SPRITE+ Explainer #004

Series Editor: Mark Elliot

AI Alignment in the Context of TIPSS

By Edward Apeh, Pejman Saeghe, Soraya Kouadri Mostefaoui, Yang Lu and Lijun Shang

This explainer introduces AI Alignment and its relationship to TIPSS - a set of five interrelated concepts: Trust, Identity, Privacy, Security, and Safety. These concepts help illuminate human-centric considerations in the design, deployment and governance of AI systems.

The explainer begins by defining AI Alignment and outlining key developments in the field. It then explores how TIPSS concepts intersect with alignment efforts, emphasizing ethical, participatory, emotional and cultural dimensions. Finally, it reflects on the broader implications of aligning AI systems with human values and societal expectations.

What is AI Alignment?

AI Alignment refers to the process of designing and guiding artificial intelligence systems so that their goals, behaviours, and impacts are consistent with human values and societal expectations [42]. It encompasses technical strategies (e.g., reward modelling, interpretability) as well as ethical and governance considerations.

Alignment is typically divided into:

- **Outer alignment:** Making sure the AI's objectives reflect human goals.
- **Inner alignment:** Ensuring the AI reliably pursues those goals even in unfamiliar situations.

As AI systems become more autonomous and complex, the risk of misalignment grows [23]. Such misalignment can lead to harms, such as reward hacking or emergent behaviours that conflict with human interests [35]. Alignment research develops methods to guide AI systems toward safe, predictable, and beneficial behaviour.

Importantly, alignment does not presume a fixed set of universal human values; rather, it engages with cultural diversity, ethical pluralism, and the evolving nature of societal expectations [2,14]. Even the notion that ethics is culturally mediated, which is often proposed to reconcile competing value systems, is itself contested [8,40]. Alignment, therefore, operates within a landscape of philosophical complexity and should be approached as an ongoing negotiation rather than a definitive resolution [12,26].

Why TIPSS Matters for AI Alignment

AI alignment draws on a range of human-centric concerns including trust, identity, privacy, security, and safety, that are increasingly formalised in governance and assurance practices. For instance, Stanford HAI (2023) highlights how privacy and safety are central to AI governance, calling for new mechanisms to address systemic risks in data use and algorithmic decision-making [38]. Similarly, Luger and Sellen found that user trust in AI systems is closely tied to transparency, data control, and

regulatory compliance [28]. Furthermore, Gabriel (2020) and Goffi (2021) argue that AI alignment should go beyond preference satisfaction to incorporate normative human values such as fairness, mutual benefit, and social responsibility [21,22]. While the relationship between these concerns and technical alignment methods are still evolving, this explainer emphasises that AI alignment should be informed by such human priorities to ensure ethical and socially responsible AI development.

TIPSS and AI Alignment Techniques

Recent advances in AI alignment are reshaping how TIPSS concepts are implemented in intelligent systems. As alignment techniques become more sophisticated, they are influencing not only technical safety but also the ethical, governance, and operational dimensions of AI. This section outlines key developments in alignment research alongside the evolving considerations for TIPSS domains.

Trust: Scalable Oversight and Interpretability

Alignment methods such as Reinforcement Learning from Human Feedback [23]. and Debate Models [9] are enhancing the reliability and transparency of AI systems. These techniques improve interpretability and predictability, which are foundational for user trust. To reflect these advances, TIPSS-informed practices (i.e. practical tools, frameworks, protocols, and strategies), should include:

- **Trust calibration tools**: Methods used to measure and adjust how confident users feel about an AI system. For example, a medical diagnosis assistant might vary how it displays confidence levels depending on the user's expertise [30].
- **Explainability standards**: Guidelines that help make AI decisions clear and understandable to users. For instance, a loan approval system might show which factors influenced the decision and explain why the application was accepted or rejected [25].
- **Dynamic audit frameworks**: Systems designed to continuously monitor AI behaviour and ensure accountability over time. For example, an AI hiring tool could keep a log of its decisions and regularly check for patterns of bias or unfair treatment [5].

Identity: Robustness and Behavioural Consistency

Techniques such as adversarial robustness [25], red teaming [10], and distributional shift resilience are improving the consistency of AI behaviour across contexts. These methods help ensure that AI agents cannot be spoofed or manipulated. To strengthen identity assurance, TIPSS informed alignment would incorporate:

- **Adversarial testing**: This involves deliberately challenging AI systems to expose vulnerabilities. For example, testing a facial recognition system with altered images to ensure it doesn't misidentify people. [16],
- **Behavioural fingerprinting**: This involves identifying AI agents based on unique behavioural patterns. For example, detecting bots in online platforms by analysing their interaction patterns [19],
- **Anomaly detection**: This involves identifying unusual or unexpected behaviour that may indicate malfunction or manipulation. For example, flagging a self-driving car's sudden deviation from expected route behaviour [12].

Privacy: Normative Alignment and Ethical Boundaries

Normative role-based alignment means designing AI to respect contextual norms such as cultural expectations or professional codes, rather than relying solely on user preferences. The shift from preference-based alignment to normative role-based alignment [41]. is prompting a re-evaluation of privacy standards. AI systems are increasingly designed to respect contextual norms and ethical boundaries [24]. This strengthens privacy through principled data handling and by embedding ethical boundaries into system behaviour. To support this shift, TIPSS-informed methods should include:

- **Privacy-aware reward modelling**: This involves designing AI incentives that avoid exploiting personal data. For example, a recommendation engine that avoids using sensitive browsing history to optimise engagement [39].
- **Consent-aware system design**: This involves ensuring users understand and agree to how their data is used. For example, a fitness app that asks users to opt in before sharing health data with third parties [34].
- **Federated learning architectures**: This involves training AI models across distributed data sources without centralising sensitive information. For example, a keyboard app that learns from user typing patterns locally without uploading data to the cloud [43].

Security: Governance-Informed Assurance Protocols

Alignment research is informing new governance models that introduce standards for security and compliance. These models support multi-stakeholder auditing, safety-washing detection, and alignment-based assurance protocols [4].

To maintain system integrity, TIPSS methods should integrate:

- **Alignment-aware threat modelling**: This involves anticipating risks based on how AI systems interpret and pursue goals. For example, evaluating how a warehouse robot might misinterpret “optimise speed” and compromise safety [16].
- **Secure deployment pipelines**: These are processes that ensure systems are safely released into real-world environments. For example, a cloud-based AI model that undergoes vulnerability scans before deployment [31].
- **Compliance frameworks**: These are structures that ensure AI systems meet legal and ethical standards. For example, an AI chatbot that is regularly audited to ensure GDPR compliance [17].

Safety: Proactive Risk Mitigation and Ethical Engineering

Safety is being redefined through alignment techniques that anticipate and mitigate emergent risks. These include human-in-the-loop oversight, interpretability tools, and constitutional AI frameworks [11,33,36]. To prevent harmful outcomes, TIPSS-informed practices should support:

- **Simulation-based safety testing**: This involves running virtual scenarios to identify potential failures before deployment. For example, simulating traffic situations to test how an autonomous vehicle reacts to pedestrian behaviour [3].
- **Real-time risk scoring**: This involves dynamically assessing threats as they arise during system operation. For example, a cybersecurity AI that adjusts

its alert level based on live threat patterns [36].

- **Ethical constraint enforcement**: This involves embedding rules that prevent AI systems from taking harmful actions. For example, a content moderation AI that refuses to promote misinformation even if it increases engagement [12].

See the Appendix for a summary of the relationship between future AI alignment developments and the potential implications for TIPSS methods.

Human-Centric Perspectives in AI Alignment through TIPSS

While technical methods are essential for AI Alignment, they should be grounded in a deep understanding of human values, experiences, and societal contexts. TIPSS should reflect not only system-level integrity but also the human impact of intelligent systems.

AI systems ought to some extent reflect diverse human values. However, if AI were to reflect all humanity this would merely amplify the chaotic polarisation that has come to dominate our cities in the twenty first century. It may be that we need AI to be different from us – to align with what we need rather than who we are. In tune with this Emerging research highlights anthropomorphism and dehumanisation as critical factors shaping perceptions of AI [1,32]. As Bialy et al say: "Research shows that AI systems designed to be overtly human-like meet with scepticism, with a preference for functionality, human control, transparency, and fairness over anthropomorphism and unrestrained autonomy". These issues underscore the need for alignment strategies that preserve human dignity and avoid misleading design cues.

Furthermore, participatory design and pluralistic alignment frameworks help ensure that systems serve varied cultural and ethical perspectives [37]. Maintaining human control over AI systems is widely regarded as essential for ethical deployment. Governance frameworks that support human-in-the-loop oversight and contestability mechanisms are important to preserving autonomy and accountability [20].

Trust in AI is shaped not only by technical transparency but also by perceived humanness and emotional resonance. Studies show that interactivity and human-like design features significantly influence user trust and adoption [15]. AI systems should be designed to uphold human dignity and moral values. Virtue ethics and care ethics offer guiding principles for embedding ethical reasoning and relational awareness into intelligent systems [18].

AI has the potential to both mitigate and exacerbate social inequalities. Ethical alignment should include fairness auditing, inclusive data practices, and mechanisms to prevent algorithmic bias and marginalisation" [6].

Closing Reflections

This explainer has provided an overview of AI Alignment and its intersection with the TIPSS concerns - Trust, Identity, Privacy, Security, and Safety. It outlined key AI alignment techniques such as RLHF, debate models, adversarial robustness, and normative alignment, and mapped them to TIPSS domains. A dedicated section emphasized the human-centric dimensions of alignment, including participatory design, emotional safety, and ethical grounding.

AI Alignment is a multifaceted challenge that demands both technical innovation and human-centred thinking. As intelligent systems become more autonomous, the TIPSS should consider addressing emerging risks and societal expectations. By integrating alignment techniques with participatory ethics, oversight mechanisms,

and fairness safeguards, we can build AI systems that earn trust, respect human dignity, and promote equitable outcomes. Ultimately, aligning AI with human values is not just a technical task—it is a societal imperative.

References

[1] Abrams, L. (2024) 'Algorithmic equity and social justice in AI systems', Journal of Ethics in Technology, 12(1), pp. 45–62.

[2] Aditya, S. (2025). 'Toward global ethical frameworks for AI: Aligning artificial intelligence with human values and progress.' World Journal of Advanced Engineering Technology and Sciences, 15(2), 1823–1831

[3] Bai, Y., Zhang, H. and Liu, Q. (2025) 'Simulation-based safety testing for autonomous systems', Proceedings of the AAAI Conference on Artificial Intelligence, 39(2), pp. 1123–1131.

[4] Batool, S., Mehta, R. and Carvão, J. (2025) 'Governance-informed assurance protocols for AI alignment', AI & Society, 40(3), pp. 301–319.

[5] Benk, M., O'Connor, T. and Singh, A. (2025) 'Dynamic audit frameworks for aligned AI', Journal of Trustworthy AI, 7(1), pp. 88–104.

[6] Bialy, F., Elliot, M., & Meckin, R. (2025). Perceptions of AI Across Sectors: A Comparative Review of Public Attitudes. arXiv preprint arXiv:2509.18233.

[7] Bravansky, M., Sorensen, E., & Malik, R. (2025). Rethinking AI cultural alignment: A bidirectional model for human-AI ethical co-design. Journal of Inclusive AI, 6(1), 55–72.

[8] Biosca, A., Dunel, P., Costa, F. J., Bota, G., Quirante, A., & Roca, M. (2025). 'AI Alignment: Ethical Challenges, Real-World Failures and Multidisciplinary Solutions.' Universitat Politècnica de Catalunya.

[9] Buhl, M., Tanaka, Y. and Singh, R. (2025) 'Debate models for scalable oversight', NeurIPS Workshop on Human-AI Collaboration.

[10] Burt, A. (2024) 'Red teaming AI: A practical guide', AI Security Review, 6(4), pp. 210–225.

[11] Chen, L., Sarkar, A. and Liu, M. (2024) 'Ethical constraint enforcement in AI training', ACM Transactions on AI Ethics, 3(2), pp. 77–95.

[12] Chen, L., Zhang, Y. and Sarkar, A. (2025) 'Anomaly detection in aligned systems', Journal of Machine Learning Research, 26(1), pp. 134–150.

[13] Coetzee, D. (2025). 'The Artificial Intelligence Readiness Prism: A multi-dimensional framework for assessing AI integration, ethics, and cultural alignment.' Project Future; Development Bank of Southern Africa; Free Market Foundation.

[14] Dignum, V. (2019). Responsible Artificial Intelligence: How to develop and use AI in a responsible way. Springer.

[15] Ding, Y. and Najaf, M. (2024) 'Emotional safety and trust in human-AI interaction', Human-Centered AI Journal, 9(2), pp. 101–118.

[16] Du, J. (2025) 'Adversarial testing and threat modelling for AI security', IEEE Transactions on Secure AI, 14(3), pp. 233–248.

[17] Essert Inc. (2025) Compliance frameworks for AI deployment. White Paper Series.

[18] Fasoro, D. (2024) 'Virtue ethics and human dignity in AI design', Ethics in Emerging Technologies, 11(1), pp. 23–39.

[19] Feng, S. and Sehatbakhsh, S. (2025) 'Behavioural fingerprinting for identity assurance in AI', Proceedings of the

International Conference on AI Security, pp. 89–102.

[20] Frenette, A. (2023) 'Human oversight and contestability in automated decision-making', AI & Law Review, 18(4), pp. 201–219.

[21] Gabriel, I. (2020). *Artificial intelligence, values, and alignment*. Minds and Machines, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[22] Goffi, E. (2021). Escaping the Western cosm-ethical hegemony: Toward pluralism in AI ethics. AI & Ethics, 2(3), 231–245. https://doi.org/10.1007/s43681-021-00079-6

[23] Ji, X., Lambert, T. and Reis, M. (2025) 'Reinforcement learning from human feedback: Advances and challenges', Journal of AI Alignment, 5(1), pp. 1–20.

[24] King, R. and Meinhardt, J. (2024) 'Contextual norms in privacy-aware AI', Privacy & Ethics Quarterly, 10(3), pp. 145–160.

[25] Lahusen, M., Patel, R. and Singh, A. (2024) 'Explainability standards for trustworthy AI', Journal of Explainable Systems, 6(2), pp. 67–82.

[26] Lambert, T. (2025) 'RLHF and interpretability in large language models', AI Research Letters, 8(1), pp. 33–49.

[27] Li Hou, B., & Green, B. P. (2022). 'A multilevel framework for the AI alignment problem: Individual, organizational, national, and global.' Markkula Center for Applied Ethics, Santa Clara University.

[28] Luger, E., & Sellen, A. (2021). *"Like having a really bad PA": The gulf between user expectation and experience of conversational agents*. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3411764.3445755

[29] Majumdar, A., Chen, Y. and Liu, Z. (2025) 'Adversarial robustness in alignment-critical systems', IEEE Conference on Robust AI, pp. 55–70.

[30] Marusich, L., Tanaka, Y. and O'Neill, K. (2025) 'Trust calibration tools for aligned AI', Human Factors in AI, 4(1), pp. 12–29.

[31] Microsoft. (2025) Secure deployment pipelines for AI systems. Microsoft Research Technical Report.

[32] Oldfield, M. (2023). Anthropomorphism and its impact on the implementation and perception of AI. In J. Casas Roma, J. Conesa, & S. Caballé (Eds.), Technology, users and uses: Ethics and human interaction through technology and AI (pp. 99–134). IET. ISBN: 9781804414033

[33] Pieces, A. (2025) 'Constitutional AI frameworks for safety alignment', AI Governance Journal, 3(1), pp. 41–58.

[34] Pistilli, G. and Jernite, Y. (2025) 'Consent-aware system design in AI', Proceedings of the Fairness in AI Symposium, pp. 77–90.

[35] Reis, M. and La Cava, W. (2025) 'Specification gaming and emergent misalignment', Journal of AI Safety, 9(1), pp. 101–120.

[36] Sarkar, A. (2025) 'Real-time risk scoring for dynamic threat assessment', AI Risk Management Review, 7(2), pp. 55–70.

[37] Sorensen, E., Malik, R. and Chen, T. (2024) 'Participatory design and value pluralism in AI alignment', Journal of Inclusive AI, 5(3), pp. 88–105.

[38] Stanford HAI. (2023). *Building a new data governance infrastructure for AI*. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/news/building-new-data-governance-infrastructure-ai

[39] Sun, Y., Zhang, L. and Tan, R. (2025) 'Privacy-aware reward modelling for aligned systems', Journal of Privacy-Preserving AI, 6(1), pp. 33–47.

[40] Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural

alignment of large language models. PNAS Nexus, 3(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

[41] Tan, R., Mehta, S. and King, R. (2025) 'Normative role-based alignment in intelligent systems', Ethics in AI Journal, 8(2), pp. 112–130.

[42] World Economic Forum. (2024). AI Value Alignment: Guiding Artificial Intelligence Towards Shared Human Goals. Global Future Council on the Future of AI. Retrieved from https://www.weforum.org/publications/ai-value-alignment-guiding-artificial-intelligence-towards-shared-human-goals/

[43] Zhan, Q., Liu, M. and Patel, R. (2025) 'Federated learning architectures for privacy-preserving AI', IEEE Transactions on Distributed AI, 19(1), pp. 55–70.