

The AI Accountability Ecosystem in the Era of Large Language Models

Chris Percy* Artur d’Avila Garcez†

May 14, 2026

Abstract

This article reviews and updates the framework for accountability in AI based on *accountability ecosystems*. We update the framework in light of the latest developments since the release of Large Language Models for general public use. We propose three interlinked updates to the original AI accountability ecosystem: (i) reorienting the accountability ecosystem to AI infrastructure and supply chains, (ii) providing greater emphasis on outcomes monitoring and identification of issues that support decentralized system improvement, and (iii) incorporating end-user accountability given the new risks of unpredictability of language models *in-the-wild*. Collectively, these updates mark a shift towards accountability as distributed, continuous, and institutionalized, away from a system in which frontier AI applications can be modeled as discrete products controlled by single identifiable actors with industry-specific oversight.

1 Introduction

This article reviews and updates the framework for accountability in AI ecosystems¹ in light of developments in Large Language Models (LLMs) since ChatGPT’s public release more than 3 years ago. The framework for an “accountability ecosystem in AI” predated the generative AI revolution with LLMs. Nevertheless, it contained relevant predictions about today’s AI landscape. Written in 2021, the paper concludes with “if we collectively fail to make progress towards these [accountability] goals, towards measurable [explainability] and implementing these accountability mechanisms, we can look forward to immature regulation operating in an unbalanced ecosystem, being either ineffective or overwhelming, and continued frustration from a civil society that lacks constructive channels through which to direct their anger and to help shape AI practice”.

Five years ago, the infrastructure layer for AI was more benign, with risks concentrated more in application development than underlying infrastructure.

*University of Warwick, Coventry, CV4 7AL, UK

†City St George’s, University of London, London, EC1V 0HB, UK

Those building AI applications did draw on an infrastructure layer, but often limited to cloud services for processing their own data, standard libraries to implement well-understood algorithms, or broad-based online communication infrastructures. In the early 2020s, it was reasonable for an accountability ecosystem to be built around the technical team or company developing the application, even if multiple stakeholder layers and mechanisms were required to hold the developers adequately to account.

Today's world is very different. The era of Large Language Models created a centralized infrastructure layer driving AI applications that is far from benign. Frontier models are revised without forward-looking schedules, producing new capabilities and limitations that are only weakly understood by their owners, let alone their API-users. Design choices are more art than science, involving externally opaque decisions on training set corpora, algorithm optimization, and applicable AI principles. Three years ago, Reinforcement Learning with Human Feedback (RLHF) was touted as the main technological innovation behind ChatGPT. Since then, there has been a retreat from openness in AI research in industry, followed by confusion around the timeline for the achievement of Artificial General Intelligence (AGI).

With the above developments, the weaknesses in our collective AI accountability ecosystem remain all too apparent today. This article explores what a strong system might look like in today's context. We argue for three interlinked adjustments to the original ecosystem argument. First, reorient the accountability ecosystem around supply chains, rather than around the development team building a given application, so as to recognize the importance of assurance markets and benchmark validation across a delivery stack. Second, apply greater emphasis on outcomes monitoring and identification of issues by stakeholders in ways that support decentralized system improvement. Third, explicitly incorporate user accountability and safeguards, given the new risks raised by jail-breaking and system unpredictability. The remaining elements of the original accountability ecosystem framework continues to be valid, with an emphasis on a tiered approach to risk management, taking into account the contributions and interaction between corporate actors, their market counterparts, civil society and government. At the same time, stronger tools are necessary to deliver on the requirements of the original proposal, as we will explore in a brief discussion on explainable AI.

Collectively, the above three adjustments mark a shift towards accountability as distributed, continuous, and institutionalized, away from a system in which frontier AI applications can be modeled as discrete products controlled by single identifiable actors with industry-specific oversight. Where the most significant risks occur in infrastructure layers, the accountability approach might draw greater inspiration from such areas as communication network infrastructure or finance. Supplier licensing, model-level certification, independent safety bodies, and incident reporting systems have a larger role to play than internal and external audit functions directed at corporate API-users.

At the same time, the ecosystem perspective from 2021 remains key to the collective approach towards AI accountability: no single mechanism is sufficient

and societal layers all have a role to play, from technical teams and companies through to markets, civil societies, and governments. Bias, explainability, transparency, and informed participation also remain essential principles underpinning accountability, now magnified by issues around large-scale data security and privacy. The focus on operationalization from 2021 becomes even more important as AI technologies are already too diffusely embedded for adherence to principles to be trusted or directly monitored.² Converting principles into practice continues to be important as proposed with the use of nested operational procedures: decision process documentation, internal audit, external audit, external accreditation.¹

This article does not question the current and potential contributions of LLMs, nor do we argue for turning back the clock. We acknowledge the value of LLMs as a very impressive engineering achievement with great potential impact to promote productivity gains across the global economy. The question is how to unlock these contributions in a safe and sustainable way. In what follows, we take the three proposed interlinked adjustments to the AI accountability ecosystem framework in turn, beginning by setting out a selection of the distinctive issues introduced by the LLM era and summarizing the AI ecosystem approach upon which this work builds.

2 New Accountability Issues in the LLM era

In an unregulated area of knowledge such as computer systems, or in a marketplace without accountability, a system like ChatGPT can be made available to billions of users for free without stringent quality checks, ethical certification, or reliability guarantees. In some corners, this is seen as creating a healthy disruption in the market. Google certainly felt that disruption when OpenAI released ChatGPT in late 2022. The ease with which internet-style searches were enabled by ChatGPT via system interactions in natural language put Google’s keyword-driven internet advertising business model in jeopardy. Apparently, OpenAI’s goal was to test their new system, collecting vast amounts of user data to refine their models. What followed was a BigTech race to the bottom with companies competing fiercely for user engagement, lawsuits about copyright violation, and a normalization of the idea that AI will *make stuff up*, in some cases with serious consequences.³

2.1 *LLM hallucinations have not gone away*

Imperfect reliability is a fact of life outside the self-contained world of axiomatic proofs and high-assurance industries. Complex socio-technical systems are very hard to verify formally and humans make mistakes. However, incentives, oversight structures, and the use of formal methods have been shown to promote error correction processes that can be deployed to improve reliability with costs proportionate to a given use case.⁴ In cases with high impact errors, such as air travel or infrastructure security, multiple layers of automated and human

oversight reduce mistakes to acceptable levels, where what counts as acceptable is continually negotiated between regulators, stakeholders, and users.

Today’s LLM architectures have not yet mastered this proportionality. ‘Hallucinations’ have reduced in recent years, but it is not yet possible to invest in larger AI infrastructure and reduce errors to a pre-determined level. The only current way to achieve this is via significant human-in-the-loop checks, which largely undo the productivity gains of AI and place unrealistic pressure on human monitors. A human finding errors in a stream of AI output that is tailored to convert unreliable source text into professional-sounding language is tantamount to looking for discolored pixels in a photo. Human cognition is not optimized for this task.

Unfortunately, the stakes of LLM hallucinations are rather higher than discolored pixels and likely to increase as more individuals rely on LLMs with increasing autonomy, so-called Agentic AI, to support their decision-making.⁵ LLM’s worldwide deployment and social media share the same business model predicated on the false premise that software is free or low cost when in fact users pay with their data or with the promise to investors that high usage numbers will yield world-changing profits in the future. Back in 2021, when referring to the *Testimony from a Facebook Whistleblower*⁶ to the US Senate, we stated it helps “highlight serious concerns of influence on topics such as the mental health of youth exposed to social media platforms, and the use of private data to manipulate information leading up to addiction to the social media platforms.”¹ Since then, we saw the CEO of Meta apologize in another US congress hearing to the parents of children who took their lives influenced by social media. The problem is now urgent and compounded by LLM slop, where a recent court ruling found Meta and YouTube guilty of building addictive social media platforms that harmed a 20-year old’s mental health.⁷

2.2 *Humans are in the loop, but disempowered*

Take for example the idea of using various LLMs in parallel to scale up the job of therapy advice, having a human therapist in hand who is alerted to *take over* whenever an LLM produces an output from a pre-defined list of red-flagged outputs. As in the case of a self-driving car where the human “driver” is expected to intervene to avoid an accident, having liability in the case of an accident, despite *automation complacency* (as drivers gain confidence in the system, their attention to the driving task diminishes), here too the therapist isn’t empowered but is made responsible when things go wrong. *Human-in-the-loop* only works when the human is empowered by the interaction with the AI system and can intervene meaningfully in the decision-making process that is partially automated.

The seductive power of LLM capabilities introduces further risks of disempowerment. From students to politicians, LLM users are increasingly passing off AI-generated content as their own. European Union (EU) regulation requires developers to declare the role of AI in their tools. That does not apply to anyone using AI informally. How can we ensure, as an alternative to regula-

tion, that people take responsibility for their decisions when they increasingly lean on AI to formulate them? Furthermore, how can we make individuals or organizations accountable in this new *AI informal economy*, where productive gains derive from employees augmenting their work with LLMs in ways that are undisclosed or untracked?

The highest risk (and arguably the highest opportunities) in AI are in Agentic AI, that is, LLMs with autonomy to take actions such as executing auto-generated code to collect and process data automatically to improve its own performance. At the same time, data quality has become a very valuable asset when it comes to fine-tuning or training domain-specific AI systems. In practical terms, Agentic AI poses a serious cyber-security risk to computer systems, with much of the latest research being focused on neuro-symbolic systems capable of providing assurances to various LLM outputs, e.g. by combining neural networks with symbolic theorem provers for the purpose of auto-generated code verification.⁸

2.3 Systemic implications are far reaching now

General purpose AI models have systemic implications that are less easy to predict and manage than the narrow, domain specific applications that dominated the landscape in the late 2010s. Where models were largely developed within an industry or use case, such as healthcare, finance, or retail, the boundaries of those organizations and industries largely curtailed spillover consequences. Systemic implications today are far reaching, connecting into geopolitics and international development. The AI race, principally between the US and China, strains trade relations and introduces concerns about data sovereignty. The increased role of drones in modern warfare exacerbates pressures for autonomous AI capabilities. An ongoing dispute between the company behind the *Claude* LLM and the US Department of War, risking a US\$200M contract, relates to whether the company or the government as its client should dictate the terms of use of Claude when it comes to safeguards preventing use for mass surveillance.

The economic disruption of AI is being increasingly analyzed by governments, with particular concerns about LLMs replacing junior level staff, disrupting the lifecycle of expertise transfer and skills ladders. As pointed out, when OpenAI released ChatGPT, RLHF was promoted widely by the company's leadership as being its major technical innovation. We now know that RLHF involved exploiting, for the purpose of model alignment, a very low paid global workforce to perform data labeling watching the worst possible content on the internet for many hours a day.⁹

Increasing levels of autonomy in AI systems with access to bank accounts, emails and calendars introduce the possibility of machines operating as independent actors in society and a range of risks that would seem outlandish if it were not for the rapid progress of recent years. A small but fast growing set of AI users are calling for independent rights and protections for AI systems, often grounded in a belief that AI systems are conscious or should be considered moral actors given their apparent agency. What will happen when these

demands grow in volume and become the focus of political campaigns?

3 Implications for the AI Accountability Ecosystem

The AI accountability ecosystem¹ argued that the then-accountability ecosystem was heavily imbalanced, being a factor in its inability to translate intensifying civil society concerns into productive changes and engender more positive attitudes towards AI. The then-ecosystem was skewed towards high-level principles (over 100 at the time) and disempowered tag-on external accountability mechanisms (such as senior boards of external observers). These external mechanisms proved unable to drive change internally, so those involved raised high-profile public concerns instead, further alienating internal processes. An explicit ecosystem approach was suggested as a way to correct these issues, with attention to developing multiple layers of interaction from actors and companies through to market counterparts, civil society, and governments.

Many of the same issues apply even more intensively now than five years ago. The traditional strong arm of modern externally-led accountability seems to matter little in the face of LLM’s technical, financial, and geopolitical reputation. In principle, society’s strongest accountability systems are independent members on corporate boards, government hearings, and political oversight. This does not seem to work in the case of disruptive AI, as best described by a former OpenAI board member.⁹

Next, we argue that AI ethics cannot be an afterthought. It needs to be a requirement considered from the start, evolving with model development, properly documented, stress-tested and ideally formally verified, a requirement to be achieved by design. Nevertheless, the AI race is on. While an AI regulatory agenda promoted some coordination around AI safety, it didn’t seem to have produced any concrete outcomes, and the risks to computer systems’ security has increased. We will discuss how an updated AI accountability ecosystem should address each of these problems.

4 Enhancing the AI Accountability Ecosystem

Three interlinked adjustments can make the original ecosystem framework fit for purpose in the LLM era: supply chain orientation, outcomes monitoring, and user accountability. The revised framework is summarized in Figure 1.

4.1 *Supply chain re-orientation*

In the era of LLMs, centering accountability on a bounded *actor* (a product or ML team) as the unit of technical agency is no longer adequate. Instead, accountability should be reoriented to multi-layered AI supply chains.¹⁰ The core supply chain has interactions among three sets of actors: frontier model

developers, application developers, and users. While some AI applications are still built from scratch, an increasing proportion either require or incorporate LLM-related functionality from frontier labs. Application developers were the central *actor* in the original proposal, but the core supply chain perspective translates responsibility into a distributed and interdependent phenomenon.

An important set of actors shapes and brings a degree of transparency to the products and behavior that emerge from this core supply chain: assessment actors. These assessment actors include those dedicated to AI developers, such as benchmark designers and promoters and red-teaming providers, as well as pre-existing solutions that can be repurposed for AI, such as external and internal audit providers and risk management consultants. Market actors continue to play an important role in holding core supply chain actors to account, ranging from investors, creditors, and insurance providers through to employees and trade bodies. This shift enables new forms of non-government accountability through assurance markets and standardized validation practices. For example, upstream actors can provide model documentation, eval results, and safety guarantees, which downstream actors rely on contractually. Independent auditors and benchmark providers can validate claims (e.g. robustness, bias) at different points in the stack. In this sense, accountability is composable: each layer is accountable for its contribution, and assurance propagates through the system via documentation, certification, and contractual obligations.

This approach aligns with emerging regulatory thinking (e.g. “value chain responsibility” in the EU AI Act), but is particularly powerful in non-regulatory contexts, where procurement requirements, platform policies, and third-party audits can enforce standards without direct state intervention. It also mirrors mature accountability systems in other domains (e.g. financial auditing, supply chain certification), suggesting a plausible pathway for scaling AI governance beyond firm-level self-regulation.

4.2 *Monitoring Outcomes*

In the LLM era, accountability must be grounded in continuous, post-deployment outcomes monitoring. This reflects the reality that many harms such as hallucinations, misuse, emergent behavior, are not predictable at development time. As a result, accountability systems must incorporate mechanisms for monitoring evolving requirements, detecting, reporting, and responding to failures often identified by users, civil society, or downstream integrators, as well as commissioned or proactive red-teaming investigations.

Recent literature also supports a shift toward continuous, post-deployment accountability grounded in observed outcomes. A central finding in the International AI Safety Report 2026 is the existence of an *evaluation gap*: systems often perform well in controlled pre-deployment testing but behave unpredictably in real-world contexts. This gap provides strong empirical reason to move beyond static audits to ongoing monitoring and feedback loops.¹¹

At the same time, policy and industry frameworks are converging on risk-tiered approaches. The International AI Safety Report and related governance

initiatives emphasize that AI risks vary widely, from low-level inaccuracies to systemic or catastrophic harms, and that oversight mechanisms should scale accordingly. In parallel, technical and organizational frameworks increasingly emphasize risk classification, proportional controls, and adaptive mitigation strategies that should be mindful of the evolving capabilities and specific use cases of AI applications.¹²

Outcomes and the associated risk-level assessment and risk management processes need to be monitored at all levels of the ecosystem. Frontier labs, developers, and users, as well as various assessment providers, market actors, and government systems, are all potential loci for risk origination or observation. The ecosystem perspective identifies a broad range of relevant actors. Frontier labs might have the most rapid reaction speeds and direct visibility/control in most cases - and do bear important responsibility - but others cannot slough off all their duties to them. Outcomes monitoring across these layers would ideally leverage a mixture of formal systems that are mandated and funded to monitor outcomes, alongside informal, civil society and volunteer systems.

4.3 *User Accountability*

In today’s world, the end user or customer is also a focal point for accountability. Model users today are active participants in shaping AI system behavior, particularly in the context of LLM *Chain of Thought* prompting. Recent research about AI risk highlights the growing importance of misuse, adversarial prompting, and interactive vulnerabilities, including scams, manipulation, and cyber-attacks facilitated by general-purpose models.¹³

This creates a new class of accountability challenges that cannot be addressed solely through developer-side controls. Instead, accountability must be shared between system providers and users, supported by a combination of technical safeguards (e.g. formal verification, alignment techniques, refusal of behaviors), platform governance (e.g. usage policies, enforcement), and user-facing interventions (e.g. warnings, time-out, education and friction). In some contexts, contractual or legal mechanisms may also be used to assign responsibility for misuse. We cannot continue with terms of service requirements that are rarely read, constantly changing, and even more rarely monitored, let alone enforced.

Importantly, incorporating user accountability does not imply shifting blame away from developers, but rather recognizing that AI systems are interactive and co-produced in use, especially Agentic AI. Indeed, suitably autonomous Agentic AI systems will also need to be held accountable similarly to other users, potentially in line with pragmatic views of AI “personhood” and metrics that quantify the robustness of AI agent identity.¹⁴ This requires designing systems that are robust to misuse while also establishing clear norms and consequences for harmful user behavior. It also reinforces the importance of monitoring and feedback loops, since users are often the first to discover system vulnerabilities.

5 An illustration of the updated ecosystem

We illustrate the importance of the three adjustments by comparing the original ecosystem with the adjusted ecosystem in a hypothetical scenario. Suppose a system is procured to support the identification of potential security risks through the analysis of publicly available communications. The system was not developed by a single actor but assembled across a delivery chain comprising a frontier model provider, a platform or API intermediary, a systems integrator, and a deploying agency. The resulting system is therefore best understood not as a bounded product but a socio-technical collaboration spanning multiple organizations and layers of control. Over time, several concerns emerge. First, the system is gradually repurposed beyond its original remit, expanding from targeted threat detection to broader population-level monitoring. Second, analysts discover that certain prompting strategies enable the model to generate inferences about individuals' political affiliations or other sensitive attributes. Third, some users begin to develop informal workarounds that circumvent built-in safeguards, effectively *jail-breaking* the system to produce outputs that would otherwise be restricted.

Traditionally, accountability would be achieved through improved transparency, clearer documentation, and the progressive layering of internal and external audit mechanisms. However, here, the behaviors of concern emerge from interactions across the supply chain and system usage. As a result, internal audit mechanisms may lack visibility over upstream model behavior or downstream user practices, documentation may not capture the full range of emergent use cases, external scrutiny may operate with limited access to the system. The likely outcome is a delayed and partial visibility of issues and a reactive pattern of accountability, characterized by incremental rather than systemic remediation.

By contrast, an accountability ecosystem reoriented around supply chains introduces a different set of mechanisms and, correspondingly, different dynamics of response. Responsibilities should be specified across the delivery stack. These responsibilities are instantiated through contractual arrangements, benchmark validation, and assurance processes that enable claims to be tested and compared across actors. In this hypothetical case, an anomalous usage pattern indicative of population-level monitoring may be detected at the platform layer, known limitations in inferring sensitive attributes may have been documented by the model provider, and deviations from intended use may be identified by the integrator. Rather than a single actor bearing diffuse responsibility, accountability is distributed but structured, ideally to allow targeted interventions at the appropriate layers.

The incorporation of continuous outcomes monitoring further differentiates the updated ecosystem from the original formulation. The updated approach places greater weight on the systematic observation of real-world system behavior. While it is challenging to engineer, we can imagine that signals from internal logs, user interactions, and external stakeholders might be aggregated and assessed through a monitoring function that supports tiered risk management. In

the original framework, users are primarily situated within the market or civil society layers, with limited emphasis on their role shaping system behavior. In contrast, the updated ecosystem recognizes that, in the context of LLMs, system outputs are co-produced through interaction. Adversarial prompting and the circumvention of safeguards are therefore not exceptional events but foreseeable behaviors that must be incorporated into accountability design. Mechanisms such as usage monitoring and access controls operate alongside technical safeguards to address the risks. Accountability thus shifts from a predominantly procedural exercise to a dynamic, evidence-based process.

Taken together, these differences in framing: distributed responsibility across the supply chain, continuous outcome monitoring, and explicit user accountability, lead to materially different patterns of accountability. None of this is straightforward to design, implement, or oversee, but having a vision for the accountability ecosystem to aim for is a prerequisite to success.

6 On Explainability, Transparency, and Neurosymbolic AI

The accountability ecosystem proposal argued and exemplified the importance of explainable AI (XAI) to increasing reliability and trust in neural networks. How does this requirement fare five years on? Although XAI continues to be relevant to improving performance and trust in industry-specific AI, achieving explainability for entire LLMs, known as global XAI, is a very difficult, if not impossible task. This is because of the current scale of LLMs with billions of trained parameters (network weights). Efforts continue, however, trying to achieve local XAI and mechanistic interpretability (explaining individual cases) as well as complex network *distillation* into simpler models, as was done by the Chinese DeepSeek LLM.¹⁵ The sector widely recognizes the lack of sustainability in continuing with the current scaling approach at the rate of the past five years and with diminishing returns.

Research in the area of neurosymbolic AI continues to make progress as an alternative to the current scaling approach, advocating the use of something known as the neurosymbolic cycle of AI system development: train a neural network with some available data, extract symbolic knowledge from the trained network using XAI, reason about what has been learned evaluating results before further training with more data. This modular and iterative approach promotes domain expert interaction with the system during training, the possibility of user intervention to fix mistakes or improve performance, formal reasoning and validation when descriptions are extracted from the system, parsimony instead of massive scale when knowledge is consolidated and reused across applications.¹¹

Using LLMs alongside formal proof assistant software or knowledge graph queries are steps in the direction of neurosymbolic integration via the application of the neurosymbolic cycle. The use of distillation in the deployment of the DeepSeek LLM, instigated apparently by a lack of access to the latest graphics

processing hardware in China, if used as a form of knowledge distillation for the sake of explainability, is also a step in the direction of a more parsimonious model with training and data efficiency, as advocated by neurosymbolic AI.

Neurosymbolic AI is an important tool in the broader toolkit to address the transparency challenges that come with today’s LLM-enabled ecosystem. Standard approaches to software documentation are a problem with the number and complexity of AI-powered software. New approaches will be needed, such as the continuous assurance framework *Audit-as-Code* that seeks to map governance requirements to auditable rules.¹⁶ The latest AI index report 2026 confirms our predictions: “AI models are achieving breakthrough results in science and complex reasoning, but at a concerning environmental toll. Today’s most capable modern models are now among the least transparent. Meanwhile, AI’s workforce disruption has moved from prediction to reality, hitting young workers first”.¹⁷

7 Conclusion and Future Work

This paper has argued that the accountability ecosystem for AI, as originally conceived in the context of machine learning prior to the release of LLMs to the general public, continues to be relevant but requires substantive adaptation to remain effective in the era of large language models. The general-purpose nature and widespread use of LLMs, their deployment through complex and multi-layered supply chains, and their susceptibility to emergent and interaction-driven behaviors challenge the assumption that accountability can be centered on a single organizational actor or addressed primarily through *ex ante* processes.

We have argued that the accountability ecosystem approach continues to be relevant because AI cannot be seen as a *ladder* from sub-human to human and super-human intelligence, but is rather an entire ecosystem of human and machine interaction, ideally in productive cooperation. As a result, accountability in AI needs to take this ecosystem into consideration, an idea that is aligned with the recently proposed *Copernican view of AI*.¹⁸ However, the original accountability ecosystem framework from 2021 requires updating to consider the new reality of the era of LLMs, with the amazing speed and scale of AI technology deployment. We therefore proposed three interrelated adjustments to the accountability ecosystem that should make it fit for purpose in the era of LLMs.

First, accountability should be reoriented around the AI supply chain, recognizing that system behavior is co-produced across multiple layers of development, integration, and deployment. This shift enables a more granular allocation of responsibility and supports the development of assurance mechanisms such as benchmarking, auditing, certification, and contractual controls that operate across organizational boundaries. Second, greater emphasis should be placed on outcomes monitoring and tiered risk management, reflecting the limits of anticipatory governance in the face of uncertain and evolving system behavior. Continuous observation of real-world performance, coupled with structured es-

calation pathways, allows accountability to be grounded in evidence and adapted over time. Third, the ecosystem should explicitly incorporate user accountability and safeguards, acknowledging that despite AI technology continuing to be centralized in BigTech infrastructure, AI’s impact is decentralized and, hence, AI accountability mechanisms should see users as active participants in shaping system outputs.

The hypothetical case presented in this paper illustrates how these adjustments can lead to materially different accountability dynamics. Under an actor-centric model, the distributed nature of responsibility and the opacity of system behavior create conditions for delayed detection, contested attribution, and reactive intervention. By contrast, an ecosystem structured around supply chain accountability, continuous monitoring, and user-level controls enables earlier identification of issues, clearer allocation of responsibility, and more coordinated responses across system components. While such an approach does not eliminate the underlying risks associated with large language models, it provides a framework within which those risks can be more effectively managed.

From a policy perspective, these findings highlight the increasing importance of non-regulatory and hybrid governance mechanisms. In particular, assurance markets comprising independent auditors, benchmark providers, certification bodies, and risk assessment services offer a pathway for translating high-level principles into operational practices without requiring immediate formal regulation. Similarly, procurement standards, contractual requirements, insurance mechanisms, and platform-level controls can exert significant influence over system design and deployment, particularly where they are embedded within supply chain relationships. These mechanisms are often more adaptive and responsive than formal regulation, and may be better suited to the pace and uncertainty associated with frontier AI development.

At the same time, the effectiveness of such non-regulatory interventions depends on a number of enabling conditions. These include the availability of credible and standardized evaluation methods, sufficient transparency and information-sharing across the supply chain, the alignment of incentives among developers, deployers of AI technology and users, and the presence of institutional actors capable of aggregating and acting on signals from monitoring systems. In the absence of these conditions, there is a risk that non-regulatory mechanisms become fragmented, performative, or captured by dominant actors. Moreover, certain categories of risk, particularly those relating to systemic harms, national security, or fundamental rights, may ultimately require formal regulatory intervention to ensure online safety, consistency and enforceability.

Future work should explore how the accountability ecosystem can be operationalized, in particular: how accountability signals can be standardized and shared across actors, and how the balance between private and public forms of governance, regulation and non-regulatory mechanisms, evolve as AI systems continue to increase in capability and societal impact. The framework presented here highlights how accountability itself adapts in response to fast, disruptive technological changes. If we can get the accountability ecosystem right, the LLM era and Agentic AI might be just the next layer of computation abstrac-

tion, allowing users to focus efforts where it matters most, the specification of systems' requirements. Computation abstractions have taken us from hand-welded hardware to assembly language, from early programming languages to object-orientation, and dynamic typing in recent scripting languages that leverage massive standard libraries. Each step has enabled a phase transition in both the number of developers and the kinds of solutions they can generate. With the right technical innovations, iterative automated workflows, and accountability mechanisms, the power of AI-generated code may become sustainably and safely available to a new generation of creatives, larger and faster than ever before.

8 Acknowledgements

The authors would like to thank Simo Dragicevic for initiating the thinking on AI accountability ecosystems within the gambling domain.

References

- [1] C. Percy, S. Dragicevic, S. Sarkar and A. d'Avila Garcez, Accountability in AI: From principles to industry-specific accreditation, *AI Communications*, 34(3):181–196, 2021, IOS Press.
- [2] C. Canca (2020). Computing Ethics: Operationalizing AI Ethics Principles. *Commun. ACM* 63, 12 (December 2020), 18–21. DOI:<https://doi.org/10.1145/3430368>.
- [3] ChatGPT falsely accuses legal academic of sexual misconduct, *Scottish Legal News*, 11 Apr 2023. <https://www.scottishlegal.com/articles/us-chatgpt-falsely-accuses-legal-academic-of-sexual-misconduct>.
- [4] How Neuro-Symbolic AI Breaks the Limits of LLMs: combining neural learning with symbolic reasoning for more reliable AI. *Wired*, 10 Dec 2025. <https://www.wired.com/sponsored/story/how-neuro-symbolic-ai-breaks-the-limits-of-llms/>.
- [5] L. B. Eliot (2026, January 29). Why agentic AI bot swarms could discombobulate widespread human mental health and crash democracy. *Forbes*. <https://www.forbes.com/sites/lanceeliot/2026/01/29/why-agentic-ai-bot-swarms-could-discombobulate-widespread-human-mental-health-and-crash-democracy/>.
- [6] Protecting Kids Online: Testimony from a Facebook Whistleblower. U.S. Senate Subcommittee on Consumer Protection, Product Safety, and Data Security Hearing. <https://www.commerce.senate.gov/2021/10/protecting%20kids%20online:%20testimony%20from%20a%20facebook%20whistleblower>, 5 Oct 2021.

- [7] Campaigners welcome Meta and YouTube’s defeat in landmark social media addiction trial, BBC News, 27 March 2026. <https://www.bbc.co.uk/news/articles/c747x7gz249o>.
- [8] A. d’Avila Garcez (2026, February 12). Neurosymbolic AI: towards sound reasoning and causal learning and the road to AGI. City St George’s, University of London. <https://openaccess.city.ac.uk/id/eprint/36826/1/NeSyAGI.pdf>.
- [9] A. Imbrie, O. Daniels, H. Toner, Decoding Intentions, <https://cset.georgetown.edu/publication/decoding-intentions/>, October 2023, Center for Security and Emerging Technology [Online; accessed 20-Jan-2025].
- [10] J. Cobbe, M. Veale and J. Singh. Understanding accountability in algorithmic supply chains. In Proc. FAccT23, ACM Conference on Fairness, Accountability and Transparency, pages 1186–1197, 2023. <http://dx.doi.org/10.1145/3593013.3594073>.
- [11] Y. Bengio et al. International AI Safety Report 2026, DSIT 2026/001, 2026. <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>.
- [12] AI Risk Management Framework (AI RMF 1.0), The National Institute of Standards and Technology (NIST), USA, Jan 2023. <https://www.nist.gov/itl/ai-risk-management-framework>.
- [13] Hagendorff, T., Derner, E. and Oliver, N. Large reasoning models are autonomous jailbreak agents. Nat Commun 17, 1435 (2026). <https://doi.org/10.1038/s41467-026-69010-1>
- [14] E. Perrier and M. T. Bennett (2025). Agent Identity Evals: Measuring Agentic Identity. arXiv preprint arXiv:2507.17257. <https://arxiv.org/abs/2507.17257>.
- [15] D. Guo et al. DeepSeek-R1: incentivizes reasoning in LLMs through reinforcement learning, Nature 645(8081):633–638, 2025. <http://dx.doi.org/10.1038/s41586-025-09422-z>.
- [16] A. E. Muhammad, K. C. Yow, S. Alsenan. Audit-as-code: a policy-as-code framework for continuous AI assurance. Front Artif Intell. 2026. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12979488/>.
- [17] The AI Index 2026 Annual Report, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2026. <https://hai.stanford.edu/ai-index/2026-ai-index-report>.
- [18] T. Klowden and T. Tao, Mathematical methods and human thought in the age of AI. arXiv preprint arXiv:2603.26524, March 2026. <https://arxiv.org/abs/2603.26524>.

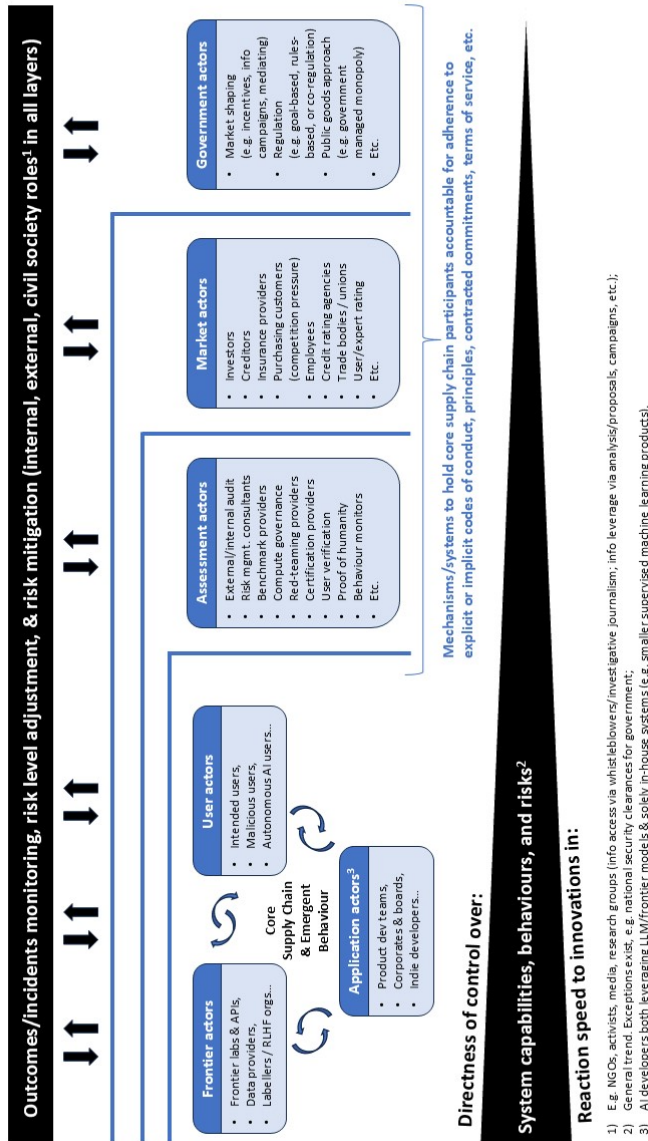


Figure 1: The Accountability Ecosystem fit for LLMs (simplified schematic for a mixed market economy in a democracy). The framework identifies the core supply chain for AI systems as encompassing three sets of actors: frontier model developers, application developers, and users. Those involved in model assessment, market participants, and government actors form three consecutive layers who interact with the core supply chain, having access to multiple accountability mechanisms and systems to hold the core supply chain to account, but with increasingly reduced reaction speed and directness of control. Outcomes need to be monitored and risks managed for all six sets of actors in the AI ecosystem.