SPRITE+ Explainer #001

Series Editor: Mark Elliot

Beyond Principles: Adaptive Governance for AI's Next Frontier

By Dr. Soraya Kouadri Mostéfaoui, Dr. Peter Winter, Dr. Zeba Khanam, Dr. Edward Chuah and Prof Lijun Shang.

This explainer focuses on AI Governance and its relationship to TIPSS (Trust, Identity, Privacy, Security, and Safety). It opens with the key challenges in current governance frameworks – drawing on the European Union (EU) AI Act, the UK's Pro-Innovation Approach to Regulation, and the United States' emerging guidance. It then examines critical TIPSS issues before concluding with likely developments and future directions in AI Governance.

Challenges with existing governance frameworks

AI ethics and governance frameworks exist in a state of productive (and sometimes counterproductive) tension. They are valuable – even essential in providing foundational principles and shared vocabulary for addressing issues such as algorithmic bias, privacy, and accountability and yet they seem inadequate in the rapidly evolving AI landscape.

Governments and organisations worldwide have developed frameworks to guide responsible development, to offer companies a moral compass to navigate complex terrain and build' build public trust. Yet, when these frameworks encounter real-world implementation, their limitations become apparent. Principles such as "fairness" and "transparency" are compelling in the abstract but difficult to operationalise in practice. What does fair treatment mean across diverse demographic groups with different needs and histories of disadvantage? How transparent should algorithmic decision-making be without undermining intellectual property, commercial advantage, or national security? These practical dilemmas expose the persistent gap between high-level ethical ideals and the messy reality of design and deployment. The global AI regulatory landscape compounds this complexity. The EU's *AI Act* (2024) [1] the UK's *Pro-Innovation Approach to AI Regulation* (2023) [2], *China's Algorithmic Recommendation Rules* (2021) [3], and emerging US guidelines (e.g., *Blueprint for an AI Bill of Rights*, 2022) [4] all reflect different strategic orientations: the EU foregrounds precaution and risk management, the UK emphasises flexibility and innovation, China focuses on social stability and control, and the US leans on standards and voluntary guidance. This heterogeneity creates compliance challenges and encourages "box-ticking" behaviour rather than deeper ethical engagement. Developers may meet regulatory requirements while neglecting broader issues of power: AI systems often entrench asymmetries between those who design and control technologies and those who are subject to their outputs, reinforcing inequalities and creating new forms of digital dependency and colonialism [5]

Adding to the challenge, AI development consistently outpaces governance. By the time comprehensive frameworks are finalised, the underlying technologies may have moved on - Large Language Models, Generative AI, and increasingly autonomous systems already stretch ethical and regulatory categories designed only a few years ago. Governance that relies on static rules risks being obsolete before it takes effect.

Recognising these limitations, governments and organisations are experimenting with more adaptive, risk-based, and participatory approaches. For example, the *EU AI Act's* [1] tiered system maps obligations to levels of risk and includes mechanisms for revision. The UK has piloted AI regulatory sandboxes, enabling safe testing of AI applications under regulatory supervision. In the US, the National Institute of Standards and Technology (NIST) has developed an AI Risk Management Framework (2023) [6] that emphasises iterative cycles of mapping, measuring, and managing risks. These approaches embody a shift from static principles to dynamic processes, recognising that ethical governance is ongoing rather than a one-off achievement.

The autonomous systems field offers some of the clearest evidence of this shift in practice. Self-driving vehicles cannot be regulated by fixed standards alone, because they face open-ended, unpredictable real-world environments. Instead, regulators have adopted adaptive models:

- In the UK, the *Code of Practice for Automated Vehicle Trialling* [7] (first issued in 2015 but frequently updated) provides a living framework that evolves alongside technological advances and safety evidence. The Code of Practice ultimately provided the empirical grounding for *The Automated Vehicles Act* (2024) [8] the legal framework informed by that trialling (e.g., AV pilots).
- In the US, states such as California have used phased licensing for AVs, moving gradually from closed-track tests to limited public road deployment as safety data accumulates [9].
- At the international level, UNECE (2024) [10] has advanced scenario-based safety testing, an approach that recognises the impossibility of prescribing a complete checklist of conditions in advance.

These examples show how principled commitments to safety, accountability, and transparency can be combined with the flexibility to adapt as new risks, contexts, and social expectations emerge. Adaptive governance does not abandon principles - it operationalises them through review, real-world testing, and participatory oversight.

Critical issues for TIPSS

Trust within organisations, and between employees, leadership, and AI systems, is foundational to productivity and morale [11; 12]. Yet, workplace surveys in the UK show that trust levels are alarmingly low, with fewer than 20% of employees trusting senior management - a trend exacerbated by fears that Generative AI might displace jobs [13; 14]. An influential analysis estimates that a rapid transition to autonomous vehicles alone could eliminate more than 1.2 million driving-related jobs, including delivery, bus, truck, and taxi drivers in UK alone [13]. These anxieties have shaken trust in employers, creating a need for a regulatory governance framework that can support social adaptation. This underlines the importance of transparent and inclusive regulations to sustain organisational trust in the face of technological disruption. Safety concerns extend beyond workplaces: Agentic AI -

systems that make autonomous decisions - raise critical questions about accountability. Some important elements in governance models might include real-time auditing to detect anomalies as they occur, secure authentication to prevent unauthorised access, and fail-safe mechanisms to mitigate system failures—all essential for upholding trust and safety (for example, NIST's AI Risk Management Framework, 2023 [6]).

Moreover, multimodal AI systems that integrate text, image, audio, and sensor data can infer highly sensitive personal information from benign inputs. This erosion of privacy and identity - through deepfakes, surveillance, or biometric inference - requires governance that is both flexible, responsive and rights-based. allowing us to be anticipatory and adaptive to whatever happens with AI, rather than merely responding to the existing technology.

In the UK, existing common law, equality, privacy, and data protection frameworks offer limited restrictions on the use of AI in the workplace [14]. This piecemeal approach, relying on older frameworks not designed for AI, illustrates a broader challenge: static regulatory frameworks struggle to keep pace with technologies evolving at an unprecedented pace [5]. In contrast, the EU has created adaptive oversight mechanisms as technologies evolve and are able to ban or restrict AI practices that threaten personal rights, especially in high-risk contexts [1]. The EU's model is an attempt to address the need for adaptive governance frameworks that can iterate with the pace of change in technology and can automatically adjust risk assessments, compliance obligations, and accountability requirements as AI capabilities evolve. Indeed, without adaptivity, conventional legal frameworks will always lag behind the pace of innovation and make workers vulnerable to new applications of AI that fall beyond the existing protections while also preventing the deployment of positive use cases through the uncertainty of regulation.

In TIPSS - Trust, Identity, Privacy, Safety and Security - terms preventing firms from deploying against workers' interests requires governance that centres these values throughout the design and rollout of workplace systems.

Christiaens [15] proposes that the state should hold ownership of core AI infrastructure (to guarantee accountability and security) without day-to-day control, while independent social institutions - labour unions, employers, workers and technology firms - jointly negotiate how AI tools are developed and implemented. This social partnership model operationalises TIPSS by mandating trustworthy processes, protecting worker identity and privacy, cultivating safety-by-design, and securing systems and data, while enabling auditable oversight of AI in the workplace.
Addressing TIPPS requires moving beyond compliance toward governance that is dynamic, context-sensitive, and co-developed. Lessons from autonomous vehicle regulation - such as living codes of practice, scenario-based testing, and participatory frameworks [e.g., 8] - demonstrate how governance might grapple with the evolving nature of AI systems. These technologies are rarely static: they are constantly updated, retrained, or redesigned, meaning that rigid, one-off certification cannot guarantee ongoing safety or accountability.

Another problematic issue is that the very notion of 'proof of safety' is contested [7;16]: headline crash statistics can obscure who absorbs the harms, black box models resist formal verification, and context matters (driving in central London is not the same as driving on rural roads in Cumbria). Adaptive governance must therefore be iterative and reflexive, recognising that safety, trust, and fairness are not endpoints

but moving targets. By integrating job-impact strategies, social dialogue, and iterative oversight, AI governance can build trust, protect identity, ensure privacy, sustain safety, and reinforce security - precisely because it remains flexible, accountable, and responsive to both technological change and democratic values, rather than fixed in outdated rules or narrow technical ideals. Although the EU has developed an adaptive governance framework, the global nature of tech industry supply chains and operations presents complex challenges like data sovereignty. These can only be addressed by consistent international efforts on how organisations and regulators implement adaptive governance in practice. While frameworks and policy pilots exist, we know little about how iteration, reflexivity, and participatory oversight might be embedded in day-to-day settings. Filling this gap will require comparative research and cross-sector collaboration, alongside shared repositories of lessons from experiments such as sandboxes, phased licensing, and scenario-based testing.

Future developments

As AI systems become more autonomous and multimodal, governance will face greater pressure to adapt. Agentic AI systems capable of initiating actions, making decisions, and pursuing goals - raise profound questions about accountability and control. Who bears responsibility when an autonomous agent causes harm? How can such agents be aligned with human values [17, 18] across diverse cultural and legal contexts?

Trust will depend not only on transparency but also on authentication and verifiability. Regulators and users alike will demand real-time auditing mechanisms for genetic behaviour. Multimodal AI systems that integrate text, images, audio, and sensor data will make privacy protections harder to enforce, since sensitive information can be inferred from seemingly benign inputs and the capacity of AI systems to process vast quantities of data dwarfs anything that was conceived when current privacy regulations were developed.

In this context, identity and cybersecurity will be critical pillars of governance. With AI agents capable of impersonation, manipulation, and deepfake production, robust safeguards for digital identity will be essential. Cybersecurity frameworks will also need to evolve beyond technical hardening, addressing sociotechnical risk such as adversarial manipulation or the exploitation of opaque decision-making.

To meet these challenges, governance must move decisively away from static compliance checklists toward dynamic, context-aware protocols. In the UK, the Institute of Science and Technology (IST) Professional Accreditation Framework for AI Practitioners [19] - developed with a steering group drawing on government, industry, and academia - offers a living benchmark grounded in recognised competency standards (ethics, professional practice, critical evaluation, etc.). Given the historical absence of best practice, there is currently no set audit regime that is genuinely fit for purpose; accordingly, governance frameworks should focus on demonstrable competence, transparent process, and continuous improvement rather than one-off inspections. In parallel with lessons from autonomous vehicle governance, AI oversight should be continuously updated, tested in practice, and responsive to emerging threats and societal expectations, fusing technical rigour, legal foresight, and participatory design, so that users remain empowered, protected, and informed.

References

[1] European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union L, 12 July 2024. ELI: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Accessed 11 January 2026.

[2] Department for Science, Innovation and Technology (UK): AI regulation: a pro-innovation approach (White Paper). Gov.uk (2023). https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper, last updated 3 August 2023). Accessed 11 January 2026.

[3] Cyberspace Administration of China et al.: Provisions on the Management of Algorithmic Recommendations in Internet Information Services (promulgated 31 December 2021; effective 1 March 2022). https://www.chinalawtranslate.com/en/algorithms/. Accessed 11 January 2026.

[4] White House Office of Science and Technology Policy (OSTP): Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People (October 2022). (Archived official copy) https://data.aclum.org/storage/2025/01/OSTP_www_whitehouse_gov_ostp_ai-bill-of-rights.pdf. Accessed 11 January 2026.

[5] Mohamed, S., Png, MT. & Isaac, W. (2020) Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philos. Technol. 33, 659–684.

[6] National Institute of Standards and Technology (NIST) AI Risk Management Framework (2023). Available at: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[7] Department for Transport. (2019). Code of Practice: Automated Vehicle Trialling. Available at: https://www.gov.uk/government/publications/trialling-automated-vehicle-technologies-in-public/code-of-practice-automated-vehicle-trialling

[8] The Automated Vehicles Act (2024). Automated Vehicles Act 2024 (c. 10). Available at: https://www.legislation.gov.uk/ukpga/2024/10/contents

[9] The California Department of Motor Vehicles (CA DMV) 2025). Available at: https://www.dmv.ca.gov/portal/news-and-media/california-dmv-releases-updated-autonomous-vehicle-regulations/?utm_source=chatgpt.com

[10] United Nations Economic and Social Council (UNECE) (2024). New Assessment/Test Method for Automated Driving (NATM) Guidelines for Validating Automated Driving System (ADS). Available at: https://unece.org/sites/default/files/2024-02/ECE_TRANS_WP.29_2022_58e.pdf?utm_source=chatgpt.com

[11] Walkowiak, E., Potts, J. (2024). Generative AI, work and risks in cultural and creative industries. Available at SSRN 4830265. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4830265

[12] Martin, K. (2024). Why trust matters to productivity: A call to action for UK employers. The Productivity Institute. Available at: https://www.productivity.ac.uk/news/why-trust-matters-to-productivity-a-call-to-action-for-uk-employers/

[13] Center for Global Policy Solutions (CGPS) (2017). Stick Shift: Autonomous Vehicles, Driving Jobs, and the Future of Work. Washington, DC: Centre for Global Policy Solutions. Available at: https://www.law.gwu.edu/sites/g/files/zaxdzs5421/files/downloads/Stick-Shift-Autonomous-Vehicles-Driving-Jobs-and-

the-Future-of-Work.pdf?utm_source=chatgpt.com

[14] Brione, P. & Rough, E. (2023). Artificial intelligence and employment law. House of Commons Library. Available at: https://commonslibrary.parliament.uk/research-briefings/cbp-9817/

[15] Christiaens, T. (2025). Nationalize Ai!. AI & SOCIETY, 40(2), 1147-1149. DOI: https://doi.org/10.1007/s00146-024-01897-0

[16] Kalra, N and Paddock, S.M (2016). Driving to Safety. Available at: https://www.rand.org/pubs/research_reports/RR1478.html.

[17] McKinlay, J. Vos, M. Hoffmann, J. and Theodorou, A. (2025). Understanding the Process of Human-AI Value Alignment. DOI: https://doi.org/10.48550/arXiv.2509.13854

[18] Apeh, E. Saeghe, P. Kouadri Mostefaoui, S. Lu, Y. and Shang, L (2026). 'AI Alignment and TIPSS' Sprite Plus Explainer. Available at: https://77265d30-ac80-43a7-a339-19564066f602.usrfiles.com/ugd/77265d_17f38282ccd340ee9128090ba1cfc0c0.pdf

[19] The Institute of Science and Technology professional Accreditation Framework for AI Practitioners (2023). Available at: https://istonline.org.uk/cms/wp-content/uploads/2024/06/The-IST-Professional-Accreditation-Framework.pdf