

## Review

Kris Carlson prompting ChatGPT4o

Prompt

Please provide a summary of Tomei et al., AI Governance through Markets (<https://arxiv.org/abs/2501.17755>), noting each market-based AI governance solution the authors propose to replace a corresponding government solution, followed by a critique of the paper, noting any market solutions that are idealistic and/or impracticable to implement in the near term (0 - 5 years).

# AI Governance through Markets

Philip Moreira Tomei, Rupal Jain, Matija Franklin

## Summary

**Overview:** Tomei et al. argue that **market-based mechanisms** should complement traditional government regulation in AI governance. They frame current AI risk as a form of **market failure** — uncertainty and information gaps prevent efficient risk management and capital allocation. To address this, the authors examine four key market governance proposals – **insurance, auditing & certification, procurement standards, and investor due diligence** – which align financial incentives with safer AI development. Crucially, they note that **market forces alone are insufficient** to protect society; instead, *standardized AI disclosure* must underpin these mechanisms, and regulation remains important in parallel.

## Market Governance vs Traditional Regulation

In traditional governance, governments use laws, oversight agencies, and formal regulations to manage AI risks (e.g. safety standards mandated by law or direct regulatory supervision). By contrast, **market governance mechanisms** leverage economic incentives and private-sector actions to encourage responsible AI behavior. Each proposal the authors discuss has a traditional counterpart it could *augment or substitute*: for example, insurance markets can complement liability laws, independent audits can supplement regulatory compliance checks, procurement policies can substitute for direct product regulations, and investor due diligence can parallel government-imposed disclosure and risk oversight. The paper's aim is not to replace government, but to **"widen the possibility space"** of AI governance beyond hard law alone. In what follows, each market-based proposal is summarized along with its analogous traditional solution.

## Risk Distribution via Insurance

**Proposal:** Use insurance markets to distribute and price AI risks. **Insurance** transfers risk from individual AI developers/users to a broad pool by charging premiums, thereby enabling transactions under uncertainty. The authors argue insurance can be a **"powerful governance mechanism"** that *shapes* AI development through **risk-based pricing, underwriting standards, and policy exclusions**. In practice, insurers would assess an AI system's risk (e.g. likelihood of

causing harm) and adjust premiums or require safety measures accordingly. This creates financial incentives for companies to reduce risks (to lower their premiums) and offers tools to handle **liability** for AI-caused damages. Notably, insurance could address “thorny liability issues” of AI – for example, if it’s unclear whether the developer, deployer, or user is at fault, an insurance payout can ensure victims are compensated. The paper cites how **high-variance, high-magnitude risks** (characteristic of advanced AI systems) historically spur new insurance products, as insurers seek to cover emerging technologies. A case study of commercial real estate is given: as building management employs AI, rising insurance premiums (due to new AI-related hazards) might incentivize owners to invest in safer AI or risk mitigations, much as fire insurance encourages installing sprinklers.

**Traditional Analog: Government liability and safety regulations.** Traditionally, if an AI system causes harm, governments rely on product liability law or negligence standards – the courts can hold a company liable and require it to pay damages. Regulators might also impose safety standards directly (e.g. requiring AI in cars to meet certain fail-safes). **Insurance augments this by internalizing risk costs upfront:** instead of (or in addition to) legal penalties after a disaster, companies carry insurance and follow insurer-mandated best practices to prevent loss. This mirrors sectors like automotive, where mandatory insurance complements traffic laws by pricing risk. In essence, the insurance market can **serve a quasi-regulatory role** by *pricing AI risk appropriately* and thus pushing firms toward safer behavior, whereas government would otherwise have to enforce safety via inspections or fines. Insurance also spreads financial risk, which pure regulation doesn’t do – if a catastrophic AI failure occurs, an insurance pool would absorb losses rather than a single victim or company bearing all costs.

## Assurance via Auditing & Certification

**Proposal:** Deploy independent **AI auditing and certification** to ensure accountability and safety. Borrowing from financial auditing, the authors describe third-party AI audits as external assessments of a system’s operations, ethics, or compliance with standards. **Certification** is the public attestation that an AI product meets certain benchmarks (often following an audit). Tomei et al. emphasize that making audit results public **creates reputational and market pressure:** firms have incentives to fix problems that auditors find, since failing to pass an audit can hurt their reputation, investor trust, or invite regulatory action. In effect, audits increase **transparency and accountability** between AI producers and stakeholders. Audits also reduce **information asymmetry** – an external audit report reveals a system’s true performance and risks to investors, customers, and regulators who might otherwise rely on the company’s own (potentially biased) claims. The paper reviews literature arguing that AI audits are needed both for **public safety** and to hold AI firms to their promised standards. Importantly, audit findings (if disclosed) give regulators and researchers insight into AI behaviors and limitations. This contributes to broader governance by informing policy and scientific understanding of advanced systems. Moreover, audits can strike a balance in the **transparency vs. intellectual property trade-off:** rather than a company fully open-sourcing an AI (which could pose IP or security risks), a confidential third-party audit can evaluate it and assure others of its safety without revealing proprietary details. A case in point is **Zoom’s security audit:** after privacy flaws were revealed, Zoom hired outside cybersecurity firms to audit and certify improved encryption. The result was strengthened security and restored user

trust – illustrating how third-party audits function as a market mechanism to identify and mitigate AI-related risks in practice.

**Traditional Analog: Regulatory oversight and certification by government.** In classical regulation, a government might mandate testing or approval for high-risk AI systems (similar to how the FDA approves medical software or how finance regulators audit banks). Government agencies could conduct inspections or require **compliance reports**. The market-based audit approach supplements this by leveraging private expertise: **independent audit firms** play a role akin to regulators, but companies engage them proactively to signal quality or to meet customer demands. For instance, rather than waiting for a government inspector to uncover an AI system's flaws, a firm might seek an **AI safety certification** (from an industry body or standards organization) to gain a competitive edge. This parallels how *financial audits of public companies are required by law but performed by private CPA firms* – a blend of regulation and market service. In AI, we see early moves in this direction (e.g. the EU's draft AI Act will require external conformity assessments for certain AI – essentially mandated audits). Thus, the authors' proposal augments regulatory structures by building an **ecosystem of accredited AI auditors** and certification standards, which can scale faster or delve deeper technically than government alone might. It is intended to bolster accountability: companies are not only answerable to regulators but also to auditors and the marketplace if they produce unsafe AI.

## Protocolization via Procurement Standards

**Proposal:** Leverage large buyers' **procurement requirements** to enforce AI safety and ethics standards. **Procurement standards** refer to the criteria and protocols that organizations (especially governments or big corporations) demand from AI vendors and contractors. The authors note that governments are major AI customers (often relying on private contractors for AI development) and thus can shape industry practices by setting rigorous conditions on the AI systems they buy. For example, a government might require that any AI system it procures **meets security and transparency guidelines**. These standards serve as **bottom-up governance**: by tying compliance to lucrative contracts, they *indirectly regulate* the industry and drive innovation toward safer designs. The paper points out that in the defense sector, detailed procurement rules (like the U.S. Department of Defense's **DFARS** regulations) impose strict safety, performance, and reporting requirements on suppliers. Vendors comply not by law per se, but because otherwise they cannot do business with the Department of Defense – effectively **creating de facto regulations via market power**. Such procurement-driven rules have broad impact: companies worldwide adopt these practices to remain eligible for contracts, spreading standards internationally. Outside of government, large tech companies also use procurement policies to influence their supply chains (e.g. **Apple or Walmart setting supplier codes of conduct**). Tomei et al. suggest this concept be applied to AI: if major buyers demand **certain safety audits, risk assessments, or alignment benchmarks** as a condition of purchase, AI developers will conform in order to access those markets. The authors draw an analogy to the FDA's pre-market certification for medical devices, which expedites approval if new devices meet established safety equivalence – similarly, an AI procurement standard could require predefined risk metrics be satisfied before an AI system is integrated into critical uses. This would incentivize a baseline of safety and **public accountability measures** (like reporting any post-deployment incidents), creating a feedback loop that encourages ongoing risk mitigation. The **NASA Apollo program** is

presented as an illustrative case: NASA's extremely stringent procurement protocols for contractors (reliant on NASA's own safety standards and the Federal Acquisition Regulations) pushed aerospace companies to achieve unprecedented reliability and safety, resulting in technological breakthroughs without sacrificing safety in a high-risk endeavor. In sum, by **using purchasing power as leverage**, procurement standards can guide AI development toward safer outcomes in the absence of, or ahead of, formal regulations.

**Traditional Analog: Direct regulatory standards for products.** Normally, governments might pass laws that *all* AI systems (not just those sold to government) must meet certain safety criteria. For instance, regulators could outright require that any AI driving car on public roads has been tested to a specific safety level. Procurement is a more selective tool: it doesn't ban unsafe AI from the market, but **filters what large buyers will accept**, thereby indirectly raising standards. In essence, it's **regulation by market demand**. Government procurement standards in AI serve as a flexible alternative to legislation – they can be updated more easily than laws and tailored per contract. This approach **augments formal regulation**: even in areas where laws are still catching up, a government agency today could decide to only buy AI systems that align with an AI Ethics Charter, effectively implementing governance immediately. Over time, these contract requirements might evolve into broadly recognized standards (or inspire future laws). For private sector buyers, their procurement preferences act similarly to **industry self-regulation** – for example, if big cloud providers refuse to deploy AI models that lack transparency documentation, it pressures the AI vendors to produce that documentation. Compared to traditional lawmaking, procurement-driven governance is faster and can be *scoped*: e.g. defense can impose ultra-high safety bars for military AI, while a smaller-scale use might have lighter requirements. However, unlike a universal law, procurement only binds participants in those transactions, so its reach depends on how influential the buyers are. The authors highlight that with enormous buyers like the U.S. Department of Defense (with a **\$916B budget**) the effect can be global, essentially **“establishing global procurement paradigms”** as suppliers worldwide adapt to meet these requirements. Thus, procurement standards operate as a market-driven regulatory structure, complementing government rules by filling gaps and pushing best practices through contractual obligation.

## Investor Behavior via Due Diligence

**Proposal:** Encourage investors to govern AI risk through rigorous **due diligence** in their funding decisions. **Investor due diligence** is the process by which investors (venture capitalists, banks, or shareholders) thoroughly investigate a company's operations, finances, and risks before investing. The authors contend that when investors scrutinize a tech company's AI practices, it becomes a **self-regulating mechanism** that complements formal regulation. By examining an AI developer's safety measures, compliance with norms, and potential liabilities, investors can **price in AI risk** – for example, demanding a lower valuation or higher returns for a company with unmitigated AI hazards. This creates a **carrot-and-stick incentive**: companies with **robust safety and governance** will attract investment more easily (the “carrot”), whereas those with sloppy or high-risk AI practices will struggle to get funding or face a risk premium (the “stick”). The paper explains that due diligence reduces **information asymmetry** between company insiders and outside investors by uncovering information not in public reports. Knowing that they will be examined, AI firms have reason to maintain accurate records and **adopt good risk management** proactively.

Several facets of due diligence are highlighted: **financial** (are AI investments financially sound), **legal** (assessing algorithmic bias, IP, liability risks), **operational** (risk management processes, cybersecurity), **technical** (evaluation of the AI model's reliability and data quality by experts), and **market** (the competitive and disruptive impact of the AI). In each facet, the presence of **standard evaluation criteria and disclosures** about the AI can help investors compare companies (e.g. using common safety benchmarks). Ultimately, the governance power of due diligence lies in the investors' ability to **allocate capital** – by funding safer, more aligned AI ventures and steering money away from reckless ones, the market collectively rewards responsible innovation. The authors give the **BP Deepwater Horizon oil spill** as an analog: after BP's disaster in 2010 (a massive risk management failure), investors responded by conducting far more stringent due diligence on oil companies' safety protocols. BP's shareholders pushed for improved disclosures and governance, leading BP to reallocate billions towards safer, sustainable operations. This illustrates how investors, motivated by avoiding future losses, forced a company to reform and reduce high-risk activities. By analogy, if an AI company had a major failure or displayed risky behavior, investors could demand changes (like independent audits, new safety expenditures) as a condition of continued financing. Thus, **market discipline via investors** can drive AI developers toward alignment with societal interests.

**Traditional Analog: Mandatory risk disclosure and corporate governance law.** In regulated markets, governments often require companies (especially public ones) to **disclose material risks** and uphold fiduciary duties to shareholders. Regulators like the SEC can sanction companies that hide risks. Due diligence as a market mechanism complements this by having investors actively seek out risk information, beyond what minimal regulations demand. It is akin to **self-policing** by the market: rather than waiting for a regulator to flag an issue, a wary investor will walk away or insist on fixes, effectively **penalizing dangerous AI practices through lack of capital**. In a sense, it parallels how **credit ratings** or **Environmental, Social, Governance (ESG) criteria** influence investment – those are private-sector evaluations that can deter investment in non-compliant firms. The authors explicitly note that due diligence *“acts as a self-regulating mechanism that complements formal regulatory frameworks”*. So, while a government might impose rules (e.g. require AI risk factors in annual reports, or hold boards liable for AI harms), the **market-based solution is investors voluntarily integrating AI risk into their investment decisions**. It's intended to augment regulatory oversight by catching issues early: if, say, there is no law (yet) against deploying a semi-aligned AI, a prudent investor might still balk and thus indirectly prevent that deployment. Additionally, investor pressure can reinforce accountability – a company's management becomes accountable not only to regulators and courts, but to its **owners and financiers** for managing AI responsibly. This mirrors traditional corporate governance where shareholders can demand changes in management or strategy if they see negligence. In summary, robust due diligence is the market's parallel to government monitoring: it **harnesses capital as leverage** to ensure AI firms align with safety and ethical expectations.

## The Foundation: Standardized AI Disclosure

All four market mechanisms above rely on **information** to function. The authors underscore that **standardized disclosure of AI development and deployment details** is a foundational enabler for market governance. Currently, there is significant **information asymmetry** – AI developers know the capabilities and risks of their systems far better than outsiders (insurers, auditors,

investors, or even regulators) do. This gap can lead to market failures: if buyers or insurers cannot assess an AI's risk, they might **underestimate or overestimate the risk**, leading to mispricing and inefficient outcomes. Tomei et al. propose creating **standard disclosure frameworks** for AI, akin to financial reporting standards, so that companies release relevant information about their AI systems in a consistent way. Such disclosures could include data on an AI model's training process, safety evaluation results, known limitations, and alignment safeguards. The paper notes that in other industries (finance, pharmaceuticals), standardized information reporting reduced transaction costs and externalities. For AI, this would mean insurers can properly price premiums, auditors have data to audit against, investors can compare risk profiles, and procurers can set clear benchmarks. Disclosure thus "**powers auditing, insurance, due diligence and procurement**" by providing a common epistemic foundation. The authors discuss how many investors globally are already pushing for **non-financial disclosures** through initiatives like the UN Principles for Responsible Investment (with signatories managing \$59 trillion in assets) – signaling appetite for transparency about things like AI ethics and risk. They also acknowledge challenges: companies fear exposing trade secrets or security details, so a balance must be struck. In their **Discussion**, the authors explore the trade-off between **intellectual property vs. public transparency**. They suggest audits and carefully scoped disclosures can mitigate this tension (e.g. reveal safety metrics without revealing the full model). Ultimately, the paper calls for **developing standard AI disclosure practices** as a critical next step, so that market mechanisms have the necessary information to function effectively in governing AI.

## Conclusion of the Paper

Tomei et al. conclude that **market-based governance mechanisms, reinforced by standardized disclosures, can create powerful incentives for safe and responsible AI development**. They stress that these approaches are *complements* to traditional regulation, not a replacement. The paper "*urges regulators, economists, and ML researchers to investigate and implement*" such market-based tools alongside laws. By aligning AI risks with financial risks, markets can nudge AI development in a pro-social direction even in the absence of perfect regulations. The authors present their proposals as **opportunities for both public and private intervention**, backed by market forces. They also acknowledge that their four highlighted mechanisms are not exhaustive – rather, they illustrate how leveraging insurance, audits, procurement, and investor scrutiny could help govern AI, and they encourage exploring other market levers as well. Finally, the paper calls for further research and pilot programs to **standardize AI risk information** and to test these market mechanisms in practice. The overarching message is optimistic: if implemented well, "**market governance mechanisms**" can incentivize AI safety, distribute risks, and *augment formal regulatory frameworks* in addressing both present AI challenges and future advanced AI systems.

## Critique and Analysis of Proposals

### Practicability in 0–5 Years

In evaluating each market-based governance proposal, we focus on **(1) AI safety and value alignment, (2) accountability mechanisms, and (3) regulatory structures for AGI**, especially regarding feasibility in the near term (the next five years). While these proposals are innovative,

some appear more **realistic** than others in the short run. We examine each in turn, noting which ideas might be **idealistic or difficult to implement** within five years and why. In general, these market mechanisms show promise for *narrow AI* applications (specific, limited AI systems), but their effectiveness and readiness for **AGI (Artificial General Intelligence)** governance is questionable. AGI poses unique challenges – potentially extreme impacts, strategic race dynamics, and hard-to-measure risks – that may stretch or outstrip what market incentives alone can manage.

### **Insurance Mechanisms – Feasibility and Implications**

**Practicability:** Using insurance to govern AI behavior is an intriguing idea, but in the next 0–5 years it faces practical hurdles. **Insurance markets for AI risk are still nascent.** Traditional insurers currently struggle to price the novel risks of AI due to limited historical data and the complex, evolving nature of AI failures. Indeed, industry analyses suggest it’s “*unlikely that a single insurance policy will cover all potential risks that AI presents.*”. Insurers are cautious: for example, certain **AI failure modes (like a rogue autonomous system causing widespread harm)** might be deemed uninsurable because the losses could be catastrophic or hard to quantify. In the near term, we may see insurance offerings for **narrower AI risks** – e.g. malpractice insurance for AI medical diagnostics, or liability coverage for autonomous vehicle accidents – because these can be at least roughly quantified using analogies to human error rates or auto accident statistics. However, for cutting-edge **AGI systems or large-scale AI decision-makers**, insurers have no actuarial baseline and might exclude such risks or charge prohibitively high premiums. The authors note that AI’s combination of high uncertainty and high stakes “*traditionally catalyse[s] insurance market development*”, implying insurers will eventually innovate products. Within five years, we might expect **pilot AI insurance programs** (perhaps led by specialty insurers or government-backed pools), but widespread adoption of AI insurance as a standard practice is likely idealistic on that timeline.

**Safety & Alignment:** If AI insurance does take hold, its impact on **AI safety and value alignment** could be positive but somewhat indirect. Insurance could **internalize externalities** – making developers financially accountable for harm – which in theory motivates them to prevent misaligned behavior. For example, if a company knows an out-of-control AI could result in huge liability costs (via insurance premiums or deductibles), it has incentive to invest in **alignment techniques and safety testing** upfront. Insurers, looking to minimize payouts, might impose their own **underwriting standards**: they could require the AI system to pass certain safety audits or to have an “alignment compliance certificate” before granting coverage. This effectively forces companies to adopt safety measures or else face no insurance (and thus difficulty selling their product). In practice, though, insurance mainly **mitigates consequences** rather than preventing them. A poorly aligned AI might still cause damage; insurance would ensure victims are compensated and the company pays financially, but that doesn’t guarantee the AI system was safe to begin with. **AGI-scale risks** further complicate this – an insurer can pay out for property damage or even loss of life in contained incidents, but if an advanced AI caused *existential* or systemic catastrophe, no payout can undo that. Insurers also might not cover extreme scenarios (“uninsurable risks”), which means those worst-case risks remain unpriced and thus not actively mitigated by insurance incentives. In summary, insurance can improve accountability and risk

management for **moderate AI failures** (encouraging safety to avoid accidents), but it is not a panacea for ensuring true alignment in advanced AI that could have unprecedented behaviors.

**Accountability:** Insurance does provide a form of **financial accountability**. It ensures there is a responsible party (the insurer and insured) who will bear the cost of an AI system's failures, rather than leaving victims uncompensated. This aligns with the principle of "**internalizing the externality**" – AI developers can't externalize all harm, since harmful outcomes drive their insurance costs up. However, insurance could also diffuse accountability in some cases: if a company is fully insured, its direct liability is cushioned (except possibly reputational damage). A firm might treat insurance as a license to take risks – akin to how some drivers drive less carefully knowing they have auto insurance. To counteract this moral hazard, insurers typically set deductibles and investigate cause of losses. In AI, if insurers actively investigate accidents and penalize negligence (e.g. voiding coverage if the company ignored known safety procedures), that bolsters accountability. Yet, **legal accountability** remains crucial – insurance works best alongside a liability regime where companies are still legally accountable for willful misconduct or gross negligence beyond just paying insurance premiums. In a five-year horizon, we are likely to see debates on how liability laws for AI (or even for autonomous AI agents) should be structured in tandem with insurance. The proposal to rely on insurance is realistic only if there's clarity on **who is liable when AI fails** (the developer? the user? the AI itself?). Without clear liability assignment from governments, insurers cannot properly underwrite the risk. So, implementing AI insurance at scale may require some regulatory foundation (e.g. laws that define AI as a product or service with certain liability rules), which in many jurisdictions is still in draft stages.

**AGI Regulatory Structures:** For **AGI-level systems**, insurance is largely inadequate as a sole governance tool. AGI could introduce **systemic, correlated risks** (for instance, if many systems share a flaw, one incident might cascade globally) that traditional insurance cannot cover. It might play a supporting role – for example, perhaps any organization developing AGI is required to carry a hefty liability insurance (as a **financial security requirement**, similar to how nuclear power plants must have insurance or backing for disaster liability). That could ensure they have skin in the game. But regulators will likely treat AGI more like an extreme hazard – possibly requiring direct **government oversight, licensing, or even international treaties** beyond what insurance markets handle. In the near term, expecting insurance markets to address **AGI existential risks** is idealistic. Instead, insurance might feasibly address **narrower issues (bias, privacy breaches, accidents)** and help gradually build a risk pricing framework. It contributes to a broader regulatory structure by informing regulators about risk levels (e.g. if insurance premiums for certain AI practices skyrocket, that's a market signal of high risk which regulators could heed). Yet, for AGI, one cannot rely on insurers to prevent a catastrophe that they themselves cannot absorb. Thus, while the **insurance proposal is innovative**, within five years its role will likely be *limited to specific domains* and primarily as a complement to, not replacement for, strong regulatory safeguards.

### **Auditing & Certification – Feasibility and Implications**

**Practicability:** Among the proposals, **third-party AI auditing and certification is one of the more immediately actionable** in the near term. There is already a growing ecosystem discussing AI audits – from bias audits of algorithms to security evaluations of AI models – spurred in part by upcoming regulations (e.g. the EU AI Act) and corporate AI ethics programs. Within the next five

years, it's plausible that **industry standards for AI audit** will emerge and a number of firms will specialize in offering AI audit services. In fact, the authors note extensive literature and ongoing efforts around AI auditing for safety and accountability. We can expect the **“Big Four” accounting firms and tech consultancies** to expand their services to include AI audits, much as they audit cybersecurity or data privacy compliance today. Some challenges, however, remain: there is not yet a consensus on **audit criteria for AI alignment**. It's straightforward to audit certain aspects (for example, an auditor can check if a model meets a client's stated accuracy and fairness metrics, or if proper data governance was followed), but **auditing an AI for latent dangers or future misbehavior is hard**. Especially with AI systems based on machine learning, which can be opaque (“black boxes”), an audit might only scratch the surface (e.g. reviewing documentation and some test results). **Tooling and expertise** for deep audits are still developing – over the next five years, the field of **AI evaluation** (red-teaming, interpretability analysis, etc.) will need to mature to make audits truly effective. Despite these difficulties, **short-term realism is fairly high** that audits will be part of AI governance: companies deploying AI in sensitive areas may voluntarily seek certifications (for competitive advantage or to assure customers), and regulators are likely to mandate audits for high-risk AI systems. Thus, implementing the auditing proposal is feasible, *with the caveat* that audit quality and consistency are concerns. Early “AI audit” labels might end up being superficial (a rubber stamp) unless rigorous standards are established. We may see industry bodies or governments defining **baseline audit requirements** to prevent a lowest-common-denominator approach.

**Safety & Alignment:** If implemented well, **auditing can significantly advance AI safety and alignment**, by introducing an external check on a developer's claims and methodologies. Auditors can be tasked with examining whether an AI system adheres to certain safety guidelines or ethical principles. For example, an audit might evaluate how an AI was trained to avoid extremist content or whether it has guardrails to prevent harmful actions. This creates accountability for alignment: a company can't just claim their AI is safe – they need to prove it to independent examiners. Moreover, the authors highlight that audit results, when shared, **“provide regulators and the scientific community with critical insights into system behaviors and limitations”**, aiding broader understanding. That means audits can uncover issues (bias, failure modes, etc.) that, if published or at least reported to regulators, contribute to the **collective knowledge on safe AI practices**. Auditing also intersects with **value alignment**: auditors could check not just technical safety, but whether the AI's outputs align with stated values or societal norms (for instance, does a chatbot respect privacy and avoid toxic language as intended?). However, we must temper expectations: an audit is often a *point-in-time or periodic* assessment. An AI model that passes an audit in January might **drift** or be repurposed in a way that makes it unsafe by June. For continuous alignment of an AGI, something more dynamic than occasional audits may be needed (e.g. ongoing monitoring or internal “AI ethics boards”). Another limitation is that **audits rely on the auditor's ability to understand the system** – with very complex AI (like a potential AGI that learns and changes itself), even expert auditors might struggle to foresee all failure modes. In five years, audits will likely improve **transparency and trust for narrow AI** (catching issues like algorithmic bias or security vulnerabilities, as in the Zoom case), but auditing for **emergent misalignment in AGI** is an unsolved problem. Still, as a safety mechanism, having independent eyes on AI systems is a valuable layer, reducing the chance that obvious warning signs are missed due to corporate

incentives. It's a step toward a culture of **"safety compliance"** in AI development, analogous to how aviation has independent safety inspectors.

**Accountability:** Auditing directly bolsters accountability by introducing **third-party oversight**. A company that undergoes an audit is implicitly **accountable to the auditor's standards and to the public or clients** who rely on the audit. If the AI system fails despite a "certified safe" audit, the auditor's reputation is on the line too, creating a chain of accountability. In the near term, we might see something like **AI audit certifications** (analogous to ISO certifications) that companies will publicly tout; failing to obtain or maintain one would signal accountability issues. Audits can also enforce **internal accountability**: knowing that an external party will verify their claims, AI developers and managers are more likely to adhere to their own stated policies (the paper cites Manheim et al., noting audits keep firms accountable to their commitments). One key aspect is that audits create **records** – documented findings that regulators or courts can later review if something goes wrong. This documentation ensures companies can't easily hide problems. However, the efficacy for accountability depends on audit transparency. If audit results are confidential, then accountability is only to the client or perhaps regulators. If they are public (or at least summaries are public), then there is broader accountability (to customers, investors, etc.). Within five years, a realistic scenario is **semi-public audit reporting**: companies might disclose that an audit was done and key findings, but not every detail (for IP or liability reasons). That still improves accountability compared to no audit. **One caveat:** auditors themselves must be held accountable (as seen in financial industries when auditors failed, e.g. in Enron's case). Regulatory oversight of the *auditing industry* might be necessary to avoid conflicts of interest (a company shopping for an easy auditor). For AGI, one can imagine a future where a specialized **"AGI auditor"** or evaluation agency is licensed by an international body to certify only those AGI systems that meet stringent safety criteria – that would be a strong accountability mechanism, akin to how the IAEA inspects nuclear facilities. In five years that's unlikely to exist, but the groundwork could be laid by expanding current audit practices. Overall, auditing is a comparatively **realistic and effective mechanism for accountability** in AI, especially for well-defined issues like bias, security, and compliance. It is not a full guarantee (audits can miss things), but it formalizes responsibility: someone is examining and vouching for the system.

**Regulatory Structures for AGI:** When it comes to **AGI** or highly autonomous AI, auditing would have to evolve into something much more intensive. **Regulators** might require *continuous auditing or monitoring* of AGI projects, not just a one-time check. Traditional markets might not spontaneously provide that level of scrutiny unless mandated because truly understanding an AGI could require months of assessment by top experts – an expensive proposition few would pay for voluntarily. So, while **third-party audits will likely be part of AGI governance**, they will need to be embedded in a larger regulatory framework (e.g. a condition of licensure to develop AGI might be frequent independent audits and publishings of safety test results). In the near term, before AGI exists, audits can help ensure the **precursors to AGI (like large language models)** are developed responsibly, potentially preventing reckless scaling by revealing alignment flaws early. The authors rightly see audits as a means to **inform regulators**. For instance, if a series of audits find that all large models have trouble with a certain type of harmful behavior, regulators can take note and perhaps issue guidance or rules. In that sense, audits *augment regulatory structures for AGI* by generating empirical evidence and standards. But, is it realistic to implement? Partially – we expect some big tech companies to willingly subject their models to external audits (OpenAI, for example,

has invited external “red teams” for GPT models). Government might also **fund independent audit teams** to evaluate frontier AI models for public safety. Within five years, establishing a **robust audit ecosystem** for current AI is feasible, which sets the stage for handling AGI. Yet, **auditing AGI for alignment** remains a largely theoretical exercise today. It may require new techniques (like AI-assisted auditing tools or interpretability breakthroughs) to be truly effective. In summary, **auditing and certification** is one of the more concrete, near-term actionable proposals – we should see its role grow quickly for narrow AI, improving safety and accountability – but scaling it up to govern an AGI will demand significant innovation and likely *integration with formal regulation*, not market forces alone.

## **Procurement Standards – Feasibility and Implications**

**Practicability: Procurement-based governance** leverages the spending power of large organizations to enforce standards. In practice, this requires those big players (especially governments) to develop and commit to **specific AI procurement guidelines**. The feasibility within 0–5 years is moderately high in certain sectors: we are already witnessing early moves. For example, some governments have started publishing **ethical AI procurement guidelines** for their agencies, and the defense sector has long had tech procurement rules (which can be adapted to AI). The authors cite that there’s an *“urgent demand for AI-specific public procurement standards outside of the defence industry.”* It’s realistic that within a few years, government bodies (national or local) will formalize procurement requirements like *“AI systems purchased must have undergone bias testing, include an audit report, and meet cybersecurity standards.”* The **UK, EU, and US** have all indicated interest in using government procurement to advance AI safety and ethics, since it doesn’t require new legislation – agencies can often set their own procurement conditions. Large corporations too (think of cloud service buyers or enterprise software purchasers) could band together to demand certain assurances from AI vendors. The main challenge is **coordination and standard-setting**: procurement officers need a clear set of metrics or checklists to evaluate AI offers. In the next five years, creating those standards is a work in progress (efforts by ISO, IEEE, or government working groups are underway to define AI risk management standards). So, by 2025-2030, we might have **published procurement standards** but their adoption will vary. Defense and critical infrastructure procurement will likely lead (due to high stakes), while general government IT procurement might lag unless mandated from the top. Private sector adoption depends on whether companies see value – e.g. a consortium of banks might all agree to not buy AI that isn’t explainable to regulators. Overall, procurement standards are **implementable in pockets** relatively quickly, but **universal procurement requirements for AI** are idealistic in five years (that would require many buyers worldwide to agree on the same standards). Instead, expect influential players (like DoD or EU government agencies) to pave the way, thereby indirectly influencing the market.

**Safety & Alignment:** The impact of procurement on **AI safety** can be significant when those standards are set with safety in mind. If a major government says “we will only procure AI systems that meet *specific safety criteria* (e.g. have passed adversarial testing, have an off-switch, comply with an AI risk management framework)”, this forces companies that want those contracts to incorporate safety from the design phase. For instance, the authors mention the **Pentagon’s CMMC cybersecurity requirements** as a model – in AI, a similar requirement could be, say, *“the system must allow human override and log all decisions for audit.”* This could push safety features

as a norm. **Value alignment** concerns (ensuring AI acts according to human values) could also be baked into procurement: a government might refuse to buy AI that hasn't been assessed for alignment with democratic values or human rights. Indirectly, procurement standards drive **competition on safety** – vendors that exceed the minimum might be favored. However, one limitation is that procurement typically affects **products delivered to the procurer**. For general-purpose AI models released to the public (like open-source models or consumer AI apps), procurement rules of a government might not influence them unless those developers also seek government business. Thus, procurement is a strong lever for specific contexts (e.g. AI in defense, or AI cloud services for government use) but not a catch-all for all AI safety issues. In the near term, we can expect safety gains in domains where government/large buyers are deeply involved. For example, any AI used in critical infrastructure might become safer due to strict procurement vetting (preventing, say, purchase of an AI system known to be unstable or unaligned). Additionally, procurement can encourage the development of **safety benchmarks** – if tenders require, for instance, a certain result on an “AI safety test suite,” companies will work to meet it, advancing the state-of-the-art. The authors draw a parallel to the **FDA's device approval**: it enforced a minimum safety/efficacy threshold, which did not stop innovation but channeled it into safer pathways. Similarly, procurement standards could ensure innovation in AI is *directed toward aligned AI* (because only aligned AI wins big contracts). For **AGI alignment**, procurement might have limited reach unless governments themselves become the primary funders or customers of AGI. If, say, a government or coalition declares they will not use or recognize any AGI that hasn't passed certain evaluations, that sets a norm, but it may not stop a private rogue project. Still, procurement could promote alignment research: large purchasers could demand evidence of alignment research investment from AGI developers as part of contracts, effectively channeling resources to safety.

**Accountability:** Procurement standards create accountability through **contractual obligation**. If a vendor fails to meet the agreed standard (for example, an AI system sold to the government malfunctions in a way that violates the contract specs), the vendor can face penalties, lose payment, or be blacklisted from future contracts. This is a strong motivator for companies to be accountable to the requirements. In contrast to regulation, which might punish after broad public harm is done, procurement contracts can be enforced more immediately and narrowly – if you don't deliver what you promised in terms of safety features, you breach the contract. In the near term, accountability via procurement is feasible where contracts are clear. We will likely see clauses like **“AI system must comply with X standard and the supplier will be liable for any deviation.”** Also, procurement can enforce **reporting obligations**: a contract might stipulate the AI provider must report any critical incident or near-miss to the buyer. This is analogous to how defense contracts often require reporting of test failures. Such measures keep AI developers on a short leash for those specific clients. However, procurement-driven accountability only works **for the parties in the contract**; it doesn't directly make a company accountable to the general public. For example, if a private AI product is unsafe but not sold to any major scrutinizing buyer, procurement standards won't catch it. Thus, as a governance tool, procurement targets accountability in *particular high-stakes relationships* (e.g. government and contractor). Within five years, one might see governments including clauses that effectively hold AI suppliers **financially and legally accountable for safety lapses** in delivered systems, which could set an industry precedent. Another aspect is that procurement standards can propagate accountability down the supply chain. If a prime contractor must answer to the government, it in turn will enforce

requirements on its subcontractors or component providers (a cascade effect the authors describe globally). This means even smaller AI component developers might become accountable to standards indirectly. For AGI, if governments intend to be significant users of AGI tech (for instance, for national security or large-scale infrastructure management), they can impose very stringent accountability measures via procurement (including perhaps oversight rights, audit rights, kill-switch requirements, etc.). It's likely that any **trusted AGI** used by government would go through a gauntlet of checks due to procurement conditions. But if a powerful AGI is developed outside these channels (e.g. by private labs for private use), procurement leverage might not apply, which is a governance gap.

**Regulatory Structures for AGI:** In a future with potential AGI, **procurement could be one tool in a larger regulatory structure.** Think of it this way: governments might decide that they will only **recognize or collaborate with AGI systems that meet certain safety criteria.** While that doesn't ban unsafe AGI globally, it creates a market incentive for alignment among those who want to do business with governments or large corporations. In five years, if we don't yet have AGI but possibly very advanced AI, governments could start trialing procurement requirements that simulate what AGI governance might need (for instance, requiring verifiable **alignment tests** for any AI above a certain capability used by the government). This could serve as a prototype for formal AGI regulations later. However, relying on procurement for something as pivotal as AGI might be **too idealistic**, because AGI development might not hinge on selling a product – it could be an end in itself or deployed directly by its creators. In such a case, more direct regulatory oversight (like licensing of large compute clusters or monitoring of AGI experiments) would be needed. Yet, procurement will still matter in the ecosystem: post-development, **whoever wants to deploy AGI widely (e.g. in industry) might need to meet the stringent demands of key customers.** To sum up, procurement standards are a **practical measure already starting** for narrow AI safety, and in five years we expect to see significant movement there. They bolster accountability and indirectly force alignment efforts for those systems within their scope. For the highest-risk AI (AGI), procurement alone is insufficient but is likely to be a *component* of a multi-faceted governance approach, ideally complemented by international agreements or formal regulations specific to AGI.

### **Investor Due Diligence – Feasibility and Implications**

**Practicability:** Harnessing **investor due diligence** for AI governance relies on changing how investors evaluate AI-related businesses. Culturally and practically, this is perhaps the hardest to guarantee, as it depends on investor awareness and priorities. In the near term, however, there are signs of change: with the surge of interest in AI, institutional investors are beginning to ask about **AI risk factors** (similar to how they consider cybersecurity or environmental risks). The fact that major asset managers have embraced **responsible investment principles** means there is at least a stated willingness to integrate AI ethics/safety into decision-making. Over the next five years, we might see **due diligence frameworks for AI** emerge – for example, checklists for venture capitalists when funding AI startups (covering questions of safety, regulatory compliance prospects, ethical implications). Some VC firms that specialize in AI or that are aware of alignment issues may already do this qualitatively. However, expecting **all investors to uniformly police AI risk** is idealistic. Many investors, especially in competitive markets, are driven by short-term returns. During an “AI boom,” the fear of missing out can override long-term risk concerns. Without

external pressure, a VC might invest in a highly profitable but somewhat risky AI venture, assuming they can exit before any negative consequences materialize (a form of moral hazard). The authors present due diligence as a “*self-regulating mechanism*”, which it can be, but it is voluntary. Feasibly, what can amplify it in five years is **regulatory nudges**: if regulators signal that AI risks are material (e.g. the SEC could require public companies to disclose AI-related risk management), then investors have more incentive to examine those disclosures. Also, if high-profile AI failures or scandals occur in the next few years, investors will take note and due diligence will tighten (much as the BP oil spill made energy investors focus on safety). In summary, improving investor due diligence is partially feasible – we’ll likely see a subset of conscientious investors lead the way, perhaps forming guidelines or coalitions (similar to ESG investor coalitions). But expecting the entire market to enforce AI governance is optimistic without either regulatory requirements or a clear business case. We should also consider that *different classes of investors behave differently*: venture capital (in startups) might be less risk-averse (willing to gamble on breakthrough AI) compared to pension funds investing in established companies (which might avoid companies that could face future regulatory troubles due to AI misuse). So the effect will be uneven in the near term.

**Safety & Alignment:** If investors rigorously evaluate AI risk, it can significantly contribute to safety. For one, due diligence might uncover whether a company has **invested in alignment research or has an internal safety team**. A well-informed investor could demand that a startup working on advanced AI allocate some budget to safety, or else they won’t invest. This directly encourages alignment efforts as part of doing business. Investors can also favor companies with **transparent and robust safety processes**, creating a competitive advantage for safety-conscious firms. Over five years, this could steer funding towards ventures that are “*safe by design*.” However, the depth of impact depends on investor expertise – many investors lack technical understanding of AI alignment and might rely on superficial indicators. There is a risk of **safety-washing**, where companies present an image of responsibility to appease investors without genuine rigor (somewhat akin to greenwashing in ESG). That said, large tech investors often hire experts or consult advisors; we may see the rise of **AI risk auditors specifically assisting investors** (merging with the auditing proposal). One positive trend is that investors hate uncertainty – and unaligned AI can produce a lot of uncertainty (unpredictable outcomes, regulatory backlash, public relation disasters). The authors note that due diligence helps in “**accurately pricing securities, reflecting all available information**”. If “all available information” includes alignment risk, then unsafe AI companies would get a risk discount (lower valuation). This market signal could deter recklessness. For actual **AGI safety**, investor diligence is a weak tool if the race for AGI is between a few well-capitalized entities who might be willing to take risks for massive rewards. It works better for **narrow AI** or general AI ethics issues like bias or privacy – investors can push companies to avoid those pitfalls to protect brand value. For extreme misalignment concerns (like AI that could become uncontrollable), many traditional investors might not even grasp the scenario enough to act, unless the discourse becomes mainstream. In five years, alignment issues might become more visible (through conferences, media, maybe minor incidents), potentially educating investors. But it’s optimistic to assume profit-driven markets will sufficiently account for low-probability existential risks; historically, markets under-invest in mitigating tail risks without regulatory prodding.

**Accountability:** Investor due diligence introduces accountability by tying corporate behavior to access to capital. Company executives become accountable to their investors (and prospective investors) for how they handle AI. If a CEO knows that reckless AI deployment will scare off investors or lower the stock price, that is a powerful incentive to be accountable for AI outcomes. We saw with BP's case that after the disaster, investors *compelled* changes in strategy and safety spending – essentially holding management accountable for improving risk management. In AI, if investors unify behind certain red lines (e.g. “we won't fund AI projects that violate privacy or that lack an ethical review board”), companies will have to answer to those expectations. In practice, however, investor pressure often comes *after* a crisis or *in reaction* to external factors. In the near term, if an AI-related scandal hits a prominent company (say a deep-learning system causes a major public harm), we could imagine its share price falling and investors demanding leadership changes or moratoria on that AI's use. That reaction would enforce accountability post hoc. Proactively, some investors might ask tough questions at board meetings or during funding rounds about AI safety. There is also the mechanism of **shareholder resolutions** in public companies – shareholders could vote to require the company to implement stronger AI governance. Such activism has precedent in areas like climate change. Near-term, only very socially-conscious or risk-averse investors are likely to go this far on AI, but a few instances could set examples. A limitation is that investors who prioritize quick gains might actually pressure companies in the opposite direction, i.e. “rush this AI product to market to beat competitors,” implicitly encouraging higher risk. This is why relying on investor due diligence alone is shaky; it's variable. It can create accountability in firms that engage with diligent, responsible investors, but companies can also seek capital from less scrupulous sources if mainstream investors are too demanding. In five years, with luck, responsible AI due diligence becomes a **norm among top-tier investors**, thereby channeling most reputable funding to aligned companies, while fringe projects struggle to find funding (except possibly from shady sources or governments). It's a scenario we can hope for but can't guarantee without aligning incentives (via regulation or cultural change).

**Regulatory Structures for AGI:** In the context of **AGI**, investor due diligence would mean that those funding AGI development (be it private investors, big tech shareholders, or government funders) impose conditions to ensure the AGI research is done safely. This could be crucial – for instance, if a venture capital fund is approached by a startup claiming it will build AGI in 3 years, will the VC consider global catastrophic risk and demand a safety plan? Ideally yes, but historically, **speculative tech bubbles** make investors less cautious. Governments might influence this by signaling how AGI will be regulated (if investors know a dangerous AGI project will be shut down by law, they won't fund it). In five years, direct AGI might still be hypothetical, but advanced AI projects (like beyond GPT-5 level) will attract big investments. Those investments will shape who leads in AGI. If market governance works, capital would shy away from any project that appears to ignore safety, thereby throttling irresponsible AGI development. However, **competition dynamics** complicate this: a nation or corporation might invest precisely in a high-risk high-reward AGI effort to leapfrog others, essentially tolerating the risk for potential dominance. This is where market logic can fail – the **race for AGI** might cause even normally prudent investors to gamble, hoping to be the winner and assume they can handle the fallout. Such scenarios suggest that **only relying on investor self-regulation is quite idealistic for AGI**, especially if international rivalries are at play (state-backed investments might not follow the same risk calculus as private investors concerned with profit). Therefore, while encouraging thorough due diligence is important,

**regulatory frameworks likely need to step in for AGI**, for example by requiring licenses or certain safety thresholds as prerequisites to even get funding or access large compute. Investor due diligence then becomes one layer: investors check that teams have that license or have done safety research as required, etc. In the absence of formal rules, we must rely on informal norms among the investor community – five years might be too short to solidify such norms globally, but initial efforts (workshops, guidelines for AI investors) could start now.

In conclusion, **investor due diligence as a governance tool** is partially realistic in the near term – some movement will happen, enhancing accountability and perhaps channeling funds toward safer AI ventures. It is not a uniformly reliable mechanism due to varied incentives, thus it should augment, not replace, **government regulation and collective agreements** when confronting AGI-level risks.

### **Idealism vs Realism and Implications for AGI**

Stepping back, it's clear that **all four market-based proposals have merits, yet also limitations, especially under the demanding scenario of AGI governance**. Within a five-year horizon, **auditing and procurement standards** seem the most **realistically implementable** and likely to show tangible progress. **Auditing** aligns with existing corporate practices and regulatory trends; we can expect to see audit frameworks and maybe even audit requirements for certain AI systems soon, which will improve transparency and safety for many narrow AI applications. **Procurement standards** can be instituted by forward-thinking governments or corporations relatively quickly, driving safer development in segments of the AI industry (particularly where public sector funding is involved). Both of these operate in a mostly **incremental fashion** – they don't revolutionize AI governance overnight, but they start embedding safety and accountability into everyday processes.

**Insurance and investor-driven governance**, on the other hand, lean more **idealistic in the short term**. Insurance has conceptual appeal but runs up against the reality that **actuarial models for AI are underdeveloped** and some AI risks might be essentially uninsurable within current market constraints (for example, no insurer will underwrite an AI that could potentially cause a trillion-dollar catastrophe, unless perhaps governments backstop it). We may see *piecemeal adoption* (e.g. including AI clauses in cyber insurance policies), but not a comprehensive risk-spreading system for AI by 2030. **Investor due diligence** is also idealistic unless bolstered by external expectations – while leading investors might incorporate AI risk, a significant portion will likely continue prioritizing growth over safety unless/until a major negative event or regulation forces a rethink. It's hard to quantify "alignment" in financial terms, so many investors could simply hope for the best or assume someone else (regulators, insurers, the companies themselves) is handling it.

For **AGI** – a hypothetical extremely advanced AI that could pose existential or transformative risks – these market mechanisms are at best partial measures. They each address certain aspects (financial incentives, information sharing, contractual control, capital allocation), which are valuable **pieces of a larger governance puzzle**. However, an **AGI capable of self-improvement or causing global-scale effects** may necessitate more direct oversight (e.g. a dedicated AGI regulatory authority, international monitoring, or strict licensing as has been proposed by some experts). Markets tend to falter in managing **very low-probability, high-impact risks** – exactly the

kind of risk AGI might pose. For example, an insurer cannot meaningfully price the risk of human extinction, and an investor might discount it entirely since it's beyond normal investment horizons. **Public-good concerns** (like preventing an AI catastrophe that affects everyone) often require collective action, which is where government regulation and international cooperation come in.

That said, one should not dismiss the market approaches as irrelevant for AGI: they can create a supportive environment that **raises the baseline of AI systems' safety** and can buy time by reducing nearer-term failures. If the industry through these mechanisms cultivates a culture of **responsible innovation**, those norms could carry over to any AGI project (for instance, an AGI team that has been subjected to rigorous audits and investor scrutiny all along is likely more safety-conscious than one operating in a laissez-faire environment). Market mechanisms can also spur **innovation in safety techniques** by making safety a competitive advantage. The authors even suggest that aligning market incentives with AI safety could **unlock funding for safety research** – for example, if insurers give discounts for provably interpretable models, suddenly there's money to be made in interpretability research. In a five-year view, we might see increased private investment into AI safety tools (auditing software, monitoring systems) because there's a market need. This indirectly benefits any future AGI oversight.

**In summary**, Tomei et al.'s market-based proposals offer **valuable additions to AI governance**: they inject flexible, incentive-driven approaches that can work alongside slower-moving formal regulations. Each proposal maps to a known governance function – insurance to liability management, auditing to standards enforcement, procurement to rule-setting via contracts, and due diligence to transparency and ethical investing – thereby augmenting or sometimes substituting for traditional government action. In the near term, **auditing and procurement** look practical and will likely strengthen AI safety and accountability soon. **Insurance and investor governance** are emerging but will need more time and perhaps regulatory nudges to reach their full potential. When it comes to **AGI or very advanced AI**, these mechanisms alone are **insufficient and somewhat idealistic** as primary safeguards within five years. They face issues of scale, knowledge, and enforceability in that context. However, they are still **crucial pieces**: they lay groundwork, address current AI challenges (narrow AI governance is itself important so we don't stumble before reaching AGI), and provide insights that can feed into formal **regulatory structures for AGI**. Policymakers and stakeholders should thus view these market approaches as **complementary tools** – realistic to deploy in the short term for incremental improvements, and instrumental in shaping a multi-layered governance regime needed for the long-term, safe trajectory of AI development.