs, while subgame-perfect equilibrium requires equilibrium behavior in every subgame [36].

Cooperative concepts allocate coalition value. The core rules out profitable coalition deviations, whereas the Shapley value assigns average marginal contributions. For a transferable-utility game with coalition value v , $v(\emptyset) = 0$, and n players, the Shapley allocation to player i is [78]

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]. \quad (4)$$

The allocation depends on how coalition value is defined. In a governance application, interpreting it as moral or legal responsibility requires additional normative assumptions.

A governance mechanism specifies rules, such as payments, reporting procedures, and access conditions. A solution concept specifies the behavior or allocation being analyzed. An algorithm computes or learns a solution under those rules [79]. These objects serve different roles. An algorithm can solve a specified game accurately while the game's assumptions provide an incomplete representation of the governance problem.

Computational feasibility is another requirement. General equilibrium computation can be difficult [80]. Therefore, practical methods can use restricted strategy spaces, approximation, sampling, or learning. Multi-agent reinforcement learning provides methods for learning policies through interaction [81]. Their use requires attention to observability, changes in other agents' policies, and the relationship between empirical performance and the claimed solution concept.

2.5. Game Models and Their Roles in AI Governance

Model selection begins with the governance decision. In non-cooperative games, actors choose strategies according to their own objectives. Normal-form representations specify complete strategies and payoffs, while extensive-form representations make decision sequences and information sets explicit. These models identify conflicts between individual incentives and collective outcomes. AI development races provide examples in which strategies for safety depend on competitors' actions [22, 23]. The relevant governance question is how a change in rules affects those strategies and the resulting risk.

Stackelberg games can represent an authority announcing an inspection policy before firms choose compliance effort. Paruchuri et al. solve Bayesian Stackelberg security games using mixed-integer linear programming [76]. Transfer to AI governance requires specifying what the authority can commit to, what participants observe, and how their responses are modeled.

Bayesian games represent private information and beliefs. Signaling models examine how observable actions affect beliefs. Bayesian persuasion studies commitment to an information structure that influences a receiver's decision [82]. These models support analysis of disclosure and sociotechnical alignment [26]. Applications must identify the private information, the credibility of a signal, and the decisions that the signal can change.

Repeated games allow strategies to depend on previous interactions. Stochastic games also include states that change with actions and uncertainty. Partial observability restricts decisions to the available observations. Repeated pricing games provide an example of persistent strategic interaction [51]. AI-Control Games use a multi-objective, partially observable, stochastic game to study deployment protocols [35]. These models are relevant when monitoring, reputation, or future access changes present incentives.

Evolutionary game theory studies changes in the prevalence of strategies through performance, imitation, or selection. Replicator dynamics describe population shares, while stochastic population models account for random changes in finite populations. AI race and auditing studies use these methods to examine the adoption of practices under different incentives [23, 27]. Their interpretation depends on the assumed strategy revision process and population structure.

Cooperative games represent the value that coalitions can produce. Bargaining additionally accounts for disagreement outcomes and negotiating power. Data Shapley applies contribution-based valuation to training data [83]. In explainable AI, SHAP and SAGE instead treat features as players and allocate a prediction or predictive performance [20, 84]. Features serve as explanatory units in a value allocation. Governance applications require a value function that represents the intended outcome.

Table 3

Game models and frameworks for AI governance. Uses include existing applications and possible applications.

Model or framework	Key structure and methods	Governance processes and implications
Non-cooperative games	Individual objectives, equilibrium analysis.	P1, P6. Competition and safety investment [22, 23].
Stackelberg games	Commitment and follower responses, bilevel or mixed-integer optimization.	P5, P6, P7. Inspection, access policies, and evidence verification [21, 76].
Bayesian games and information design	Private information, equilibrium or signal optimization.	P6, P8. Disclosure and strategic reporting [26, 82].
Repeated and stochastic games	History and evolving states, dynamic programming and policy optimization.	P7, P10. Monitoring and deployment protocols [35].
Evolutionary games	Strategy revision in populations, replicator dynamics or stochastic processes.	P1, P7. Adoption of safety practices and auditing incentives [23, 27].
Cooperative games and bargaining	Coalition value and disagreement outcomes, allocation rules.	P4, P8, P9. Contribution rewards, shared costs, and explanation evidence [20, 83].
Mechanism design and contract theory	Strategic participation and hidden effort, optimization under incentive constraints.	P4, P8. Rewards, contracts, and reporting mechanisms [85, 86].

Mechanism design selects rules for strategic participants. Incentive compatibility makes prescribed behavior optimal under the specified model. Individual rationality makes participation preferable to an outside option [85]. Contract theory addresses hidden information, hidden actions, and control rights [86]. These issues arise when development or evaluation effort is costly and difficult to observe [26, 27]. A governance application must specify which outcomes can be verified and which commitments can be enforced.

Feedback between deployment and future observations is also important. Perdomo et al. study performative prediction, where deploying a predictor changes the distribution on which it is evaluated [87]. This provides a complementary account of feedback. A strategic model additionally specifies the actors' objectives and responses. For governance, the connection motivates analysis of how evaluation rules, contracts, and deployment decisions change the evidence available in later periods [88, 89]. Table 3 summarizes the main models and their governance uses.

3. Incentives and Responsibility

AI governance requires incentives for responsible development and a basis for assigning responsibility [26, 47]. These tasks concern who bears costs, who receives benefits, and who controls actions. They arise when development effort is difficult to observe, several participants contribute to a system, or multiple actors can influence an outcome [24, 49, 83]. The decisions range from setting responsibilities and assessing risk to governing suppliers, evaluating systems, approving deployment, documenting contributions, and assigning remedies. Table 2 summarizes the corresponding process coverage.

This section examines safety investment, delegated learning, contribution rewards, responsibility allocation, and liability and insurance. The discussion connects each task to its game model, solution method, and governance role. Table 4 summarizes the game details and governance benefits.

3.1. Incentives for Safety Investment

Safety investment creates a conflict between completing AI development first and avoiding shared harm [90]. Competing teams choose precautions that improve safety but reduce their competitive advantage [22]. Payoffs depend on the winner, safety, and preferences over another team's success. Nash equilibrium analysis shows how more competitors, greater enmity, and capability information can increase risk. This conflict extends across sequential development steps. SAFE and UNSAFE strategies trade safety costs and slower progress against intermediate returns, a final prize, and disaster risk [23]. Fixation probabilities and stationary distributions identify conditions under which unsafe strategies spread despite collective harm.

Governance can address this dilemma through commitments followed by enforcement. Developers can voluntarily agree to act safely, with peer or institutional sanctions for violations [46]. Evolutionary analysis shows how conditional sanctions support safety in the social-dilemma region while preserving beneficial risk-taking in low-risk settings. Enforcement can also change development speed. Safe developers bear a speed cost to slow unsafe competitors or accelerate safe competitors [91]. Punishment promotes safety but can cause over-regulation, whereas rewards avoid this

AI race dynamics: development progress and strategy evolution

Illustrative simulation based on the published model [91] • Two distinct timescales

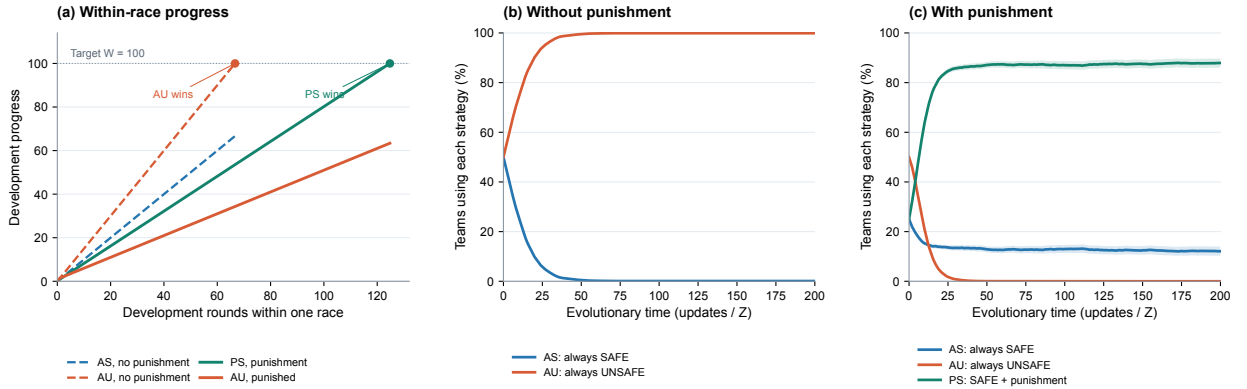


Figure 7: Illustrative AI race dynamics based on the development-race payoffs and pairwise comparison rule in [91]. AS always follows safety precautions. AU always ignores them. PS follows them and punishes an unsafe opponent. (a) Development progress using the fixed-target representation of the payoff calculation. Dashed lines show AS against AU. Solid lines show PS against AU. Each pair stops when its winner reaches $W = 100$. After the first round, punishment reduces AU's speed from 1.5 to 0.5 and PS's speed from 1 to 0.8. AU wins the unpunished race at approximately 66.7 rounds, whereas PS wins the punished race at 124.75 rounds. (b,c) Population trajectories under Fermi imitation, with $Z = 100$, $\beta = 0.01$, and no mutation. Initial counts (AS, AU, PS) are (50, 50, 0) and (25, 50, 25), respectively. Lines show means over 1,000 independent simulations. Shading gives pointwise 95% Monte Carlo confidence intervals for the means. One unit of evolutionary time comprises Z update attempts. Other parameters are $B = 10,000$, $b = 4$, $c = 1$, and $p_r = 0.6$. These transient simulations use illustrative punishment parameters $s_a = 0.2$ and $s_\beta = 1$. They do not reproduce the source's rare-mutation stationary analysis. The remaining AS and PS frequencies reflect neutral drift once AU disappears.

effect in some regimes. These results connect agreement formation and enforcement costs to the safety–innovation trade-off. Figure 7 illustrates development progress and population-level strategy evolution under the punishment model.

These incentives operate within a competitive setting shaped by entry and interaction. In an all-pay contest, teams choose participation and costly R&D investment for a prize [92]. Equilibrium participation favors sufficiently competitive teams, connecting taxation and procurement to race composition. Among participating developers, heterogeneous networks determine whose payoffs are observed and whose strategies are imitated [93]. Large differences in participants' network degree can reduce the safety dilemma. The contest model concerns who enters the race, while the network model concerns how safety practices spread among participants. These findings also motivate the collective-investment questions in Section 6.1.

Incentives may also need to respond as the race progresses. In a framed experiment, paired participants choose safe or unsafe development, trading faster progress and higher returns against accumulating private risk [94]. The preregistered comparison between risk conditions was not supported. Exploratory analyses nevertheless associate unsafe actions with falling behind and an opponent's previous unsafe action. The experiment points to competitive position as a factor to test when designing adaptive incentives. Applying the race models to operational safeguards still requires estimates of safety costs, risk reduction, competitive delay, and enforcement capacity (Section 5).

3.2. Contracts for Delegated Learning

Delegation introduces a measurement problem because a principal benefits from learning but cannot fully observe a supplier's or developer's effort [24, 26, 42]. Contracts link payments to observable outcomes, accounting for effort costs, participation, and incentive compatibility [24, 86]. This approach connects AI alignment to incentives within a sociotechnical system [26].

The payment rule first depends on what can be measured. For strategic data sources, costly effort reduces observation noise, and payments compare each report with an estimate excluding it [42]. Under the model's assumptions, the mechanism induces a unique dominant-strategy equilibrium and individual rationality. Worker and effort allocation minimizes weighted estimation error plus payments, although computing the optimum can be NP-hard. Moving from

data provision to delegated classification, providers choose training-sample numbers and receive validation-accuracy payments [24]. The principal minimizes the maximum payment needed to induce a target action. Under the Concave Monotone Likelihood Ratio Property, the Simple Binding Algorithm yields an optimal threshold contract. Thus, procurement guarantees depend on the statistical relationship between effort and measured performance.

Participant heterogeneity then introduces selection and private information. In federated learning, a task publisher combines reputation-based selection with computing-resource–reward contracts for devices with private data-quality characteristics [95]. The owner values timely training less payments, while devices value rewards less training and communication energy costs. Incentive compatibility and individual rationality constrain the menu. Zhan et al. instead model a server that announces a total reward before edge nodes choose their training-data volumes [96]. The reward is shared in proportion to those volumes, assuming independently and identically distributed data of equal quality. The server values learning benefits less payments, while nodes value their reward shares less computation and communication costs. The paper derives a Stackelberg solution with known information and uses deep reinforcement learning to adapt rewards when that information is unavailable.

Procurement can further select among competing offers. FMore uses a multidimensional auction in which devices bid resource qualities and prices, and earn payments less contribution costs [97]. Its analysis characterizes equilibrium bidding and efficient-selection conditions. Learned auction design provides a related method for revenue maximization, with allocation and payment networks penalizing bidders' gains from misreporting [98]. Generalization bounds and experiments support approximate incentive compatibility.

Across these mechanisms, the implementable contract is limited by the quality of the performance measure [24, 86] and the credibility of payment commitments. Studies of algorithmic hiring illustrate why prediction targets and validation procedures require scrutiny [99]. For evaluation services, test coverage and evidence quality are plausible contract variables, but their relationship to costly assessment effort needs validation. Section 7.2 develops this measurement problem alongside audit design. Section 7.4 considers the additional changes needed to adapt learned auctions to procurement.

3.3. Contribution Measurement and Reward Allocation

Reward allocation begins by defining who contributes and what their joint contribution produces. Cooperative games represent data contributors as players and data subsets as coalitions. The Shapley value allocates the specified coalition value through marginal contributions, satisfying efficiency, symmetry, additivity, and the null-player property [78]. Data Shapley applies this allocation to predictive performance, using Monte Carlo and gradient-based approximations to reduce evaluation costs [83]. Identifying helpful or harmful data can support supplier assessment and contribution records. Related methods address efficient Shapley computation and valuation relative to a data distribution [100, 101]. Robust valuation addresses noisy performance scores and uncertain validation distributions [102, 103]. Influence functions provide a complementary way to trace predictions to training data [104].

Scaling this approach changes how contributions are measured. In-Run Data Shapley accumulates values across gradient updates within one training run [105]. First- and second-order Taylor approximations reduce retraining costs, supporting foundation-model contribution records. These values describe the observed training process, while conventional Data Shapley evaluates subsets through retraining. The subsequent decision also matters. Without restrictions on the utility function, Shapley scores need not distinguish better data subsets more reliably than random guessing. They support optimal selection for monotonically transformed modular functions [106]. The contribution studies grouped in Table 4 therefore distinguish valuation from selecting data or suppliers.

Turning values into rewards adds participation, budget, and incentive constraints [107, 108]. Replication and contribution splitting require separate analysis [109]. A proposed governance application would value robustness or failure detection and test whether the resulting payments encourage useful contributions. This connects valuation to contract design (Section 3.2). Feature attribution (Section 5.2) uses related allocation principles, while responsibility additionally requires information about actions and prevention capacity.

3.4. Responsibility Allocation in Multi-Agent Systems

Responsibility analysis moves from measuring contributions to asking who could prevent an outcome. Extensive-form games represent actors' actions, decision order, and information [47]. Forward responsibility concerns possible plays, while backward responsibility concerns a realized play. Coalition responsibility supplies the value function for individual Shapley allocations. A specified coalition's responsibility can be checked in polynomial time, although full allocation can require many coalitions. These results support prevention duties and incident analysis.

Table 4

Representative models for incentives and responsibility.

Study and task	Game details	Governance processes and implications
Armstrong et al. [22] and Han et al. [23, 46]: safety investment	Type: non-cooperative development race, stochastic evolutionary games. Players: competing developers. Strategies: precautions, safe/unsafe development, and commitments. Payoff: development benefits, costs, and disaster losses. Solution: Nash equilibria, fixation probabilities and stationary strategy distributions.	P1, P5, P6. Explain how competition changes safety investment and when commitments support responsible development.
Cai et al. [42] and Saig et al. [24]: delegated learning	Type: principal-agent mechanism and contract design. Players: estimator/buyer and data workers/model provider. Strategies: effort or sample acquisition, payment rule. Payoff: payment less effort cost, estimation quality and procurement cost. Solution: unique dominant-strategy equilibrium (Cai), optimal incentive-compatible threshold contract under the likelihood-ratio condition (Saig).	P4, P5. Reward costly data or training effort through measured outcomes, connect procurement to learning performance.
Ghorbani and Zou [83] and Wang et al. [105, 106]: contributions	Type: cooperative valuation. Players: data contributors. Coalitions: subsets of training data or update contributions, the coalition value measures joint performance. Value: specified learning performance. Solution: Shapley allocation and approximations.	P4, P8. Attribute measured learning value and support contribution records and reward design.
Baier et al. [47, 48]: responsibility	Type: extensive-form games and cooperative attribution in parametric Markov models. Players: decision-makers or parameter contributors. Strategies: outcome-preventing plans, parameter changes in attribution. Objective/value: avoid an outcome or allocate a property change. Solution: responsibility criteria and Shapley/Aumann-Shapley allocations.	P1, P3, P9. Identify prevention capacity and allocate responsibility or changes in modeled system properties.
Karimian et al. [49]: concurrent responsibility	Type: Situation Calculus Synchronous Game Structures. Players: concurrent agents. Strategies: actions and outcome-preventing alternatives. Objective: causal involvement and strategic ability to prevent an outcome, no monetary payoff is required. Solution: causal, strategic, and combined responsibility predicates.	P1, P9. Analyze responsibility when concurrent actions, preemption, or multiple sufficient causes affect an outcome.

When the concern is a change in system risk, parametric Markov chains provide a complementary representation [48]. Parameters become attribution units, and path-based schemes, including Aumann-Shapley attribution, allocate changes in a specified system property. The authors establish decidability results for threshold questions about these allocations. Applied to a system model, the method can attribute changes in modeled failure probability to component parameters. Causal claims about an observed incident require additional evidence.

Concurrent actions introduce a further problem because causal involvement and the ability to prevent harm can yield different responsibility assignments. Situation Calculus Synchronous Game Structures represent concurrent agents and combine causal and strategic responsibility [49]. The analysis handles preemption and multiple sufficient causes and shows that causal, strategic, and combined notions need not coincide. The result characterizes responsibility through actions and available alternatives. For AI services, permission records, tool calls, and incident traces could help reconstruct actions and feasible alternatives. Additional evidence about each actor's knowledge and control is needed to connect the findings to contractual duties (Section 3.2) and liability (Section 3.5).

3.5. Liability Allocation for AI Services

Responsibility analysis identifies prevention capacity, while liability allocation changes incentives by assigning losses. Obidzinski and Oytana model a sequence in which a policymaker sets liability sharing, a manufacturer chooses autonomy investment and price, and users decide whether to adopt the service and intervene when needed [50]. Manufacturer payoff combines revenue, investment costs, and assigned liability, while user payoff combines service benefits, intervention costs, and expected harm. In the baseline model, attention decreases with autonomy. An extension makes attention depend on costly cognitive effort. Equilibrium comparisons characterize liability sharing and show how behavioral inattention changes investment and adoption. The distinction between the baseline and extension matters when using the model to design supervision requirements and remedies.

Insurance extends this incentive problem by transferring part of the remaining loss to an insurer. Section 7.2 develops it as an open modeling question, with attention to precaution incentives, audit evidence, and correlated failures. Table 5 compares additional game models for incentives and risk allocation.

Table 5

Additional AI-related game models for incentives and risk allocation. Payoffs and solutions refer to the cited model.

Study and task	Game details	Governance processes and implications
Naudé and Dimitri [92]: development race	Type: all-pay contest. Players: competing developers. Strategies: entry and R&D investment. Payoff: winning value less investment, the alternative objective maximizes success probability subject to non-negative profit. Solution: contest equilibrium and entry conditions.	P1, P6. Connect competition and entry to policy incentives, qualify conclusions through the development objective.
Han et al. [91] and Cimpanu et al. [93]: safety incentives	Type: evolutionary games. Players: developers in populations or networks. Strategies: safe/unsafe development and incentive responses. Payoff: development returns, safety costs, disaster losses, and incentive-dependent speed. Solution: evolutionary prevalence and stability under the modeled dynamics.	P1, P5, P6. Identify risk regimes and network conditions in which incentives support safety.
Kang et al. [95]: federated-learning participation	Type: contract design with reputation. Players: task publisher and workers with private types. Strategies: computing-resource–reward contracts and contract choice. Payoff: timeliness benefit less payment, payment less energy cost. Solution: individual-rationality and incentive-compatibility constrained optimization.	P4. Support reliable participation and cost screening in distributed model development.
Zhan et al. [96] and Zeng et al. [97]: distributed training	Type: Stackelberg incentive game (Zhan), multidimensional procurement auction (Zeng). Players: server/publisher and devices. Strategies: total reward and training-data volume (Zhan), bids and resource attributes (Zeng). Payoff: learning value and transfers net of costs. Solution: leader–follower optimization, auction equilibrium analysis.	P4, P5. Connect resource allocation and participant incentives to training quality and efficiency.
Obidzinski and Oytana [50]: AI liability	Type: sequential non-cooperative game. Players: policymaker, manufacturer, users. Strategies: liability sharing, autonomy, price, adoption, intervention; attention effort in an extension. Payoff: welfare, profit, or service value net of costs and assigned losses. Solution: sequential optimization and liability-rule comparison.	P1, P6, P9. Analyze how loss allocation changes autonomy, supervision, and adoption across users.
Dütting et al. [98]: learned auction design	Type: learned auction mechanism. Players: designer and bidders. Strategies: allocation/payment rules and value reports. Payoff: revenue for the designer and allocated value less payment for bidders. Solution: neural optimization with regret penalties and generalization bounds.	P4, P5. Inform procurement of AI evaluation services with strategic suppliers and quality constraints.

3.6. Summary and Lessons Learned

Safety incentives depend on competitive pressure, the cost of delay, and the credibility of enforcement. Contracts require an observable outcome that distinguishes the costly actions that the principal wants to induce. Contribution values and responsibility criteria can inform rewards or remedies, but the value being allocated and the actor’s ability to prevent harm answer different questions. Their joint use therefore requires a clear account of effort, measurement, and control.

4. Regulation and AI Markets

Regulation changes the conditions under which AI systems are developed, released, and used [25, 55, 59, 110]. Its effects depend on how model providers, application developers, and users respond. Game theory helps analyze these responses and identify conditions under which a requirement improves governance outcomes. The decisions span policy and risk classification, supplier obligations, assessment, deployment, monitoring, and disclosure, as summarized in Table 2. This section examines safety standards, model openness, compliance incentives, privacy, explainability, and algorithmic competition.

4.1. Safety Standards across the AI Development Chain

Sequential games are useful when regulation affects several stages of AI development. A general-purpose model provider determines the initial capabilities and safety properties of an AI model. A domain specialist then adapts the AI model for an application. A requirement imposed at either stage can change investment at both stages.

Laufer et al. model a regulator setting minimum safety requirements, followed by a general-purpose AI creator investing in safety and performance and a specialist adapting the model [25]. Returns combine shared application revenue

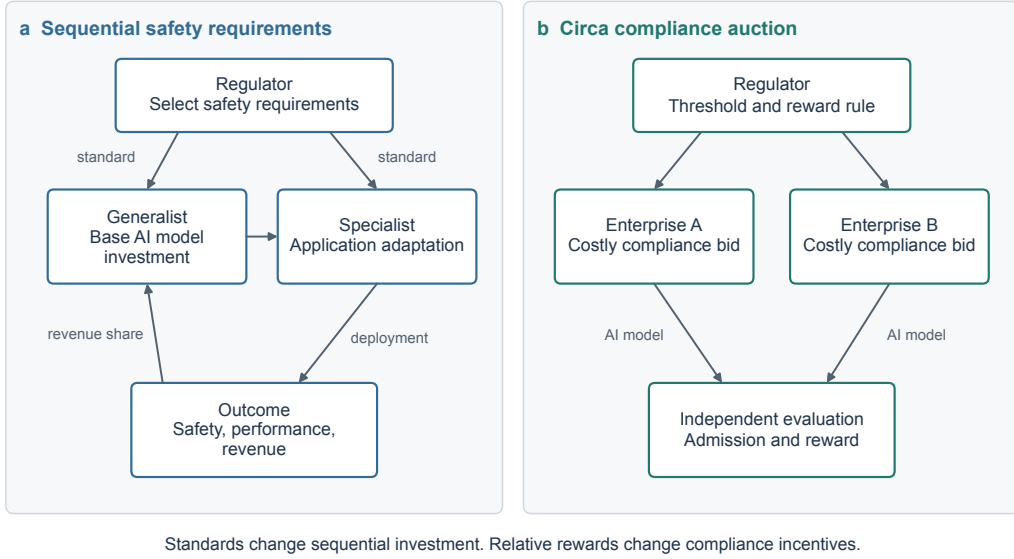


Figure 8: Game models for regulation. (a) Requirements affect generalist and specialist investment [25]. (b) Enterprises invest in compliance before assessment and reward allocation [59]. Arrows indicate decision and evaluation dependencies. The return arrow from the outcome to the generalist indicates revenue sharing. Game-specific functional assumptions are omitted.

and investment costs. Their subgame-perfect equilibrium analysis shows that, with quadratic costs and the specified safety–performance interaction, weak specialist-only requirements can reduce safety. Appropriate requirements on both actors can improve safety and performance through commitment. The result illustrates why requirements on different parts of the development chain need to be analyzed jointly.

Figure 8 illustrates the decision sequence. A regulator needs to consider which actor controls each safety activity and how a requirement changes subsequent investment. For example, requiring an application developer to correct an AI model limitation can affect the provider’s incentive to address that limitation during development.

A similar sequence of decisions arises during deployment. Ahmed et al. add human reliance to a regulator–organization hierarchy [111]. The regulator chooses audit and penalty intensity, the organization selects deployment and oversight settings, and an operator chooses reliance on the AI recommendation. Backward induction and synthetic simulations illustrate the interaction between these choices. The example broadens the set of modeled actors, while its simulated risk reduction remains specific to the chosen utility and risk functions.

Across these models, the allocation of safety duties matters because training safeguards, application testing, and deployment controls can substitute for or complement one another. This gives the contracts and responsibility analyses in Section 3 a regulatory role because they identify who controls an activity and who bears its cost.

4.2. Model Openness and Access

Openness decisions affect model access, adaptation costs, and revenue. These decisions involve several technical dimensions, including access to weights, training information, documentation, and deployment interfaces [55, 112]. A game model specifies which dimensions are governed and how they affect participants.

Qiu et al. model a generalist and specialist bargaining over revenue, with the generalist choosing openness and the specialist choosing adoption and improvement [55]. Payoffs combine shared revenue, costs, and penalties for release below an openness threshold. They derive closed-form best responses and subgame-perfect equilibria for fixed revenue sharing. Initial performance determines when thresholds or penalties change release. The scalar openness representation limits interpretation across technical access dimensions.

For governance, the game model connects release requirements to upstream and downstream incentives. Evaluation of an openness requirement considers adaptation costs, provider revenue, and operating obligations. Therefore, its evaluation includes participation and the allocation of costs across the development chain.

The scalar model leaves open the practical question of how access to weights, evaluation interfaces, and deployment environments should be allocated separately. Section 7.2 considers this extension alongside the evidence needed to authorize each permission.

4.3. Competitive Incentives for Compliance

A minimum requirement establishes an acceptance condition, but it does not necessarily reward further improvement [59]. Auction and contest models analyze how additional rewards affect compliance investment and participation.

Bornstein et al.'s Compliance-Incentivized Regulatory-Centered Auction (*Circa*) treats enterprise compliance investment as a costly all-pay bid [59]. The regulator sets an acceptance threshold and rewards relative compliance through pairwise comparisons among participating enterprises. Enterprise payoff combines market access and comparison rewards, minus investment cost. Equilibrium strategies and computational evaluation relate this assessment rule to participation and investment beyond the minimum requirement. The baseline analysis assumes that enterprise and regulator evaluation data follow the same distribution and that enterprises share a strictly increasing compliance-cost function.

Figure 9 illustrates how a comparative reward changes enterprises' incentives from meeting the minimum requirement to investing in enhanced compliance.

This approach makes the reward structure part of regulatory design. Applying it requires a measurable criterion, a credible assessment procedure, and resources for rewards. A higher compliance score supports governance when it measures a relevant system property. Therefore, the assessment procedure is part of the incentive problem.

Further analysis can examine heterogeneous supplier costs, as modeled in federated-learning contracts [95], payment budgets, as in delegated classification [24], and multiple criteria, as in regulation games for fairness and privacy [113]. It also needs to examine strategic optimization of the score. A provider can improve a measured criterion while leaving other risks unchanged. Connecting the auction to the auditing models in Section 5 permits joint analysis of investment incentives and measurement reliability.

4.4. Regulatory Objectives, Policy Mixes, and Privacy

Regulatory design also concerns the information and objectives of the regulator [60, 113]. The models below connect regulatory objectives and policy instruments to investment, assessment, and disclosure.

Chan et al. model a representative domestic AI firm facing foreign competition [114]. Regulatory stringency is the probability that exploitative practices are revealed to consumers. Foreign competition enters as a parameter rather than a strategic player. When the regulator maximizes actual consumer welfare, their equilibrium comparisons favor stringent regulation at high or low levels of foreign competition and lighter intervention at intermediate levels. Other regulatory objectives can produce different choices. This non-monotonic result links regulatory effects to the level of competition. This emphasis on market competition complements the development-chain analysis of Laufer et al., which represents sequential safety investment.

Cha uses a two-stage game in which a policymaker selects direct regulation and incentives before firms select compliance and innovation investment [60]. The policy objective accounts for sector risk and innovation. Firms balance benefits against compliance and investment costs. Optimization and simulations compare policy mixes. In the numerical scenarios, an incentive-centered policy is preferred with high risk, high compliance costs, and a strong innovation effect. A mixed policy is preferred with high risk, low costs, and a moderate innovation effect. Regulation and incentives can be substitutes or complements. These comparisons motivate attention to all three sector characteristics rather than a policy prescription based on risk alone.

Privacy protection illustrates how these policy instruments affect investment along the application chain. Lv et al. use an evolutionary game involving large-model providers, application providers, and privacy regulators [56]. Each provider chooses whether to invest in protection, and the regulator chooses strict or lax regulation. Payoffs include profit, protection costs, free-riding benefits, fines, subsidies, and regulatory reputation. Stability analysis identifies parameter-dependent outcomes, including strict regulation without investment when protection returns remain too low. The game model thus shows that supervision alone need not induce protection. Yao et al. instead model a government, a GenAI company, and users [57]. Their strategies concern supervision, protection, and participation. The payoffs include costs, reputation, free-riding benefits, subsidies, rewards, and fines. Replicator dynamics and system-dynamics simulations show how these parameters change convergence. Moderate subsidies and penalties support protection in their simulations, with subsidies more effective under the studied settings. The two studies connect upstream technical decisions to different downstream incentives.

(a) Game formulation



M: minimum compliance; cost = 1

E: enhanced compliance; cost = 2

Both levels meet the acceptance threshold. Deployment benefit = 4 for each enterprise.

Payoff = deployment benefit + expected reward - investment cost

Each enterprise bears its investment cost whether or not it wins the reward.

(b) Minimum requirement only

No additional reward

		Enterprise B	
		M	E
Enterprise A	M	(3, 3) Nash equilibrium	(3, 2)
	E	(2, 3)	(2, 2)

Expected payoffs: (Enterprise A, Enterprise B)

Minimum compliance is strictly dominant.

Against M: choosing M gives $3 > 2$.

Against E: choosing M gives $3 > 2$.

Best-response comparisons apply to both enterprises.

(c) Comparative compliance reward

Prize = 4 for higher compliance. A tie gives each a 50% chance.

		Enterprise B	
		M	E
Enterprise A	M	(5, 5)	(3, 6)
	E	(6, 3)	(4, 4) Nash equilibrium

Expected payoffs: (Enterprise A, Enterprise B)

Enhanced compliance is strictly dominant.

Against M: choosing E gives $6 > 5$.

Against E: choosing E gives $4 > 3$.

Best-response comparisons apply to both enterprises.

Governance role: comparative rewards encourage compliance beyond the minimum requirement.

Figure 9: An illustrative compliance-investment game inspired by Circa [59]. (a) Two enterprises choose minimum compliance (M), with investment cost 1, or enhanced compliance (E), with cost 2. Both strategies meet the acceptance threshold and give deployment benefit 4. Each enterprise pays its investment cost regardless of whether it receives a reward. Its payoff is deployment benefit plus expected reward minus investment cost. (b) Without an additional reward, M gives $4 - 1 = 3$ and E gives $4 - 2 = 2$, regardless of the competitor's strategy. Thus, M is strictly dominant and (M, M) is the unique Nash equilibrium. (c) A prize of 4 is awarded to the enterprise with higher compliance. A tie is resolved with equal probability, giving each an expected reward of $0.5 \times 4 = 2$. For an E–M pair, payoffs are $(4 + 4 - 2, 4 - 1) = (6, 3)$. Mutual M gives each $4 + 2 - 1 = 5$, while mutual E gives each $4 + 2 - 2 = 4$. Against M, choosing E gives 6 instead of 5. Against E, choosing E gives 4 instead of 3. Therefore, E is strictly dominant and (E, E) becomes the unique Nash equilibrium. Payoff pairs list enterprise A first and B second. The two compliance levels, numerical parameters, and tie-breaking rule are illustrative assumptions. Participation and accurate assessment are assumed, and the regulatory rule is fixed. This discrete game illustrates the incentive mechanism and does not reproduce the full model's equilibrium analysis.

Zhang et al. compare self-assessment and independent assessment as hypothetical approaches to foundation-model conformity assessment [115]. Their evolutionary analysis includes market and innovation-ecosystem effects. Its main finding is that ex ante assessment alone may be insufficient because market incentives and subsequent surveillance affect trustworthiness. Where the privacy models examine incentives to invest in protection, this study examines who assesses whether requirements are met. The proposed assessment arrangements connect system evaluation to subsequent monitoring.

Disclosure adds the question of what information firms must provide to consumers. For explainability, Mohammadi et al. study a regulator and competing firms serving heterogeneous consumers [116]. Firms choose aspects of AI model performance, explanation, and pricing to maximize profit. The published model evaluates regulation through consumer surplus. Equilibrium comparisons show that explanation mandates can reduce welfare in some parameter regimes. The authors also distinguish equal explanation levels (vertical XAI fairness) from equal explanation utility (horizontal XAI fairness). A mandate can ensure the former without ensuring the latter across all parameter regimes. Neither criterion is a general guarantee of fair AI decisions. The technical lesson is to represent the cost, availability, and use of an explanation. Section 5.2 examines the technical content and reliability of such explanations. These issues complement

the market model because disclosure policy also needs to specify what an explanation establishes and how users can check it.

4.5. Algorithmic Competition and Collusion

Repeated games represent market decisions made by interacting algorithms over time. Algorithms choose prices or output quantities, observe market outcomes, and update their policies [51–53]. The resulting behavior depends on learning rules, available information, and the market setting.

Calvano et al. study Q-learning algorithms in repeated price competition [51]. Their computational experiments show that the algorithms can maintain supracompetitive prices without direct communication. Deviations trigger a period of lower prices, followed by a return toward the previous outcome. The study demonstrates that learning dynamics can support collusive behavior within the modeled market.

Several studies refine the algorithmic-collusion problem. Klein examines sequential pricing with Q-learning [53]. Learning firms choose prices to maximize discounted profit. Simulations identify supracompetitive outcomes and asymmetric pricing behavior. Hansen et al. study independent bandit learners that treat the market as a monopoly and omit rivals from their demand model [54]. Informative feedback can correlate their price experiments, bias estimated price sensitivity, and sustain prices near the joint-monopoly level. Calvano et al. instead examine Q-learning in Cournot quantity competition with imperfect monitoring [52]. Firms observe past market prices but not rivals' outputs, so demand shocks can therefore trigger punishment even without a deviation. These studies show why information, timing, and exploration belong in the governance model. A simulated price level alone does not identify its mechanism.

These findings create a technical governance task of evaluating an algorithm in interaction with other adaptive algorithms. Testing against a fixed environment can overlook strategies that emerge during repeated competition. Evaluation needs to vary competitors, information access, learning horizons, and market conditions. This connects assessment before release to monitoring competitive behavior during operation.

Cooperation among AI agents can improve their shared task while imposing costs on other participants. Therefore, the governance objective must specify whose outcomes are included. This issue connects market competition to the coordination mechanisms in Section 6.1. Section 5.3 examines statistical tests of algorithmic non-collusion.

Table 6 compares the regulatory and market game models. In contrast to the safety-race models, these studies make the allocation of market access, transfers, or pricing opportunities part of the strategic setting.

4.6. Summary and Lessons Learned

Regulatory effects depend on responses throughout the development chain and market. Safety and explainability models show why a stronger requirement can have an adverse effect when it changes upstream investment or participation. Privacy and compliance models add the costs of protection and assessment to this comparison. Repeated pricing studies further show that policy evaluation must examine behavior learned through interaction, alongside equilibrium comparisons of fixed rules.

5. Auditing, Disclosure, and Enforcement

Auditing produces evidence, disclosure determines who receives it, and enforcement connects it to corrective action and deployment decisions [16, 17, 117]. Game models address the incentives to produce evidence, manipulate assessment, or evade controls. Cooperative explanations and strategic verification support evidence interpretation [20, 21]. The analysis moves from incentives to produce reliable evidence, through its interpretation and use in strategic assessment, to intervention during deployment. Together, the applications touch all ten processes in Table 2. Table 8 compares game models for auditing, disclosure, enforcement, and deployment control.

5.1. Incentives for Auditors and Oversight Providers

Oversight requires incentives for costly assessment, a policy for allocating inspections, and evidence that can support enforcement. In regulatory markets, firms choose development behavior and oversight providers invest in detection capability, with payoffs combining returns, costs, and contingent payments [27]. *Bounty Incentives* reward detected violations but can weaken as firms become safer. *Vigilant Incentives* permit withdrawal of rewards following evidence of missed wrongdoing and can sustain investment under suitable conditions. Equilibrium and evolutionary analyses connect oversight effort to both payment rules and later observations.

Adding users introduces trust into this interaction. An evolutionary model lets creators choose safe development, regulators choose effective oversight, and users decide whether to trust [118]. Development benefits, risk, costs,

Table 6

Comparison of regulation and market game models.

Study and task	Game details	Governance processes and implications
Laufer et al. [25]: safety standards	Type: sequential game. Players: regulator, generalist, specialist. Strategies: requirements, safety and performance investment. Payoff: development returns net of costs, with safety outcomes evaluated by the regulator. Solution: equilibrium investment and comparative statics.	P1, P4, P6. Show why a downstream requirement can change upstream incentives and reduce safety if poorly designed.
Qiu et al. [55]: model openness	Type: bargaining and sequential game. Players: generalist and specialist under regulation. Strategies: revenue sharing, openness, adoption and improvement. Payoff: revenues less adaptation costs and regulatory penalties. Solution: best responses and subgame-perfect equilibrium for fixed revenue sharing.	P4, P6. Relate release requirements to adoption, adaptation, and the allocation of development benefits.
Bornstein et al. [59]: compliance competition	Type: all-pay auction mechanism. Players: regulator and enterprises. Strategies: threshold/reward rule and costly compliance investment. Payoff: compliance-dependent rewards net of investment. Solution: equilibrium bidding/investment and computational evaluation.	P2, P5, P6. Reward compliance improvements beyond a minimum acceptance threshold.
Cha [60]: policy mix	Type: two-stage policy game. Players: policymaker and AI firms. Strategies: regulation and incentives, compliance and innovation. Payoff: sector-risk and innovation objectives, firm returns net of costs. Solution: sequential optimization and numerical policy comparisons.	P1, P2, P6. Identify cost- and risk-dependent combinations of regulation and incentives.
Lv et al. [56] and Yao et al. [57]: privacy	Type: evolutionary games. Players: providers, application providers, regulators (Lv), government, company, users (Yao). Strategies: protection, supervision, and participation. Payoff: benefits, costs, reputation, fines, and subsidies. Solution: fixed-point stability and simulated strategy evolution.	P3, P4, P7. Explain how privacy protection depends on incentives at several points in the service chain.
Mohammadi et al. [116]: explainability	Type: regulated competition. Players: regulator, firms, heterogeneous consumers. Strategies: regulation, performance/explanation, pricing, adoption. Payoff: consumer surplus, firm profit, consumer utility. Solution: equilibrium comparisons across disclosure requirements.	P5, P6, P8. Compare welfare effects; distinguish vertical and horizontal XAI fairness.
Calvano et al. [51, 52] and Klein [53] and Hansen et al. [54]: algorithmic competition	Type: repeated price or quantity competition with learning. Players: algorithms/firms. Strategies: prices or outputs and learned response policies. Payoff: profit from repeated sales. Solution: computational learning outcomes and analytical comparisons.	P5, P7. Identify collusion risks and the roles of monitoring, timing, and information in evaluation.
Ahmed et al. [111]: human–AI oversight	Type: hybrid principal-agent and Stackelberg model. Players: regulator, organization, human operator. Strategies: policy/audit/penalty, risk/explanation/oversight, reliance. Payoff: risk and decision quality, deployment returns, reliance benefits, net of costs and penalties. Solution: backward induction and synthetic simulations.	P1, P3, P6, P7, P8. Illustrate the dependence of human reliance on enforcement and system configuration; evidence is simulation-based.

and rewards determine payoffs. Government rewards and trust conditioned on regulatory effectiveness support safer development within specified parameter ranges. These models motivate audit contracts tied to credible evidence about assessment quality, connecting oversight to delegated learning (Section 3.2).

The information available to the government constrains enforcement. In the regulatory-market model, auditor capability is not directly observed, but later evidence of missed wrongdoing can affect rewards [27]. This connects auditor incentives to the statistical feasibility of the evidence used for enforcement. Section 7.2 examines sampled reassessment as a possible way to address this problem.

Alongside these effort incentives, oversight requires a policy for allocating limited inspection resources. Kim et al. propose Stackelberg security games as a perspective on governance of language-model systems [119]. A defender commits scarce protective resources and an attacker responds to exposed opportunities. Utility combines attack success, damage, and defense cost. The proposed applications include threats across the AI model lifecycle. The perspective supports resource allocation for monitoring and auditing. A useful extension would combine this game model's resource-allocation problem with statistically valid audit tests and an explicit model of auditor reliability.

For evidence to guide market access, the assessment must also distinguish compliant from non-compliant providers. Singh et al. formalize this relationship through a mechanism-design framework [63]. Providers hold private information and decide whether to enter a regulated market under evidence-dependent incentives. The regulator seeks a *perfect market outcome*, retaining compliant providers and deterring non-compliant ones. Under the framework, this outcome is attainable exactly when the set of non-compliant evidence distributions is closed and convex. Mechanisms can then be constructed from hypothesis tests. Experiments examine spurious-feature and fairness requirements. The result connects deployment approval to statistically testable evidence.

5.2. Explainable AI as Governance Evidence

Explainable AI provides information that helps people understand a model's predictions or behavior. A local explanation concerns an individual prediction, while a global explanation summarizes behavior over a population [20, 34, 43, 120]. Game-theoretic methods allocate model outcomes to input features or analyze interactions between evidence providers and evaluators. These methods support evaluation, monitoring, disclosure, and incident review. Table 7 summarizes their governance uses.

Local prediction attribution starts with SHAP (SHapley Additive exPlanations) [20]. Features are players and feature subsets are coalitions. Model outputs under a specified missing-feature rule define coalition values. Features have no strategic actions or financial payoffs. Shapley values allocate the prediction difference from a reference through marginal contributions. For a fixed simplified-input mapping, the authors characterize the unique additive attribution satisfying local accuracy, missingness, and consistency. KernelSHAP estimates this allocation by weighted regression. For governance, an explanation record can document the inputs associated with a decision and support audit review.

For tree models, Tree SHAP solves these feature games efficiently [43]. It computes exact attributions in polynomial time for the specified value function. The authors demonstrate subgroup analysis and monitoring of a hospital model. This supports repeated checks of model behavior, provided that the feature rule and reference population remain appropriate.

Moving to population-level behavior, SAGE uses a cooperative predictive-power game [84]. Features form coalitions whose value is the reduction in expected prediction loss relative to the empty coalition. Shapley allocation measures global feature importance. The analysis establishes allocation properties and consistency as sampling budgets grow, assuming the correct conditional feature distribution and sufficient data. Replacing conditional sampling with marginal sampling can change the target allocation and its properties [84]. SAGE can identify inputs requiring data-quality checks and monitor population-level model dependence. It complements Tree SHAP by allocating predictive performance instead of an individual prediction. Related dependence issues also arise in local Shapley explanations [121].

Sundararajan and Najmi compare Baseline Shapley with Conditional Expectation Shapley, showing how the value function affects interpretation [122]. Both allocate marginal contributions, but their different coalition values can produce different explanations for the same model. This finding supports recording the value function and reference data when explanations are disclosed. Causal Shapley values instead define coalition values using interventions in causal graphs [123]. Causal Shapley values allocate prediction differences, separate direct and indirect effects, and admit an interventional formula for causal chain graphs. They can guide investigation of harmful dependencies when the causal assumptions are justified.

Joint dependencies motivate the Shapley–Taylor interaction index, which allocates coalition value to feature groups up to a selected order [124]. Its axiomatic characterization includes an interaction-distribution requirement. Faith-Shap uses weighted polynomial approximation of the coalition-value function [125]. Extended dummy, symmetry, linearity, and efficiency axioms identify a unique index within its faithful-interaction mechanism. These solutions expose joint dependencies that individual feature scores can miss. A proposed governance application selects input combinations for risk assessment and joint testing.

FastSHAP learns an explainer from the Shapley weighted least-squares objective [66]. Features remain the players and masked subsets define coalition values. A learned approximation returns the allocation in one forward pass. Covert et al. extend amortized explanation to vision transformers through image-patch coalitions and attention-masked outputs [126]. Their computational results support frequent model evaluation and monitoring, with checks of approximation error after model changes.

When the evidence provider is strategic, Prover-Verifier Games analyze checkable answers through an untrusted message-generating prover and a trusted answer-predicting verifier [21]. The verifier minimizes prediction loss, while the prover seeks acceptance of a designated answer, including incorrect cases. Considering a complete and

Table 7

Game-theoretic explanations and verification methods considered for governance. Feature inclusion defines coalitions in attribution games.

Study and task	Game details	Governance processes and implications
Lundberg and Lee [20] and Lundberg et al. [43]: local attribution	Type: cooperative feature game. Players: input features. Coalitions: subsets made available to the explained model. Value: prediction under the missing-feature rule. Solution: Shapley allocation with KernelSHAP regression and polynomial-time Tree SHAP for specified tree games.	P5, P7, P8. Explain individual predictions and support monitoring of model behavior.
Covert et al. [84]: global importance	Type: cooperative predictive-power game. Players: features. Coalitions: observed feature subsets. Value: expected loss reduction relative to the empty coalition. Solution: SAGE allocation with a sampling estimator, consistency with correct conditional sampling and increasing sample budgets.	P4, P5, P7. Prioritize checks of inputs that contribute to population-level predictive performance.
Heskes et al. [123]: causal attribution	Type: cooperative game with causal interventions. Players: features. Coalitions: intervention sets. Value: expected prediction under the causal graph. Solution: causal Shapley allocation, including direct and indirect effects, chain-graph identification formula.	P3, P9. Guide investigation of model dependencies when causal assumptions are justified.
Sundararajan et al. [124] and Tsai et al. [125]: interactions	Type: cooperative interaction attribution. Players: features, allocations also concern feature groups. Coalitions: jointly observed subsets. Value: model response to a subset. Solution: Shapley–Taylor or Faith-Shap under different axiomatic characterizations.	P3, P5, P8. Identify joint dependencies that may be hidden by individual feature scores.
Jethani et al. [66] and Covert et al. [126]: fast explanation	Type: amortized solution of cooperative attribution games. Players: features or image patches. Coalitions: masked subsets. Value: explained model output. Solution: learned approximation to Shapley allocation through weighted least squares, one explainer forward pass.	P5, P7. Reduce the cost of repeated explanations for high-dimensional AI models.
Anil et al. [21]: checkable answers	Type: non-cooperative prover–verifier game. Players: untrusted prover and trusted verifier. Strategies: messages and answer policies. Objectives: target-answer acceptance versus correct prediction. Solution: verifier-leading Stackelberg characterization under protocol-existence and timing assumptions, Nash equilibrium alone is insufficient.	P5, P8. Design evidence-checking rules that account for an adversarial information provider.
Irving et al. [128]: debate	Type: zero-sum debate game. Players: two debating agents. Strategies: competing arguments for a judge. Payoff: favorable judgment against the opponent. Solution: idealized optimal-play complexity analysis.	P5, P8. Motivate challenges to assessment claims before a governance review.

sound protocol and instances revealed after verifier commitment, the authors characterize verifier-leading Stackelberg equilibrium as necessary and sufficient for such a protocol. They also derive a necessary Nash-equilibrium condition. The sequential model offers a basis for evaluating evidence from a strategically motivated provider. An organizational application can specify who controls the prover and verifier and how their outputs affect review decisions. Explanation procedures also require checks against adversarial manipulation [127].

In AI safety via debate, two agents present competing arguments and seek a favorable judgment [128]. Their opposed objectives define a zero-sum debate game. Under idealized optimal-play assumptions, the computational argument explains how debate can help a limited judge evaluate a harder question. It does not establish truthful debate for trained LLMs. A proposed governance use is to challenge AI-generated assessment claims before review. This use requires checking the reliability of the trained debaters and judge.

The feature-attribution methods share marginal-contribution analysis with data valuation (Section 3.3), but allocate different quantities. Attribution evidence can inform incident analysis. Responsibility additionally requires decision rights and feasible alternatives (Section 3.4). Verification games examine the reliability of evidence supplied by strategic actors, connecting explanation to auditor incentives (Section 5.1).

5.3. Auditing Strategic Algorithmic Behavior

Auditing interactions among algorithms requires a testable behavioral criterion. A game-theoretic criterion for algorithmic non-collusion connects pricing traces to a statistical audit with characterized data requirements [64].

Its refinement uses *pessimistic calibrated regret*, the highest calibrated regret among outcomes compatible with the observations [65]. This permits tests without requiring positive probability for every price. Low best-in-hindsight regret alone is insufficient, and sufficiently inflated cost reports can make a firm appear compliant. The test concerns the observed run under specified cost and information assumptions. It does not certify the algorithm in every future setting. Logging, exploration, and access to costs therefore determine what the audit can establish. Extending these tests to other AI services requires explicit counterfactual actions and a corresponding behavioral criterion, connecting assessment to market rules (Section 4).

5.4. Strategic Classification and Fairness Constraints

An evaluation rule can change the behavior that it measures. Hardt et al. formulate strategic classification as a Stackelberg interaction [129]. A classifier is selected before individuals change their features in response. The classifier seeks accurate decisions. Individuals trade the benefit of a favorable classification against modification cost. The resulting optimization accounts for their best responses. Brückner and Scheffer use Stackelberg prediction games to model similar anticipation of data generators [130]. Dalvi et al. analyze classification against cost-sensitive adversarial modifications [131]. These models are transferable to compliance scoring when assessed providers can alter submitted evidence.

Ahmadi et al. develop the strategic perceptron for online classification with strategic responses [132]. The standard perceptron can make unbounded mistakes when agents manipulate otherwise separable inputs. Their modified algorithm achieves finite mistake bounds for ℓ_2 and weighted ℓ_1 manipulation costs, and an extension learns unknown costs. The result supports repeated assessment when evaluated actors adapt their reported features. Ghalme et al. analyze strategic classification when individuals have imperfect knowledge of the classifier [133]. Their results show that transparency affects manipulation and classification outcomes. Shavit et al. connect regression incentives to causal improvement, while Miller et al. link strategic classification to causal inference [134, 135]. These works explain why changing an observed feature and improving the underlying property are different governance outcomes.

Where the preceding models analyze responses to assessment rules, the following methods use games to compute classifiers subject to fairness constraints. Agarwal et al. reduce fair classification to cost-sensitive classification through a Lagrangian game [136]. The learner selects a classifier, while a dual player places weight on constraint violations. Approximate saddle-point solutions provide bounds on prediction error and fairness violations under the stated optimization assumptions. Kearns et al. use a learner–auditor game to address fairness across many structured subgroups [137]. The auditor searches for a violated subgroup constraint. No-regret interaction supports approximate solutions when the required oracles are available. Here the learner and auditor are computational roles within a constrained optimization procedure.

Yaghini et al. instead model builders and regulators with competing fairness and privacy objectives [113]. Regulators set penalties, builders select models, and payoffs balance performance, compliance, and participation. Their SpecGame formulation and ParetoPlay method search for mutually acceptable outcomes on a performance frontier. Unlike the learner–auditor reductions, this formulation assigns model choice and penalty decisions to institutional actors. Its equilibrium and computational claims should be interpreted within that builder–regulator setting.

5.5. Deployment Control and Human Intervention

Deployment control connects observations to intervention strategies. Control evaluations test protocols against intentional subversion, including concerns about deceptive behavior that persists through safety training [138, 139]. *AI-Control Games* model a protocol designer and adversary through a multi-objective, partially observable stochastic game [35]. Reductions to zero-sum games support protocol search across safety and usefulness objectives. This formalizes what a control evaluation assumes about adversarial behavior and available observations.

Human oversight also depends on agent incentives. Cooperative inverse reinforcement learning represents a human and robot with a shared reward initially known only to the human [40]. Actions both teach and affect outcomes in a cooperative partially observable game. The Off-Switch Game then identifies conditions under which uncertainty about that reward gives the robot an incentive to preserve human intervention [41]. These results support oversight design under their shared-objective assumptions, without guaranteeing shutdown acceptance by arbitrary deployed agents.

Containment models further specify the controlled resources. A sequential hypothetical AGI–defender interaction represents autonomy, resource acquisition, monitoring, and access restrictions along an *AGI Kill Chain* [140]. Benefits combine detection or evasion, longevity, and resource share. Stationary-point calculations and simulations explore these trade-offs, but first-order conditions alone establish neither optimal controls nor a joint equilibrium. Feasible

Table 8

Comparison of game models for AI auditing, disclosure, enforcement, and deployment control. Computational players are identified where applicable.

Study and task	Game details	Governance processes and implications
Bova et al. [27]: oversight incentives	Type: regulatory-market game. Players: firms and oversight providers. Strategies: safe behavior and oversight investment under payment rules. Payoff: development/oversight returns, effort costs, and contingent rewards. Solution: equilibrium and evolutionary analysis.	P7, P8. Explain why rewards for detected violations can weaken as compliance improves, compare vigilant incentives.
Hartline et al. [64, 65]: non-collusion audits	Type: repeated-game behavioral certification. Players: pricing firms, auditor evaluates traces. Strategies: price policies and deviations. Payoff: profit under feasible costs. Solution: calibrated-regret criteria and statistical tests, including pessimistic calibrated regret.	P5, P7, P8. Connect a competition requirement to observable data, uncertainty, and audit sample complexity.
Hardt et al. [129] and Ghalme et al. [133]: strategic assessment	Type: Stackelberg classification. Players: classifier designer and assessed individuals. Strategies: decision rule and feature manipulation. Payoff: accuracy, classification benefit less modification cost. Solution: optimization against best responses under the modeled information structure.	P2, P5, P8. Anticipate responses to compliance scores and examine effects of transparency.
Agarwal et al. [136] and Kearns et al. [137]: fairness	Type: zero-sum/Lagrangian learning game. Players: computational learner and dual player/auditor. Strategies: classifiers and constraint weights or subgroup tests. Payoff: error combined with fairness violations. Solution: approximate saddle points using optimization oracles and no-regret learning.	P3, P5. Turn specified fairness requirements into computable training and auditing procedures.
Yaghini et al. [113]: competing requirements	Type: builder-regulator game. Players: model builder and regulators. Strategies: model selection and penalties. Payoff: performance, fairness/privacy compliance, and participation. Solution: SpecGame analysis and ParetoPlay search on a performance frontier.	P1, P5, P6. Analyze trade-offs between several governance requirements and builder participation.
Griffin et al. [35]: AI control	Type: multi-objective partially observable stochastic game. Players: protocol designer and adversary. Strategies: control protocol and adversarial behavior. Payoff: safety and usefulness objectives. Solution: reductions to zero-sum games and protocol search.	P5, P6, P10. Formalize deployment-control trade-offs and assumptions behind adversarial evaluation.
Ahmadi et al. [132]: online assessment	Type: sequential strategic classification. Players: classifier and arriving individuals. Strategies: classification rule and feature modification. Payoff: prediction accuracy and positive classification less manipulation cost. Solution: finite mistake bounds for a strategic perceptron for the analyzed manipulation costs.	P5, P7. Support repeated assessment when evaluated actors change their submitted features.
Singh et al. [63]: incentive-aware regulation	Type: mechanism design under statistical uncertainty. Players: regulator and privately informed providers. Strategies: evidence-dependent regulation and market participation. Objectives: retain compliant providers and deter non-compliant entry while providers seek favorable market outcomes. Solution: credal-set characterization and hypothesis-test constructions.	P5, P6, P8. Link deployment approval to statistical evidence and strategic participation.
McIntosh et al. [140]: AGI containment	Type: sequential attacker-defender model. Players: hypothetical AGI and human defenders. Strategies: autonomy/resource acquisition and containment. Payoff: detection/evasion, longevity, resource share. Solution: stationary-point analysis and simulations; optimal controls and a joint equilibrium are not established.	P6, P7, P9. Identify infrastructure access, monitoring, and response targets for containment.
Hadfield-Menell et al. [40, 41]: human oversight	Type: cooperative partially observable game, Off-Switch Game. Players: human and robot. Strategies: teaching, action, and intervention decisions. Payoff: shared human objective, initially uncertain to the robot. Solution: joint-policy reasoning and incentive conditions for preserving human intervention.	P1, P6, P10. Explain how uncertainty about objectives affects incentives to accept oversight or shutdown.

action bounds, second-order conditions, and joint strategic responses require further analysis before the results can justify containment decisions. Control records can inform responsibility analysis (Section 3.4).

5.6. Summary and Lessons Learned

Reliable oversight requires both incentives to produce evidence and a valid rule for interpreting it. Feature attribution describes model dependence, verification games examine strategic claims, and behavioral audits test actions against an explicit criterion. Deployment control then determines what intervention follows from an observation. Combining these methods requires agreement on the claim being assessed, the information available, and the authority to act on the result.

6. Coordination among Organizations, Communities, and Jurisdictions

Coordination games connect individual contributions to shared evaluation, platform oversight, and common policies. These activities require participants to bear costs whose benefits may be distributed widely [141]. The models cover policy, evaluation, deployment, monitoring, disclosure, and incident response, as mapped in Table 2. Table 9 compares the institutional models.

6.1. Shared Safety Investment and Collective Action

Shared evaluation creates a contribution problem. Voluntary participation also makes coalition stability important, as shown by model-sharing games for federated learning [142, 143]. In a stochastic evolutionary public-goods game, developers choose costly ethical cooperation or defection, with development benefits exposed to collective failure [68]. Smaller groups, lower costs, and greater perceived failure risk favor cooperation. A proposed governance extension would include contribution quality and verification costs, connecting collective investment to valuation and auditing (Sections 3.3 and 5).

6.2. Coordination of Platform Oversight

Platform oversight adds interactions among providers, users, and government. A prospect-theoretic evolutionary game models active or passive moderation, user participation, and stringent or lenient supervision [58]. Perceived payoffs include benefits, costs, externalities, rewards, penalties, and reputation losses. Replicator dynamics and Jacobian eigenvalues characterize the local stability of eight pure-strategy states. Active governance, user participation, and stringent supervision are locally stable when their three eigenvalues are negative.

In the reported simulations, platform penalties accelerate all three actors' responses, while user rewards mainly accelerate participation. These findings connect moderation, reporting incentives, and enforcement. They extend the privacy games in Section 4.4 by representing probability weighting and loss aversion. The stability result is local, so the starting distribution of strategies remains relevant to implementation.

6.3. International Coordination and Regulatory Differences

International agreements require both coordination incentives and compatible regulatory interests. Qualitative policy accounts emphasize credible commitments and shared research [144]. Formal models help specify the conditions under which these arrangements can persist. A Stag Hunt represents a higher-payoff joint action alongside a safer individual alternative [145]. Woods uses illustrative Q-learning simulations to examine cooperation under assigned rewards and discount factors, and interprets these patterns through perceived risk and institutional legitimacy [62]. The study motivates attention to the acceptance of common policies, but its simulations do not establish an empirical effect of legitimacy on international cooperation.

Technological differences can nevertheless produce divergent regulatory preferences. A two-stage model lets a regulator choose stringency and consumers allocate use between domestic and foreign AI services [61]. The regulator pursues digital sovereignty, while consumers seek service benefits. Sequential optimization shows that stringent regulation supports the sovereignty objective only under particular technological and regulatory conditions. A proposed repeated-game extension would let heterogeneous jurisdictions share incidents, recognize evaluations, and fund common assessment capacity. Contribution costs, access benefits, and expected future participation would determine whether cooperation persists. Section 7.1 relates these requirements to measurement and long-term commitments.

6.4. Summary and Lessons Learned

Shared safety investment depends on how costs and benefits are distributed among participants. Platform models add reporting incentives and perceived losses. International governance discussions highlight legitimacy, while regulatory models show how technological position affects policy preferences. A workable arrangement must make participation worthwhile and its contributions verifiable. Once those institutional choices and the intended outcome

Table 9

Comparison of institutional coordination models.

Study and task	Game details	Governance processes and implications
LaCroix and Mohseni [68]: ethical development	Type: stochastic evolutionary public-goods game. Players: developers. Strategies: costly ethical cooperation or defection. Payoff: contribution cost and benefits exposed to collective failure. Solution: strategy evolution and cooperation conditions.	P1, P5. Identify how group size, contribution cost, and perceived failure risk affect cooperation.
Lei et al. [58]: disinformation governance	Type: evolutionary game with prospect theory. Players: platform, users, government. Strategies: active/passive governance, participation, stringent/lenient supervision. Payoff: perceived benefits, costs, externalities, rewards, penalties, reputation. Solution: replicator dynamics and local stability of pure-strategy states.	P1, P7, P9. Coordinate moderation, user reporting, and enforcement while accounting for loss aversion.
Woods [62]: international coordination	Type: Stag Hunt-inspired learning model. Players: modeled governance participants. Strategies: cooperation or non-cooperation. Payoff: assigned coordination rewards. Solution: illustrative Q-learning simulations.	P1, P8. Explore cooperation under reward and discount assumptions, with legitimacy discussed as a governance interpretation.
Papyshev and Chan [61]: regulatory differences	Type: two-stage regulatory game. Players: regulator and consumers of domestic/foreign AI services. Strategies: regulatory stringency and service allocation. Payoff: digital-sovereignty objective and consumer utility. Solution: sequential optimization under technological differences.	P1, P6. Explain why different technological positions produce different regulatory preferences.

are explicit, computational methods for learning transfers or reconciling preferences can support implementation, as discussed in Section 7.4.

7. Synthesis and Research Directions

The models examined here connect governance decisions to incentives, information, and computational methods. Further research should address ethical objectives, adaptation of governance mechanisms, and models that remain useful at realistic scale. Figure 1 summarizes the shared process coverage. Building on the strategic problems introduced in Section 1, the task comparisons reveal dependencies between incentives, evidence, and decisions over time. A payment rule needs a performance measure, an audit needs incentives for reliable assessment, and a deployment rule needs an actor able to enforce it. We examine these dependencies before developing the research directions.

7.1. Dependencies across Tasks, Governance Levels, and Horizons

Across the four tasks, one dependency concerns measurement. Delegated-learning contracts rely on a statistical relationship between effort and observed performance, while compliance auctions reward an assessed score [24, 59]. Auditing supplies those observations, but its own quality depends on the evaluator's incentives [27]. Strategic classification adds the possibility that participants change reported features in response to the rule [129]. A mechanism that rewards the wrong observable can therefore remain incentive compatible while failing to induce the intended safeguard. Measurement validity and incentive analysis need to be evaluated together [146].

Another dependency concerns commitment and authority. A regulator's requirements alter later investment by providers and specialists, and liability rules alter adoption and supervision [25, 50]. These are examples of the interactions across governance levels in Figure 4. A national rule changes organizational choices, which in turn change the options available to a user or operator. A model intended to guide policy should preserve this chain of responses. Aggregating all private actors into one representative provider can hide the actor whose incentive undermines implementation.

A further dependency concerns time. Operational monitoring generates information for later contract renewal, continued access, and investment. Conversely, expectations about those later decisions shape present effort. Figure 5 helps distinguish the speed at which a control responds from the horizon over which its incentives operate. A rapid audit update may improve detection but weaken investment if participants cannot predict future requirements. A long-term commitment may encourage investment but become ineffective when capabilities or risks change. The resulting research question is how to preserve credible commitments while allowing justified revision.

Coordination combines these dependencies. Shared evaluations and incident reporting can reduce duplicated effort, yet participants need credible quality standards and reasons to disclose adverse information. International cooperation adds differences in technological position and regulatory objectives [61, 62]. A repeated-game analysis of common assessment capacity should therefore include who pays, who verifies contributions, which access benefits depend on participation, and how cooperation changes after a reported failure.

7.2. Combining Incentives, Evidence, and Access Rules

Combining incentives, evidence, and access rules changes the game. Insurance alters the losses that motivate precautions, an audit affects access as well as payment, and a disclosure rule changes what another actor can infer. Three extensions follow directly from the task analyses.

First, an insurance model can extend the manufacturer–user liability game in Section 3.5. An insurer chooses premiums, coverage, and evidence requirements, a provider chooses safeguards, and users choose adoption and supervision. Returns then include premium costs, retained losses, and claims payments. The useful question is whether audit-contingent terms preserve incentives for precautions while making participation worthwhile. Shared dependence on a foundation model also requires correlated losses, since pooling many applications may leave common failure modes largely undiversified.

Second, model access can be represented as a vector of permissions for weights, evaluation interfaces, and deployment environments, extending the scalar openness decision in Section 4.2. The authority allocates permissions using intended use and assessment results. Providers and downstream developers respond through disclosure, adaptation, and deployment choices. Analysis should identify how each permission changes benefits, exposure to misuse, and the ability to verify compliance. Evaluation can then compare access decisions, participation, and risk across alternative permission rules.

Third, an audit contract can combine routine compensation with independent reassessment of sampled cases. The principal chooses payment and sampling rules, while the auditor chooses effort and reporting behavior. Reassessment cost, detection error, and dispute resolution determine whether a missed violation provides a useful payment signal. This formulation joins the effort-incentive problem in regulatory markets to the statistical evidence problem in incentive-aware regulation [27, 63]. It should be tested against collusion, selective case reporting, and changes in the underlying case distribution.

These are research formulations requiring analysis and validation. Their common contribution is to make the interaction between mechanisms explicit. A useful test must produce evidence that an enforceable rule can use, and the rule must leave participants an incentive to produce that evidence accurately.

7.3. Ethical Objectives and Representation

AI ethics examines the values and responsibilities that should guide AI development and use, including fairness, privacy, transparency, and the prevention of harm [9, 11]. AI governance translates these objectives into responsibilities, incentives, assessment procedures, and deployment decisions. The modeling challenge is to connect ethical requirements to strategic behavior without assuming that every actor shares the same values or bears the same costs.

Existing game models provide starting points. LaCroix and Mohseni show how costly ethical development can create a collective-action problem [68]. Agarwal et al. and Kearns et al. formulate fairness requirements through learner–auditor games that support approximate constrained solutions [136, 137]. Yaghini et al. model competing fairness and privacy requirements through interactions between builders and regulators [113]. These examples connect ethical objectives to participation incentives, computational constraints, and institutional trade-offs. These approaches differ in who chooses the requirement and who can challenge it.

Future models should represent affected groups as participants or give their interests explicit weight in governance objectives. Candidate mechanisms include contribution-dependent incentives for ethical development and constrained bargaining over acceptable model choices. Challenges include conflicting definitions of fairness, unequal bargaining power, and ethical requirements that cannot be exchanged for higher aggregate utility. Evaluation should measure subgroup outcomes [147], privacy loss, participation, and gains from strategic deviation. An equilibrium establishes stability under the modeled incentives. Its ethical acceptability still depends on the chosen requirements and affected parties [148].

7.4. Adaptive Governance and Computational Mechanism Design

Adaptive governance updates requirements, audit allocations, and deployment controls as systems, risks, and participant behavior change. Automated mechanism design searches computationally for rules that meet a governance

objective while accounting for strategic responses. Their potential benefits include faster reassessment, better use of limited oversight resources, and incentives that respond to changing costs and capabilities.

When participants adapt, incentives must account for their learning. *Learning to Incentivize Other Learning Agents* learns reward transfers by differentiating through recipients' policy updates [67]. Agents choose action and incentive policies in general-sum Markov games, with environmental rewards and transfer costs determining returns. Experiments show improvements in collective return and division of labor. Applying this mechanism to shared evaluation would require explicit safety objectives, budget constraints, and checks against manipulation. Learned auctions provide a further approach to searching allocation and payment rules under incentive constraints [98]. Applying revenue-maximizing auctions to evaluation procurement requires a procurement objective and renewed incentive analysis.

Alongside changes to institutional rules, computational methods can support model revision using participants' preferences. *Nash Learning from Human Feedback* represents response policies as players, with payoffs based on pairwise preferences and reference-policy regularization [44]. For tabular policies, the last iterate of Nash-MD converges to the regularized Nash equilibrium. *Direct Nash Optimization* uses preference-based self-improvement and batched policy updates [45]. These games support preference-sensitive evaluation and revision, including cyclic preferences. Their players are computational policies. Governance still determines whose preferences enter the game and which safety requirements remain fixed constraints.

Computational convergence does not resolve disagreements about governance objectives. The main challenges are nonstationary responses, delayed evidence, strategic manipulation of the learning signal, and instability when rules change too quickly. Possible solutions include separating provider and regulator update timescales, restricting policy changes to a tested rule set, and validating candidate mechanisms against held-out participant strategies. A governance mechanism should preserve participation and budget constraints and enforce minimum safety requirements during learning. Performance should be measured through risk reduction, oversight cost, adjustment speed, and robustness to strategic responses.

7.5. Scalability and Empirical Validation

Practical governance involves many developers, suppliers, auditors, users, and AI agents. Each can choose from a large action space, while payoffs combine revenue, computation costs, delay, privacy, safety, and external harms. The resulting game can be difficult to specify and solve. Population, network, and hierarchical approximations offer ways to reduce computation. Their evaluation should check whether aggregation preserves influential actors, rare failures, and the incentives that determine governance outcomes.

Real-world data are needed to estimate effort costs, detection probabilities, loss distributions, and responses to incentives. Empirical studies of machine behavior [149], procurement records, audit logs, incident reports [150], and controlled experiments can inform these quantities. Missing reports and strategic disclosure create selection bias. Models should therefore report uncertainty, test sensitivity to poorly identified parameters, and validate predictions on later observations or different organizations. Evaluation should also compare decisions with simpler policy baselines. A small equilibrium error is useful only when the resulting governance decisions remain reliable.

Incomplete information concerns unknown types, objectives, or payoff parameters. Imperfect information concerns actions or histories that a participant cannot observe. Governance games often contain both. A regulator may not know a provider's safety costs or observe its development effort. Bayesian models, signaling and screening mechanisms, and partially observable stochastic games can represent these limits [35, 36, 79]. Research should develop scalable belief updates and robust mechanisms when prior distributions are uncertain. Tests should combine gains from deviation with detection accuracy, computational cost, safety failures, and participation under misspecified information.

Table 10 translates these issues into tractable research questions. Each row separates a candidate formulation from the evidence needed to judge it. This provides a route from theoretical or computational progress to an assessment of governance outcomes.

8. Conclusion

This perspective paper connects game models to four AI governance tasks, namely (i) incentives and responsibility, (ii) regulation and AI markets, (iii) auditing, disclosure, and enforcement, and (iv) coordination. The comparison identifies how players, information, payoffs, and solution assumptions determine a model's governance use. Cooperative explanations and interactive verification add tools for producing and checking evidence. Their application

Table 10

Open problems derived from the cross-task synthesis. Formulations are research proposals; the final column identifies evidence needed to evaluate them.

Open problem	Candidate formulation	Validation and outcomes
Audit quality under hidden effort	Principal-agent contract with sampled independent reassessment, noisy detection, and reporting choices.	Reassessment records and effort costs; detection error, reporting accuracy, audit cost, and gains from deviation.
Insurance and correlated AI failures	Sequential insurer–provider–user game with audit-contingent premiums and shared-model loss dependence.	Loss and safeguard data; precaution effort, retained losses, coverage, participation, and sensitivity to common shocks.
Permissions beyond scalar openness	Sequential allocation of weight, interface, and environment access with privately known uses.	Permission and evaluation records; access decisions, misuse exposure, downstream participation, and verification cost.
Ethical constraints and representation	Constrained bargaining or builder–regulator games with affected-group objectives and protected requirements.	Subgroup outcomes, privacy loss, participation, bargaining sensitivity, and constraint violations.
Stable adaptive oversight	Bilevel or stochastic mechanism design with delayed signals and separate update timescales.	Held-out strategic responses; risk, budget feasibility, adjustment speed, oscillation, and safety during adaptation.
Preference-sensitive model revision	Regularized preference games within externally set safety constraints.	Preference coverage and subgroup agreement; convergence error, model behavior, and retained safety properties.
Shared international assessment	Repeated contribution game with verification, access benefits, and heterogeneous outside options.	Contribution quality and cost; participation persistence, incident disclosure, withdrawal, and benefits by jurisdiction.
Scalable decisions under uncertainty	Network or population approximations with Bayesian or robust responses.	Held-out organizations and periods; computation time, calibration, rare-failure sensitivity, and decision quality against baselines.

requires explicit claims and validation conditions. Across the tasks, measurements must be informative about the outcome being rewarded, commitments must be credible to the actors they affect, and evaluation must account for strategic adaptation. These requirements connect the models to practical choices about contracts, deployment approval, access permissions, monitoring, and incident response. Several operational applications remain proposals that require empirical validation. Future work should connect game models to AI ethics, develop adaptive governance mechanisms, and scale analysis using real-world data and explicit information assumptions. These priorities connect computational methods to organizational decisions in computer science, business, and information systems.

CRediT authorship contribution statement

Dusit Niyato contributed to conceptualization, supervision, project administration, and writing through review and editing. **Yinqiu Liu** contributed to methodology, investigation, data curation, visualization, and writing of the original draft. **Wanpracha “Art” Chaovalitwongse** contributed to conceptualization and writing through review and editing. **Zehui Xiong** contributed to methodology, validation, and writing through review and editing. **Zhu Han** contributed to conceptualization, supervision, and writing through review and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

Generative AI tools were used to improve the language and readability of the manuscript. The authors take full responsibility for its content.

References

- [1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the opportunities and risks of foundation models, arXiv:2108.07258 [preprint] (2021). <https://arxiv.org/abs/2108.07258>.
- [2] Infocomm Media Development Authority, Model AI governance framework for agentic AI, Tech. rep., Infocomm Media Development Authority, version 1.5 (2026). <https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf> [accessed 23 September 2026].
- [3] E. Tabassi, Artificial intelligence risk management framework (AI RMF 1.0), Tech. Rep. NIST AI 100-1, National Institute of Standards and Technology (2023). <https://doi.org/10.6028/nist.ai.100-1>.
- [4] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall, K. Roberts, Artificial intelligence risk management framework: generative artificial intelligence profile, Tech. Rep. NIST AI 600-1, National Institute of Standards and Technology (2024). <https://doi.org/10.6028/NIST.AI.600-1>.
- [5] Microsoft, 2026 Responsible AI transparency report, Tech. rep., Microsoft (2026). <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Responsible-AI-Transparency-Report-2026.pdf> [accessed 23 September 2026].
- [6] European Commission, General-purpose AI code of practice now available (2025). <https://digital-strategy.ec.europa.eu/en/news/general-purpose-ai-code-practice-now-available> [accessed 23 September 2026].
- [7] A. Batool, D. Zowghi, M. Bano, AI governance: a systematic literature review, AI Ethics 5 (3) (2025) 3265–3279. <https://doi.org/10.1007/s43681-024-00653-w>.
- [8] E. Papagiannidis, P. Mikalef, K. Conboy, Responsible artificial intelligence governance: a review and research framework, J. Strateg. Inf. Syst. 34 (2) (2025) 101885. <https://doi.org/10.1016/j.jsis.2024.101885>.
- [9] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nat. Mach. Intell. 1 (9) (2019) 389–399. <https://doi.org/10.1038/s42256-019-0088-2>.
- [10] A. Reuel, B. Bucknall, S. Casper, T. Fist, L. Soder, O. Aarne, L. Hammond, L. Ibrahim, A. Chan, P. Wills, M. Anderljung, B. Garfinkel, L. Heim, A. Trask, G. Mukobi, R. Schaeffer, M. Baker, S. Hooker, I. Solaiman, A. S. Luccioni, N. Rajkumar, N. Moës, J. Ladish, D. Bau, P. Bricman, N. Guha, J. Newman, Y. Bengio, T. South, A. Pentland, S. Koyejo, M. J. Kochenderfer, R. Trager, Open problems in technical AI governance, Trans. Mach. Learn. Res. (2025). <https://openreview.net/forum?id=1nO4qFMiS0>.
- [11] B. Mittelstadt, Principles alone cannot guarantee ethical AI, Nat. Mach. Intell. 1 (11) (2019) 501–507. <https://doi.org/10.1038/s42256-019-0114-4>.
- [12] T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddharth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano, A. Dafoe, Model evaluation for extreme risks, arXiv:2305.15324 [preprint] (2023). <https://arxiv.org/abs/2305.15324>.
- [13] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, Trans. Mach. Learn. Res. (2023). <https://openreview.net/forum?id=iO4LZibEqW>.
- [14] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, J. Clark, Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned, arXiv:2209.07858 [preprint] (2022). <https://arxiv.org/abs/2209.07858>.

- [15] E. Perez, S. Ringer, K. Lukosiute, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. El Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, J. Kaplan, Discovering language model behaviors with model-written evaluations, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023, pp. 13387–13434. <https://doi.org/10.18653/v1/2023.findings-acl.847>.
- [16] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, 2020, pp. 33–44. <https://doi.org/10.1145/3351095.3372873>.
- [17] J. Mökander, J. Schuett, H. R. Kirk, L. Floridi, Auditing large language models: a three-layered approach, *AI Ethics* 4 (4) (2024) 1085–1115. <https://doi.org/10.1007/s43681-023-00289-2>.
- [18] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O’Keefe, M. Koren, T. Ryffel, J. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. O. hEigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, M. Anderljung, Toward trustworthy AI development: mechanisms for supporting verifiable claims, *arXiv:2004.07213 [preprint]* (2020). <https://arxiv.org/abs/2004.07213>.
- [19] S. Casper, C. Ezell, C. Siegmund, N. Kolt, T. L. Curtis, B. Bucknall, A. Haupt, K. Wei, J. Scheurer, M. Hobbhahn, L. Sharkey, S. Krishna, M. Von Hagen, S. Alberti, A. Chan, Q. Sun, M. Gerovitch, D. Bau, M. Tegmark, D. Krueger, D. Hadfield-Menell, Black-box access is insufficient for rigorous AI audits, in: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, ACM, 2024, pp. 2254–2272. <https://doi.org/10.1145/3630106.3659037>.
- [20] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
- [21] C. Anil, G. Zhang, Y. Wu, R. Grosse, Learning to give checkable answers with prover-verifier games, *arXiv:2108.12099 [preprint]* (2021). <https://arxiv.org/abs/2108.12099>.
- [22] S. Armstrong, N. Bostrom, C. Shulman, Racing to the precipice: a model of artificial intelligence development, *AI & Soc.* 31 (2) (2016) 201–206. <https://doi.org/10.1007/s00146-015-0590-y>.
- [23] T. A. Han, L. M. Pereira, F. C. Santos, T. Lenaerts, To regulate or not: a social dynamics analysis of an idealised AI race, *J. Artif. Intell. Res.* 69 (2020) 881–921. <https://doi.org/10.1613/jair.1.12225>.
- [24] E. Saig, I. Talgam-Cohen, N. Rosenfeld, Delegated classification, in: Advances in Neural Information Processing Systems, Vol. 36, Curran Associates, Inc., 2023, pp. 13200–13236. <https://doi.org/10.52202/075280-0579>.
- [25] B. Laufer, J. Kleinberg, H. Heidari, The backfiring effect of weak AI safety regulation, *Proc. Natl. Acad. Sci. U.S.A.* 123 (30) (2026) e2509768123. <https://doi.org/10.1073/pnas.2509768123>.
- [26] Z. Zhang, F. Bai, M. Wang, H. Ye, C. Ma, Y. Yang, Roadmap on incentive compatibility for AI alignment and governance in sociotechnical systems, in: Artificial General Intelligence: AGI 2025, Vol. 16058 of Lecture Notes in Computer Science, Springer, 2026, pp. 370–380. https://doi.org/10.1007/978-3-032-00800-8_33.
- [27] P. Bova, A. Di Stefano, T. A. Han, Both eyes open: Vigilant Incentives help auditors improve AI safety, *J. Phys. Complex.* 5 (2) (2024) 025009. <https://doi.org/10.1088/2632-072X/ad424c>.
- [28] W. Wei, L. Liu, Trustworthy distributed AI systems: robustness, privacy, and governance, *ACM Comput. Surv.* 57 (6) (2025) 1–42. <https://doi.org/10.1145/3645102>.
- [29] J. Ji, T. Qiu, B. Chen, J. Zhou, B. Zhang, D. Hong, H. Lou, K. Wang, Y. Duan, Z. He, L. Vierling, Z. Zhang, F. Zeng, J. Dai, X. Pan, H. Xu, A. O’Gara, K. Ng, B. Tse, J. Fu, S. Mcaleer, Y. Wang, M. Yang, Y. Liu, Y. Wang, S.-C. Zhu, Y. Guo, Y. Yang, W. Gao, AI alignment: a contemporary survey, *ACM Comput. Surv.* 58 (5) (2026) 1–38. <https://doi.org/10.1145/3770749>.
- [30] T. A. Han, Emergent behaviours in multi-agent systems with evolutionary game theory, *AI Commun.* 35 (4) (2022) 327–337. <https://doi.org/10.3233/AIC-220104>.
- [31] H. Sun, Y. Wu, Y. Cheng, X. Chu, Game theory meets large language models: a systematic survey, in: Proceedings of the 34th International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 10669–10677. <https://doi.org/10.24963/ijcai.2025/1184>.
- [32] H. Sun, Y. Wu, P. Wang, W. Chen, Y. Cheng, X. Deng, X. Chu, Game theory meets large language models: a systematic survey with taxonomy and new frontiers, *arXiv:2502.09053v2 [preprint]*, expanded version (2025). <https://arxiv.org/abs/2502.09053v2>.
- [33] Y. Cai, Z. Zhao, Z. Lu, I. Lee, W. E. Zhang, M. Xue, Y. Zhang, S. Pang, W. Xiang, AI alignment through a game-theoretic lens: a survey, *arXiv:2608.27910v2 [preprint]* (2026). <https://arxiv.org/abs/2608.27910v2>.
- [34] I. Covert, S. Lundberg, S.-I. Lee, Explaining by removing: a unified framework for model explanation, *J. Mach. Learn. Res.* 22 (209) (2021) 1–90. <https://jmlr.org/papers/v22/20-1316.html>.
- [35] C. Griffin, L. Thomson, B. Shlegeris, A. Abate, Games for AI control: models of safety evaluations of AI deployment protocols, *arXiv:2409.07985v2 [preprint]* (2026). <https://arxiv.org/abs/2409.07985v2>.
- [36] Y. Shoham, K. Leyton-Brown, Multiagent systems: algorithmic, game-theoretic, and logical foundations, Cambridge University Press, 2009. <https://www.masfoundations.org/>.

- [37] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, D. Dennison, Hidden technical debt in machine learning systems, in: *Advances in Neural Information Processing Systems*, Vol. 28, Curran Associates, Inc., 2015, pp. 2503–2511. <https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcdf2674f757a2463eba-Abstract.html>.
- [38] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. K. Paritosh, L. M. Aroyo, “Everyone wants to do the model work, not the data work”: data cascades in high-stakes AI, in: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, 2021, pp. 1–15. <https://doi.org/10.1145/3411764.3445518>.
- [39] A. Reuel, L. Soder, B. Bucknall, T. A. Undheim, Position: technical research and talent is needed for effective AI governance, in: *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 42543–42557. <https://proceedings.mlr.press/v235/reuel24a.html>.
- [40] D. Hadfield-Menell, A. Dragan, P. Abbeel, S. J. Russell, Cooperative inverse reinforcement learning, in: *Advances in Neural Information Processing Systems*, Vol. 29, Curran Associates, Inc., 2016. https://papers.nips.cc/paper_files/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html.
- [41] D. Hadfield-Menell, A. Dragan, P. Abbeel, S. Russell, The off-switch game, in: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, 2017, pp. 220–227. <https://doi.org/10.24963/ijcai.2017/32>.
- [42] Y. Cai, C. Daskalakis, C. Papadimitriou, Optimum statistical estimation with strategic data sources, in: *Proceedings of the 28th Conference on Learning Theory*, Vol. 40 of *Proceedings of Machine Learning Research*, PMLR, 2015, pp. 280–296. <https://proceedings.mlr.press/v40/Cai15.html>.
- [43] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, *Nat. Mach. Intell.* 2 (1) (2020) 56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
- [44] R. Munos, M. Valko, D. Calandriello, M. G. Azar, M. Rowland, Z. D. Guo, Y. Tang, M. Geist, T. Mesnard, C. Fiegel, A. Michi, M. Selvi, S. Girgin, N. Momchev, O. Bachem, D. J. Mankowitz, D. Precup, B. Piot, Nash learning from human feedback, in: *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 36743–36768. <https://proceedings.mlr.press/v235/munos24a.html>.
- [45] C. Rosset, C.-A. Cheng, A. Mitra, M. Santacroce, A. Awadallah, T. Xie, Direct Nash Optimization: teaching language models to self-improve with general preferences, *arXiv:2404.03715 [preprint]* (2024). <https://arxiv.org/abs/2404.03715>.
- [46] T. A. Han, T. Lenaerts, F. C. Santos, L. M. Pereira, Voluntary safety commitments provide an escape from over-regulation in AI development, *Technol. Soc.* 68 (2022) 101843. <https://doi.org/10.1016/j.techsoc.2021.101843>.
- [47] C. Baier, F. Funke, R. Majumdar, A game-theoretic account of responsibility allocation, in: *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 1773–1779. <https://doi.org/10.24963/ijcai.2021/244>.
- [48] C. Baier, F. Funke, R. Majumdar, Responsibility attribution in parameterized Markovian models, *Proc. AAAI Conf. Artif. Intell.* 35 (13) (2021) 11734–11743. <https://doi.org/10.1609/aaai.v35i13.17395>.
- [49] M. Karimian, S. M. Khan, Y. Lespérance, Causal, strategic, and combined responsibility attribution in Situation Calculus concurrent game structures, *Proc. AAAI Conf. Artif. Intell.* 40 (23) (2026) 19178–19188. <https://doi.org/10.1609/aaai.v40i23.38992>.
- [50] M. Obidzinski, Y. Oytana, Artificial intelligence, inattention and liability rules, *Int. Rev. Law Econ.* 79 (2024) 106211. <https://doi.org/10.1016/j.irle.2024.106211>.
- [51] E. Calvano, G. Calzolari, V. Denicolò, S. Pastorello, Artificial intelligence, algorithmic pricing, and collusion, *Am. Econ. Rev.* 110 (10) (2020) 3267–3297. <https://doi.org/10.1257/aer.20190623>.
- [52] E. Calvano, G. Calzolari, V. Denicolò, S. Pastorello, Algorithmic collusion with imperfect monitoring, *Int. J. Ind. Organ.* 79 (2021) 102712. <https://doi.org/10.1016/j.ijindorg.2021.102712>.
- [53] T. Klein, Autonomous algorithmic collusion: Q-learning under sequential pricing, *RAND J. Econ.* 52 (3) (2021) 538–558. <https://doi.org/10.1111/1756-2171.12383>.
- [54] K. T. Hansen, K. Misra, M. M. Pai, Frontiers: algorithmic collusion: supra-competitive prices via independent algorithms, *Mark. Sci.* 40 (1) (2021) 1–12. <https://doi.org/10.1287/mksc.2020.1276>.
- [55] T. Qiu, B. Laufer, J. Kleinberg, H. Heidari, Modeling the economic impacts of AI openness regulation, in: *Advances in Neural Information Processing Systems*, Vol. 38, Curran Associates, Inc., 2025, pp. 41647–41679. <https://doi.org/10.52202/085713-1246>.
- [56] Y. Lv, J. Yang, X. Sun, H. Wu, Evolutionary game analysis of stakeholder privacy management in the AIGC model, *Oper. Res. Perspect.* 14 (2025) 100327. <https://doi.org/10.1016/j.orp.2025.100327>.
- [57] L. Yao, Q. Zhang, S. Deng, A combined approach of evolutionary game and system dynamics for user privacy protection in human intelligence interaction, *Sci. Rep.* 15 (1) (2025) 13730. <https://doi.org/10.1038/s41598-025-98805-3>.
- [58] L. Lei, Y. Wu, S. Gao, Evolutionary game analysis of AI-generated disinformation governance on UGC platforms based on prospect theory, *Systems* 14 (4) (2026) 416. <https://doi.org/10.3390/systems14040416>.
- [59] M. Bornstein, Z. Che, S. Julapalli, A. Mohamed, A. S. Bedi, F. Huang, Advancing regulation in artificial intelligence: an auction-based approach, in: *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2026, pp. 467–496, earlier version: ICLR 2026 Workshop on AI for Mechanism Design and Strategic Decision Making. <https://doi.org/10.1145/3805689.3812397>.
- [60] S. Cha, Policy mix framework for AI regulation: a game-theoretic analysis of risk-based direct regulation and standardization approaches, *Inf. Dev.* (2025) 02666669251352455, advance online publication. <https://doi.org/10.1177/02666669251352455>.
- [61] G. Papishev, K. J. D. Chan, AI regulatory strategies for digital sovereignty: the role of geopolitics and technological disparities, *Electron. Mark.* 36 (1) (2026) 8. <https://doi.org/10.1007/s12525-025-00870-z>.

- [62] D. Woods, Stag hunt in the digital wilds: legitimizing global AI governance amidst diverse terrains, *Fudan J. Humanit. Soc. Sci.* 19 (1) (2026) 65–96. <https://doi.org/10.1007/s40647-025-00438-3>.
- [63] A. Singh, J. Rodemann, R. Verma, S. L. Chau, K. Muandet, Incentive aware AI regulations: a credal characterisation, *arXiv:2603.05175v2 [preprint]* (2026). <https://arxiv.org/abs/2603.05175v2>.
- [64] J. D. Hartline, S. Long, C. Zhang, Regulation of algorithmic collusion, in: *Proceedings of the 2024 Symposium on Computer Science and Law*, ACM, 2024, pp. 98–108. <https://doi.org/10.1145/3614407.3643706>.
- [65] J. D. Hartline, C. Wang, C. Zhang, Regulation of algorithmic collusion, refined: testing pessimistic calibrated regret, in: *Proceedings of the 2025 Symposium on Computer Science and Law*, ACM, 2025, pp. 108–120. <https://doi.org/10.1145/3709025.3712217>.
- [66] N. Jethani, M. Sudarshan, I. Covert, S.-I. Lee, R. Ranganath, FastSHAP: real-time Shapley value estimation, in: *Proceedings of the 10th International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=Zq2G_VTV53T.
- [67] J. Yang, A. Li, M. Farajtabar, P. Suneag, E. Hughes, H. Zha, Learning to incentivize other learning agents, in: *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., 2020, pp. 15208–15219. <https://papers.nips.cc/paper/2020/hash/ad7ed5d47b9baceb12045a929e7e2f66-Abstract.html>.
- [68] T. LaCroix, A. Mohseni, The tragedy of the AI commons, *Synthese* 200 (4) (2022) 289. <https://doi.org/10.1007/s11229-022-03763-2>.
- [69] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, pp. 220–229. <https://doi.org/10.1145/3287560.3287596>.
- [70] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, K. Crawford, Datasheets for datasets, *Commun. ACM* 64 (12) (2021) 86–92. <https://doi.org/10.1145/3458723>.
- [71] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, I. Gabriel, Taxonomy of risks posed by language models, in: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2022, pp. 214–229. <https://doi.org/10.1145/3531146.3533088>.
- [72] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in AI safety, *arXiv:1606.06565 [preprint]* (2016). <https://arxiv.org/abs/1606.06565>.
- [73] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn, Direct preference optimization: your language model is secretly a reward model, in: *Advances in Neural Information Processing Systems*, Vol. 36, Curran Associates, Inc., 2023, pp. 53728–53741. https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.
- [74] M. G. Azar, Z. D. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, D. Calandriello, A general theoretical paradigm to understand learning from human preferences, in: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, Vol. 238 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 4447–4455. <https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html>.
- [75] J. Clymer, N. Gabrieli, D. Krueger, T. Larsen, Safety cases: how to justify the safety of advanced AI systems, *arXiv:2403.10462 [preprint]* (2024). <https://arxiv.org/abs/2403.10462>.
- [76] P. Paruchuri, J. P. Pearce, J. Marecki, M. Tambe, F. Ordonez, S. Kraus, Playing games for security: an efficient exact algorithm for solving Bayesian Stackelberg games, in: *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, Vol. 2, IFAAMAS, 2008, pp. 895–902. https://ifaamas.org/Proceedings/aamas08/proceedings/pdf/paper/AAMAS08_0057.pdf.
- [77] M. Jagadeesan, C. Mendler-Düner, M. Hardt, Alternative microfoundations for strategic classification, in: *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 4687–4697. <https://proceedings.mlr.press/v139/jagadeesan21a.html>.
- [78] L. S. Shapley, A value for n-person games, in: *Contributions to the Theory of Games II*, Princeton University Press, 1953, pp. 307–317, earlier version: RAND Paper P-295, 1952. <https://doi.org/10.1515/9781400881970-018>.
- [79] N. Nisan, T. Roughgarden, É. Tardos, V. V. Vazirani (Eds.), *Algorithmic game theory*, Cambridge University Press, 2007. <https://doi.org/10.1017/CBO9780511800481>.
- [80] C. Daskalakis, P. W. Goldberg, C. H. Papadimitriou, The complexity of computing a Nash equilibrium, *Commun. ACM* 52 (2) (2009) 89–97. <https://doi.org/10.1145/1461928.1461951>.
- [81] S. V. Albrecht, F. Christianos, L. Schäfer, *Multi-agent reinforcement learning: foundations and modern approaches*, MIT Press, 2024. <https://www.marl-book.com/>.
- [82] E. Kamenica, M. Gentzkow, Bayesian persuasion, *Am. Econ. Rev.* 101 (6) (2011) 2590–2615. <https://doi.org/10.1257/aer.101.6.2590>.
- [83] A. Ghorbani, J. Zou, Data Shapley: equitable valuation of data for machine learning, in: *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 2242–2251. <https://proceedings.mlr.press/v97/ghorbani19c.html>.
- [84] I. Covert, S. M. Lundberg, S.-I. Lee, Understanding global feature contributions with additive importance measures, in: *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., 2020, pp. 17212–17223. https://proceedings.neurips.cc/paper_files/paper/2020/file/c7bf0b7c1a86d5eb3be2c722cf2c746-Paper.pdf.
- [85] R. B. Myerson, Optimal auction design, *Math. Oper. Res.* 6 (1) (1981) 58–73. <https://doi.org/10.1287/moor.6.1.58>.
- [86] P. Bolton, M. Dewatripont, *Contract theory*, MIT Press, 2004. <https://mitpress.mit.edu/9780262025768/contract-theory/>.

- [87] J. Perdomo, T. Zrnic, C. Mendler-Dünnér, M. Hardt, Performative prediction, in: Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 7599–7609. <https://proceedings.mlr.press/v119/perdomo20a.html>.
- [88] C. Mendler-Dünnér, J. Perdomo, T. Zrnic, M. Hardt, Stochastic optimization for performative prediction, in: Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 4929–4939. <https://proceedings.neurips.cc/paper/2020/hash/33e75ff09dd601bbe69f351039152189-Abstract.html>.
- [89] G. Brown, S. Hod, I. Kalemaj, Performative prediction in a stateful world, in: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, Vol. 151 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 6045–6061. <https://proceedings.mlr.press/v151/brown22a.html>.
- [90] A. Asbell, M. Brundage, G. Hadfield, The role of cooperation in responsible AI development, arXiv:1907.04534 [preprint] (2019). <https://arxiv.org/abs/1907.04534>.
- [91] T. A. Han, L. M. Pereira, T. Lenaerts, F. C. Santos, Mediating artificial intelligence developments through negative and positive incentives, PLoS One 16 (1) (2021) e0244592. <https://doi.org/10.1371/journal.pone.0244592>.
- [92] W. Naudé, N. Dimitri, The race for an artificial general intelligence: implications for public policy, AI & Soc. 35 (2) (2020) 367–379. <https://doi.org/10.1007/s00146-019-00887-x>.
- [93] T. Cimpanu, F. C. Santos, L. M. Pereira, T. Lenaerts, T. A. Han, Artificial intelligence development races in heterogeneous settings, Sci. Rep. 12 (1) (2022) 1723. <https://doi.org/10.1038/s41598-022-05729-3>.
- [94] E. F. Domingos, T. A. Han, Falling behind drives unsafe development in an idealised AI race experiment, arXiv:2607.26034 [preprint] (2026). <https://arxiv.org/abs/2607.26034>.
- [95] J. Kang, Z. Xiong, D. Niyato, S. Xie, J. Zhang, Incentive mechanism for reliable federated learning: a joint optimization approach to combining reputation and contract theory, IEEE Internet Things J. 6 (6) (2019) 10700–10714. <https://doi.org/10.1109/jiot.2019.2940820>.
- [96] Y. Zhan, P. Li, Z. Qu, D. Zeng, S. Guo, A learning-based incentive mechanism for federated learning, IEEE Internet Things J. 7 (7) (2020) 6360–6368. <https://doi.org/10.1109/jiot.2020.2967772>.
- [97] R. Zeng, S. Zhang, J. Wang, X. Chu, FMore: an incentive scheme of multi-dimensional auction for federated learning in MEC, in: Proceedings of the 40th IEEE International Conference on Distributed Computing Systems, IEEE, 2020, pp. 278–288. <https://doi.org/10.1109/icdcs47774.2020.00094>.
- [98] P. Dütting, Z. Feng, H. Narasimhan, D. Parkes, S. S. Ravindranath, Optimal auctions through deep learning, in: Proceedings of the 36th International Conference on Machine Learning, Vol. 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 1706–1715. <https://proceedings.mlr.press/v97/duetting19a.html>.
- [99] M. Raghavan, S. Barocas, J. Kleinberg, K. Levy, Mitigating bias in algorithmic hiring: evaluating claims and practices, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2020, pp. 469–481. <https://doi.org/10.1145/3351095.3372828>.
- [100] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, C. J. Spanos, Towards efficient data valuation based on the Shapley value, in: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, Vol. 89 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 1167–1176. <https://proceedings.mlr.press/v89/jia19a.html>.
- [101] A. Ghorbani, M. Kim, J. Zou, A distributional framework for data valuation, in: Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 3535–3544. <https://proceedings.mlr.press/v119/ghorbani20a.html>.
- [102] J. T. Wang, R. Jia, Data Banzhaf: a robust data valuation framework for machine learning, in: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, Vol. 206 of Proceedings of Machine Learning Research, PMLR, 2023, pp. 6388–6421. <https://proceedings.mlr.press/v206/wang23e.html>.
- [103] X. Lin, X. Xu, Z. Wu, S.-K. Ng, B. K. H. Low, Distributionally robust data valuation, in: Proceedings of the 41st International Conference on Machine Learning, Vol. 235 of Proceedings of Machine Learning Research, PMLR, 2024, pp. 30362–30391. <https://proceedings.mlr.press/v235/lin24t.html>.
- [104] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 1885–1894. <https://proceedings.mlr.press/v70/koh17a.html>.
- [105] J. T. Wang, P. Mittal, D. Song, R. Jia, Data Shapley in one training run, in: Proceedings of the 13th International Conference on Learning Representations, 2025, pp. 12358–12395. <https://openreview.net/forum?id=HD6bWcj87Y>.
- [106] J. T. Wang, T. Yang, J. Zou, Y. Kwon, R. Jia, Rethinking data Shapley for data selection tasks: misleads and merits, in: Proceedings of the 41st International Conference on Machine Learning, Vol. 235 of Proceedings of Machine Learning Research, PMLR, 2024, pp. 52033–52063. <https://proceedings.mlr.press/v235/wang24cg.html>.
- [107] R. H. L. Sim, Y. Zhang, M. C. Chan, B. K. H. Low, Collaborative machine learning with incentive-aware model rewards, in: Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 8927–8936. <https://proceedings.mlr.press/v119/sim20a.html>.
- [108] A. Agarwal, M. Dahleh, T. Sarkar, A marketplace for data: an algorithmic solution, in: Proceedings of the ACM Conference on Economics and Computation, ACM, 2019, pp. 701–726. <https://doi.org/10.1145/3328526.3329589>.
- [109] D. Han, M. Wooldridge, A. Rogers, O. Ohrimenko, S. Tschitschek, Replication robust payoff allocation in submodular cooperative games, IEEE Trans. Artif. Intell. 4 (5) (2023) 1114–1128. <https://doi.org/10.1109/TAI.2022.3195686>.

- [110] M. Anderljung, J. Barnhart, A. Korinek, J. Leung, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, B. Chang, T. Collins, T. Fist, G. Hadfield, A. Hayes, L. Ho, S. Hooker, E. Horvitz, N. Kolt, J. Schuett, Y. Shavit, D. Siddarth, R. Trager, K. Wolf, Frontier AI regulation: managing emerging risks to public safety, arXiv:2307.03718 [preprint] (2023). <https://arxiv.org/abs/2307.03718>.
- [111] S. R. Ahmed, Z. M. I. Hossain, R. R. Pranta, K. P. Diba, Hierarchical governance equilibrium for human–AI decision systems: a hybrid principal–agent and Stackelberg game-theoretic model, *J. Artif. Intell. Technol. Dev.* 2 (3) (2026) 31–53. [https://doi.org/10.59324/jaitd.2026.2\(3\).03](https://doi.org/10.59324/jaitd.2026.2(3).03).
- [112] T. Shevlane, Structured access: an emerging paradigm for safe AI deployment, in: *The Oxford Handbook of AI Governance*, Oxford University Press, 2022, pp. 604–618. <https://doi.org/10.1093/oxfordhb/9780197579329.013.39>.
- [113] M. Yaghini, P. Liu, F. Boenisch, N. Papernot, Regulation games for trustworthy machine learning, arXiv:2402.03540 [preprint] (2024). <https://arxiv.org/abs/2402.03540>.
- [114] K. J. D. Chan, G. Papsyshev, M. Yarime, Balancing the tradeoff between regulation and innovation for artificial intelligence: an analysis of top-down command and control and bottom-up self-regulatory approaches, *Technol. Soc.* 79 (2024) 102747. <https://doi.org/10.1016/j.techsoc.2024.102747>.
- [115] X. Zhang, W. Zhou, W. Hu, S. Zhou, X. Hu, L. Yang, Can ex ante conformity assessment regulations contribute to trustworthy foundation models? An evolutionary game analysis from an innovation ecosystem perspective, *Technol. Soc.* 82 (2025) 102900. <https://doi.org/10.1016/j.techsoc.2025.102900>.
- [116] B. Mohammadi, N. Malik, T. Derdenger, K. Srinivasan, Regulating explainable artificial intelligence (XAI) may harm consumers, *Mark. Sci.* 44 (3) (2025) 711–724, earlier author version: arXiv:2209.03499v3. <https://doi.org/10.1287/mksc.2022.0396>.
- [117] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney, FactSheets: increasing trust in AI services through supplier's declarations of conformity, *IBM J. Res. Dev.* 63 (4/5) (2019) 6:1–6:13. <https://doi.org/10.1147/jrd.2019.2942288>.
- [118] Z. Alalawi, P. Bova, T. Cimpeanu, A. Di Stefano, M. H. Duong, E. Fernandez Domingos, T. A. Han, M. Krellner, B. Ogbo, S. T. Powers, F. Zimmario, Trust AI regulation? Discerning users are vital to build trust and effective AI regulation, *Appl. Math. Comput.* 508 (2026) 129627. <https://doi.org/10.1016/j.amc.2025.129627>.
- [119] C. W. Kim, D. Choo, T. Y. Neoh, M. Tambe, Incentive-aware AI safety via strategic resource allocation: a Stackelberg security games perspective, in: *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems, IFAAMAS, 2026*, pp. 3887–3892, Blue Sky Ideas Track. <https://arxiv.org/abs/2602.07259>.
- [120] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328. <https://proceedings.mlr.press/v70/sundararajan17a.html>.
- [121] K. Aas, M. Jullum, A. Løland, Explaining individual predictions when features are dependent: more accurate approximations to Shapley values, *Artif. Intell.* 298 (2021) 103502. <https://doi.org/10.1016/j.artint.2021.103502>.
- [122] M. Sundararajan, A. Najmi, The many Shapley values for model explanation, in: *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 9269–9278. <https://proceedings.mlr.press/v119/sundararajan20b.html>.
- [123] T. Heskes, E. Sijben, I. G. Bucur, T. Claassen, Causal Shapley values: exploiting causal knowledge to explain individual predictions of complex models, in: *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., 2020, pp. 4778–4789. https://proceedings.neurips.cc/paper_files/paper/2020/file/32e54441e6382a7fbacbbbf3c450059-Paper.pdf.
- [124] M. Sundararajan, K. Dhamdhere, A. Agarwal, The Shapley Taylor interaction index, in: *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 9259–9268. <https://proceedings.mlr.press/v119/sundararajan20a.html>.
- [125] C.-P. Tsai, C.-K. Yeh, P. Ravikumar, Faith-Shap: the faithful Shapley interaction index, *J. Mach. Learn. Res.* 24 (94) (2023) 1–42. <https://jmlr.org/papers/v24/22-0202.html>.
- [126] I. Covert, C. Kim, S.-I. Lee, Learning to estimate Shapley values with vision transformers, in: *Proceedings of the 11th International Conference on Learning Representations*, 2023. https://openreview.net/forum?id=5ktFNz_pJLK.
- [127] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ACM, 2020, pp. 180–186. <https://doi.org/10.1145/3375627.3375830>.
- [128] G. Irving, P. Christiano, D. Amodei, AI safety via debate, arXiv:1805.00899 [preprint] (2018). <https://arxiv.org/abs/1805.00899>.
- [129] M. Hardt, N. Megiddo, C. Papadimitriou, M. Wootters, Strategic classification, in: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ACM, 2016, pp. 111–122. <https://doi.org/10.1145/2840728.2840730>.
- [130] M. Brückner, T. Scheffer, Stackelberg games for adversarial prediction problems, in: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2011, pp. 547–555. <https://doi.org/10.1145/2020408.2020495>.
- [131] N. Dalvi, P. Domingos, Mausam, S. Sanghai, D. Verma, Adversarial classification, in: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2004, pp. 99–108. <https://doi.org/10.1145/1014052.1014066>.
- [132] S. Ahmadi, H. Beyhaghi, A. Blum, K. Naggita, The strategic perceptron, in: *Proceedings of the 22nd ACM Conference on Economics and Computation*, ACM, 2021, pp. 6–25. <https://doi.org/10.1145/3465456.3467629>.
- [133] G. Ghalme, V. Nair, I. Eilat, I. Talgam-Cohen, N. Rosenfeld, Strategic classification in the dark, in: *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 3672–3681. <https://proceedings.mlr.press/v139/ghalme21a.html>.

- [134] Y. Shavit, B. Edelman, B. Axelrod, Causal strategic linear regression, in: Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 8676–8686. <https://proceedings.mlr.press/v119/shavit20a.html>.
- [135] J. Miller, S. Milli, M. Hardt, Strategic classification is causal modeling in disguise, in: Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 6917–6926. <https://proceedings.mlr.press/v119/miller20b.html>.
- [136] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, H. Wallach, A reductions approach to fair classification, in: Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 60–69. <https://proceedings.mlr.press/v80/agarwal18a.html>.
- [137] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: auditing and learning for subgroup fairness, in: Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 2564–2572. <https://proceedings.mlr.press/v80/kearns18a.html>.
- [138] R. Greenblatt, B. Shlegeris, K. Sachan, F. Roger, AI control: improving safety despite intentional subversion, in: Proceedings of the 41st International Conference on Machine Learning, Vol. 235 of Proceedings of Machine Learning Research, PMLR, 2024, pp. 16295–16336. <https://proceedings.mlr.press/v235/greenblatt24a.html>.
- [139] E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, A. Jermyn, A. Askill, A. Radhakrishnan, C. Anil, D. Duvenaud, D. Ganguli, F. Barez, J. Clark, K. Ndousse, K. Sachan, M. Sellitto, M. Sharma, N. DasSarma, R. Grosse, S. Kravec, Y. Bai, Z. Witten, M. Favaro, J. Brauner, H. Karnofsky, P. Christiano, S. R. Bowman, L. Graham, J. Kaplan, S. Mindermann, R. Greenblatt, B. Shlegeris, N. Schiefer, E. Perez, Sleeper agents: training deceptive LLMs that persist through safety training, arXiv:2401.05566 [preprint] (2024). <https://arxiv.org/abs/2401.05566>.
- [140] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, A. Ng, M. N. Halgamuge, A game-theoretic approach to containing artificial general intelligence: insights from highly autonomous aggressive malware, IEEE Trans. Artif. Intell. 5 (12) (2024) 6290–6303. <https://doi.org/10.1109/TAI.2024.3394392>.
- [141] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, T. Graepel, Open problems in cooperative AI, arXiv:2012.08630 [preprint] (2020). <https://arxiv.org/abs/2012.08630>.
- [142] K. Donahue, J. Kleinberg, Model-sharing games: analyzing federated learning under voluntary participation, Proc. AAAI Conf. Artif. Intell. 35 (6) (2021) 5303–5311. <https://doi.org/10.1609/aaai.v35i6.16669>.
- [143] K. Donahue, J. Kleinberg, Optimality and stability in federated learning: a game-theoretic approach, in: Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, Inc., 2021, pp. 1287–1298. <https://proceedings.neurips.cc/paper/2021/hash/09a5e2a11bea20817477e0b1dfe2cc21-Abstract.html>.
- [144] Y. Bijoor, AI governance in the United States: using game theory to promote innovation and uphold ethical standards, SSRN working paper 6463740, written 14 November 2024; posted 20 April 2026 (2024). <https://doi.org/10.2139/ssrn.6463740>.
- [145] B. Skyrms, The stag hunt and the evolution of social structure, Cambridge University Press, 2004. <https://www.cambridge.org/core/books/stag-hunt-and-the-evolution-of-social-structure/AB3181FA469D6EA411E05D6CBF644507>.
- [146] I. D. Raji, I. E. Kumar, A. Horowitz, A. D. Selbst, The fallacy of AI functionality, in: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, ACM, 2022, pp. 959–972. <https://doi.org/10.1145/3531146.3533158>.
- [147] J. Buolamwini, T. Gebru, Gender Shades: intersectional accuracy disparities in commercial gender classification, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Vol. 81 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 77–91. <https://proceedings.mlr.press/v81/buolamwini18a.html>.
- [148] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 59–68. <https://doi.org/10.1145/3287560.3287598>.
- [149] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. Pentland, M. E. Roberts, A. Shariff, J. B. Tenenbaum, M. Wellman, Machine behaviour, Nature 568 (2019) 477–486. <https://doi.org/10.1038/s41586-019-1138-y>.
- [150] S. McGregor, Preventing repeated real world AI failures by cataloging incidents: the AI Incident Database, Proc. AAAI Conf. Artif. Intell. 35 (17) (2021) 15458–15463. <https://doi.org/10.1609/aaai.v35i17.17817>.