Studia Iuridica Lublinensia vol. 34, 2, 2025
DOI: 10.17951/sil.2025.34.2.441-479

Karol Kasprowicz
Maria Curie-Skłodowska University (Lublin), Poland
ORCID: 0000-0001-6328-052X
karol.kasprowicz@mail.umcs.pl

Alignment Problem as Cultural and Legal Challenge: Artificial Intelligence, Interpretability, and Searching for Sense

Problem dostosowania jako wyzwanie kulturowe i prawne. Sztuczna inteligencja, interpretowalność i poszukiwanie sensu

ABSTRACT

The article examines the AI alignment problem as a fundamental challenge of cross-cultural communication between human interpretive frameworks and algorithmic optimization. The author argues that effective AI alignment requires integrating cultural sense-making practices and legal frameworks that vary across societies. The analysis reveals how current regulatory attempts, including the EU AI Act and national AI strategies, struggle with three interconnected challenges: ensuring the interpretability of algorithmic decisions, managing the indeterminism inherent in AI systems, and addressing knowledge extraction controversies. Through examination of emerging AI agents, Big Tech's regulatory capture, and the rise of AI nationalism, the study demonstrates that alignment failures stem not from technical limitations alone, but from inadequate engagement with diverse cultural logics of interpretation. The author proposes frameworks that adapt AI systems to varied contexts while maintaining core functionality and concludes that solving alignment requires computational cultural modelling capable of navigating value pluralism. The analysis warns that without integrating technical safety mechanisms with cultural frameworks of societies, AI systems risk becoming tools of extraction and control rather than beneficial partners for societies.

Keywords: alignment; artificial intelligence; interpretability; regulations; sense-making; culture

CORRESPONDENCE ADDRESS: Karol Kasprowicz, PhD, Assistant Professor, Maria Curie-Skłodowska University (Lublin), Faculty of Law and Administration, Institute of Legal Sciences, 5 Maria Curie-Skłodowska Square, 20-031 Lublin, Poland.

The Zeroth Law of Robotics
A robot may not injure humanity, or,
through inaction, allow humanity to come to harm.

I. Asimov, *Robots and Empire* (1985)

INTRODUCTION

In 1942, I. Asimov proposed the Three Laws of Robotics as a fictional foundation for safe coexistence between humans and intelligent machines.[1] Eight decades later, in an era of advanced artificial intelligence (AI) systems, we face far more complex challenge: ensuring AI systems operate according to human values and goals. Science fiction has long explored potential futures where intelligent machines interact with humanity, but Asimov's Laws represent perhaps the most enduring attempt to codify principles governing such interactions.[2] Nevertheless, fictional guidelines, while elegant in their simplicity,[3] fail to address the nuanced challenges of modern AI systems that operate through statistical patterns rather than deterministic rules.[4]

Thus, the AI alignment problem is essentially a problem of cross-cultural communication between the world of human interpretation and the world of algorithmic optimization. I presume that this framing reveals a key dimension of usefulness of AI to human life – interpretability as more than a technical problem and sense-making of real uses of AI. It represents a challenge of translation between two distinct forms of intelligence: human understanding built on cultural contexts, emotional resonance, and embodied experience vs machine learning systems operating through statistical pattern recognition across massive datasets. Interpretability challenges emerge from technical opacity as well as from fundamental differences in how humans and AI systems process information. While humans interpret through contextual understanding, cultural frameworks, and embodied experience, AI systems operate through statistical correlations that may lack causal understanding. This gap creates profound challenges for ensuring AI systems genuinely align with human intentions rather than merely optimizing for specified objectives that incompletely capture human values.[5]

---

[1] I. Asimov, *Runaround*, "Astounding Science Fiction" 1942, no. 3, pp. 94–103.

[2] Cf. K. Mamak, *Whether to Save a Robot or a Human: On the Ethical and Legal Limits of Protections for Robots*, "Frontiers in Robotics and AI" 2021, vol. 8.

[3] For example, see J. Zajdel, *Limes Inferior*, Warszawa 1982; N. Bostrom, *Deep Utopia: Life and Meaning in a Solved World*, 2024; S. Lem, *Golem XIV*, Kraków 1981.

[4] Cf. M. de Sautoy, *The Creativity Code: Art and Innovation in the Age of AI*, Cambridge 2020.

[5] A. Elliott, *Making Sense of AI: Our Algorithmic World*, Cambridge 2022, pp. 41–44.

Y. Bengio emphasizes that without a deep understanding of cultural mechanisms of sense-making, even the most advanced AI systems may remain fundamentally misaligned with human values.[6] Furthermore, the 2025 *International AI Safety Report* identifies interpretability, knowledge extraction, and managing indeterminism as one of the key challenges for AI safety in the coming decade.[7] The survey and report both emphasize the insufficiency of purely technical approaches, highlighting instead the need for socio-legal frameworks that can accommodate rapid technological evolution.[8] This urgency is underscored by D. Kokotajło's *AI 2027* scenario, which projects transformative AI capabilities emerging within just two years – a timeline that suggests current alignment research may be racing against technological development.[9]

Bengio's advocacy for slowing AI development reflects similar concerns about the temporal mismatch between capability advancement and safety research, echoing Tegmark's Future of Life Institute position that regulatory breathing room is essential for developing adequate governance structures.[10] These calls are directly reflected in the growing emphasis on transparency as a key element of AI regulatory policy.[11] This is not, however, merely a call for a slowdown. In parallel, the first

[6] His conclusions are much more alarmistic: "I feel strongly that it is critical to invest immediately and massively in research endeavours to design systems and safety protocols that will minimize the probability of yielding rogue AIs, as well as to develop countermeasures against the possibility of undesirable scenarios. There is a great need and opportunity for innovation in governance research to design adaptable and agile regulations and treaties that will safeguard citizens and society as the technology evolves and new unexpected threats may arise. I believe we have the moral responsibility to mobilize our greatest minds and major resources in a bold, coordinated effort to fully reap the economic and social benefits of AI, while protecting society, humanity, and our shared future against its potential perils. And we need to do so urgently, with the United States playing the same leadership role in protecting humanity as it is in advancing AI capabilities" (Y. Bengio, *Government Interventions to Avert Future Catastrophic AI Risks*, "Harvard Data Science Review" 2024, no. 5, Special Issue).

[7] Y. Bengio (ed.), *International AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, January 2025, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf (access: 17.10.2025).

[8] We can refer here also to technological developments of innovation, e.g. Gartner hype cycle – framework introduced by J. Fenn which provides a visual model for tracking how technologies evolve through stages of maturity and adoption within society. It maps the lifecycle of emerging technologies from initial breakthrough to mainstream application. However, the model's reliability remains questionable. Research examining its predictive power has revealed significant limitations – empirical evidence suggests the framework's accuracy is sporadic and unreliable.

[9] D. Kokotajło, S. Alexander, T. Larsen, E. Lifland, R. Dean, *AI 2027*, 3.4.2025, https://ai-2027.com (access: 19.10.2025).

[10] Future of Life Institute, *Pause Giant AI Experiments: An Open Letter*, 22.3.2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments (access: 20.7.2025).

[11] See also RenAIssance Foundation, *The Rome Call for AI Ethics*, 28.2.2020, https://www.romecall.org/the-call (access: 20.6.2025). Cf. S. Hastings-Woodhouse, D. Kokotajło, *We Should Not Allow Powerful AI to Be Trained in Secret: The Case for Increased Public Transparency*, 27.5.2025,

promising research avenues are emerging, aimed at increasing control over the internal processes of models. One such avenue is *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*, which offers a new, albeit still fragile, opportunity for real progress in safety (e.g. open source initiatives).[12] The significance of this direction is underscored by the fact[13] that it is becoming the focus of flagship government initiatives (e.g. such as the UK's AI Safety Institute – AISI).[14] Thus, the debate on AI safety is transitioning from a phase of manifestos and appeals to a stage of institutional support for concrete solutions to the alignment problem – encapsulated in the first attempts to regulate it.

Culture and, in particular, law can provide frameworks and tools to address challenges of aligning AI to humanity. Legal frameworks provide essential structures for governing technology, but law itself represents a cultural technology evolving alongside the systems it regulates. Traditional legal approaches assuming deterministic causation face significant challenges when applied to probabilistic AI systems operating through statistical inference rather than explicit rules. Human societies have historically developed sophisticated mechanisms for coordinating diverse agents with potentially conflicting interests through shared norms, institutions, and collaborative frameworks.[15] These social practices

https://www.aipolicybulletin.org/articles/we-should-not-allow-powerful-ai-to-be-trained-in-secret-the-case-for-increased-public-transparency (access: 20.6.2025).

[12] T. Korbak et al., *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*, 15.7.2025, https://arxiv.org/abs/2507.11473 (access: 20.6.2025).

[13] For this kind of action initiatives one of the key opportunities to succeed are open source (open software/models) movements. They are an extremely important catalyst for advances in AI security. Closed models, accessible only through APIs (such as GPT-4 from OpenAI), allow only their "behaviour" to be studied. Open source models (such as Llama from Meta, Mistral) give researchers full access to their "brain". The open source community creates and provides tools for analysing and interpreting AI models (e.g. libraries such as TransformerLens or platforms like Hugging Face). This speeds up research for everyone because no one has to "reinvent the wheel". Also open source initiatives like automated toolkits, such as PyRIT, systematize the process of "red teaming" in search of gaps. Initiatives such as the AI Alliance further standardize these efforts, providing a framework for the secure development of open artificial intelligence. A key strength of this ecosystem is the global community, which uses open access to conduct independent audits and public testing. From organized "jailbreaking" competitions to academic publications exposing new vulnerabilities, the "many eyes see more" principle is at work here. Not only does this enable verification of security claims made by developers, but also ensures the reproducibility of research, which is fundamental to scientific progress. In this way, grassroots pressure and open source collaboration create a dynamic cycle of discovering, documenting and fixing vulnerabilities, realistically accelerating the development of safer and more trustworthy AI systems. We will refer to open source later, defining major problems with AI regulations.

[14] AI Security Institute, *The Alignment Project*, https://alignmentproject.aisi.gov.uk (access: 20.7.2025).

[15] Cf. M. Bennett, *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs That Made Our Brains*, New York–Boston 2023, pp. 344–358. On social dimension, see M. Pasquinelli,

of sense-making, value negotiation, and collective decision-making represent centuries of evolutionary adaptation to the challenge of aligning individual and group interests. Within this broader social context, law emerges as a particularly refined tool for fostering cooperation among agents and facilitating joint actions. W. Załuski's game-theoretic analysis of law as a cooperation-fostering mechanism offers potential pathways forward for AI alignment, suggesting that legal frameworks might provide coordination tools for aligning multiple AI and human agents around shared values and goals.[16]

With all these considerations in mind, this article addresses the AI alignment problem as a cultural and legal challenge, focusing on three key aspects: aligning AI as a social practice of taming technological uncertain outcomes, interpretability of algorithmic decisions, and cultural practices of sense-making related to AI systems' actions. "Cultural practices of making sense" are a set of shared, socially inherited schemas by which people interpret AI actions, judging their legitimacy, fairness and credibility. In the context of this article, these practices explain why the same regulatory framework for AI may be accepted as a necessary tool for protecting fundamental rights in one jurisdiction, and rejected as a barrier to economic progress in another.

Hypotheses:

1. Cultural practices of sense-making and interpretation hold fundamental importance for effective AI alignment. Technical safety mechanisms are inadequate and risk failure if not integrated into cultural frameworks of societies.
2. Effective AI alignment requires a new, transdisciplinary approach which integrates technical, cultural, social and legal dimensions of diverse phenomena (e.g. interpretability, indeterminism, and knowledge extraction).

The aim of this analysis is to propose an integrated model of AI alignment. This article employs a qualitative, interdisciplinary methodology, and the core method is a critical analysis of a diverse range of texts, spanning technical AI safety research, socio-legal theory, and contemporary policy documents. This approach facilitates a theoretical synthesis that addresses the purely technical view of alignment by foregrounding often overlooked cultural frameworks of sense-making.

ALIGNMENT PROBLEM

From a theoretical standpoint, the alignment problem manifests as a key agency problem – a situation where an agent (AI system) must act on behalf of and in accordance with the intentions of a principal (human operator). The complexity of

The Eye of the Master: A Social History of Artificial Intelligence, London–New York 2023.

[16] W. Załuski, *Game Theory in Jurisprudence*, Kraków 2014, p. 81.

this problem stems from the impossibility of precisely encoding the full spectrum of human values, intentions, and preferences into formal systems. Moreover, even the very concept of "human values" is not monolithic. It varies across cultures, social groups, and even individuals.[17] This inherent gap between human intent and code makes interpretability and explainability central to the alignment problem, because culture – in general – is not universal and homogenic. Culture is rather diverse and relativistic at its core.[18]

If we cannot perfectly specify our objectives a priori, we must be able to audit the agent's reasoning post hoc. Interpretability – the ability to scrutinize a model's internal mechanics and understand how it reaches its conclusions – becomes a crucial diagnostic tool. It allows us to verify that the AI has not developed a flawed or dangerously simplified proxy for our intended values. Furthermore, as AI systems make decisions that impact a pluralistic society, explainability – the capacity of an agent to justify its actions in human-understandable terms – becomes a prerequisite for legitimacy and trust. An unexplainable decision that navigates a complex ethical trade-off cannot be debated, contested, or democratically governed. Therefore, solving alignment is not solely about creating an obedient agent; it is about creating a transparent one whose internal logic and external justifications are open to human scrutiny, ensuring it remains a truly accountable partner rather than an inscrutable black box.

We cannot align what we cannot interpret, and we cannot interpret what we view through a single cultural lens, suggesting that true interpretability must encompass the diverse ways human communities construct and communicate meaning. I reckon then, that one of possible answers to question of how to align with "thinking machines" lies in interpretability. D. Amodei highlights that understanding how AI systems process and represent is not just a technical challenge but a prerequisite for meaningful alignment.[19] He emphasizes this urgency by noting that "we are thus

[17] For example, see B. Whorf, *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, ed. J.B. Carroll, Cambridge 1956, pp. 246–254; G. Hofstede, G.J. Hofstede, M. Minkov, *Cultures and Organizations: Software of the Mind*, New York 2010; J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, New York 2012. See also very important for training autonomous systems MIT Media Lab Moral Machine, theoretically based on P. Foot ethical thought experiment – trolley problem: Moral Machine Platform, *MIT Media Lab*, https://www.moralmachine.net (access: 5.8.2025).

[18] Cf. S. Natale, F. Biggio, P. Arora, J. Downey, R. Fassone, R. Grohmann, A. Guzman, E. Keightley, D. Ji, V. Obia, A. Przegalinska, U. Raman, P. Ricaurte, E. Villanueva-Mansilla, *Global AI Cultures: How a Cultural Focus Can Empower Generative Artificial Intelligence*, 8.8.2025, https://cacm.acm.org/opinion/global-ai-cultures (access: 15.8.2025).

[19] D. Amodei, *The Urgency of Interpretability*, 2025, https://www.darioamodei.com/post/the-urgency-of-interpretability (access: 14.5.2025). We need also to explain the background of this researcher. Amodei is the CEO and co-founder of Anthropic, an AI safety company that created the Claude models. He previously served as Vice President of Research at OpenAI, where he led the development of GPT-2 and GPT-3, and was also co-inventor of the RLHF (Reinforcement Learning

in a race between interpretability and model intelligence. It is not an all-or-nothing matter: as we've seen, every advance in interpretability quantitatively increases our ability to look inside models and diagnose their problems".[20] The stakes of this race are particularly high because, as he warns, we could have AI systems equivalent to a "country of geniuses in a data center" as soon as 2026 or 2027.[21] Amodei considers it "basically unacceptable for humanity to be totally ignorant of how they work" when deploying such systems.[22]

The urgency of interpretability directly addresses the alignment problem by providing a potential solution to the opacity that characterizes modern AI systems.[23] As Amodei explains, "Modern generative AI systems are opaque in a way that fundamentally differs from traditional software". Unlike conventional programs where "a human specifically programmed them in", generative AI systems are "grown more than they are built – their internal mechanisms are 'emergent' rather than directly designed [emphasis – K.K.]".[24] This opacity creates a cascade of alignment challenges: we cannot predict harmful behaviours, cannot provide meaningful explanations for decisions, and cannot systematically prevent deception or power--seeking behaviours. And that's the key to address the challenges of alignment, because most of risks are in the end consequences of this opacity. Interpretability means to make our eyes fully open – make AI systems interpretable.

Hence, the practical solution Amodei proposes – developing an "MRI for AI"[25] – represents a concrete approach to bridging the alignment gap. Interpretability framework

---

from Human Feedback) method. He is a physicist by training (PhD from Princeton) and one of the leading researchers in AI safety and alignment. As CEO of Anthropic, Amodei occupies an inherently conflicted position – on one hand, he is an advocate for AI safety who warns about existential risks, while on the other, he runs a company competing in the commercial market. This structural conflict of interest may influence his public statements: emphasizing AI risks can justify Anthropic's approach to developing "safer" AI, while simultaneously promoting Claude's capabilities serves business objectives. His perspective on regulation, the pace of AI development, or the definition of "safe" development is inevitably shaped by his company's market position and strategy, making him both a valuable yet non-objective voice in the debate about AI's future.

[20] *Ibidem*.

[21] It corresponds with mentioned *AI 2027* document. Although this is a very optimistic assumption. Other executives are much more conservative in this kind of predictions (e.g. D. Hassabis, S. Altman).

[22] D. Amodei, *op. cit*.

[23] The problem of opacity is twofold, encompassing not only the technical "black box" of AI models but also the social opacity of human interaction with them.

[24] D. Amodei, *op. cit*. Cf. R. Mishra, G. Varshney, *Exploiting Jailbreaking Vulnerabilities in Generative AI to Bypass Ethical Safeguards for Facilitating Phishing Attacks*, 16.7.2025, https://arxiv.org/abs/2507.12185 (access: 2.8.2025).

[25] However, this "MRI for AI" metaphor itself warrants scrutiny. While interpretability tools promise insight into AI systems' inner workings, they simultaneously risk creating an illusion of insight. The very act of observation introduces a crucial epistemological problem: when we prompt models to explain their

would enable practitioners to conduct comprehensive "brain scans" of AI systems, identifying "tendencies to lie or deceive, power-seeking, flaws in jailbreaks, cognitive strengths and weaknesses".[26] Such capabilities would transform alignment from a theoretical problem into a manageable engineering challenge, allowing for iterative testing and refinement of AI systems before deployment. The urgency stems from the temporal mismatch between AI capability advancement and interpretability research – we risk deploying systems before developing adequate tools to understand and control them.

The alignment problem in AI is one of the most profound challenges at the crossroads of technology, philosophy, ethics, and governance. At its core, this problem concerns how to ensure that increasingly powerful AI systems act in accordance with human values, intentions, and welfare. The fundamental challenge lies not in creating intelligent systems, but in creating systems whose goals remain aligned with human flourishing even as their capabilities expand beyond human comprehension.[27] This challenge is particularly acute because powerful optimization processes directed toward misspecified objectives may produce catastrophic[28] outcomes despite achieving their formal goals.[29]

Nevertheless, the multidimensional nature of the alignment problem extends beyond technical specifications into profound questions of interpretation and meaning. B. Christian highlights how the challenge involves translating vague, context-dependent, and culturally variable human values into precise mathemat-

reasoning chains, we potentially alter their behaviour – a kind of observer effect in AI. Models may generate plausible-sounding explanations that bear little relationship to the actual computational processes driving their outputs. When we reward interpretability as a metric, we may inadvertently train models to appear interpretable rather than genuinely being so. They learn to produce the linguistic artifacts of transparency – step-by-step reasoning, coherent justifications, apparent logical structures – without these narratives necessarily reflecting their true decision-making processes. As Anthropic researchers have noted in their work on constitutional AI and interpretability, this performative transparency can become another layer of opacity. Amodei and his team have highlighted how models can learn to satisfy our interpretability criteria while their actual mechanisms remain as inscrutable as ever, turning the quest for alignment into a sophisticated game of appearances rather than genuine understanding.

[26] D. Amodei, *op. cit*. We can see this already in many Anthropic team experiments and research papers. For example, see R. Chen, A. Arditi, H. Sleight, O. Evans, J. Lindsey, *Persona Vectors: Monitoring and Controlling Character Traits in Language Models*, 5.8.2025, https://arxiv.org/abs/2507.21509 (access: 19.10.2025).

[27] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford 2014, pp. 127–144.

[28] "(…) it seems that the march towards superhuman intelligence is unstoppable, but success might be the undoing of the human race. Not all is lost, however. We have to understand where we went wrong and then fix it" (S.J. Russel, *Human Compatible: Artificial Intelligence and the Problem of Control*, New York 2019, p. 11). Cf. E.P. Torres, *Human Extinction: A History of the Science and Ethics of Annihilation*, New York 2024. See more about cultural dimension of the catastrophic visions in E. Horn, *The Future as Catastrophe: Imagining Disaster in the Modern Age*, New York 2018.

[29] See S.J. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, New Jersey 2010, pp. 1034–1039.

ical specifications that AI systems can optimize for.[30] As systems become more capable, the gap between their formal objectives and the intended human values they should serve can widen, creating what Amodei and others term the "specification-reality gap" – a challenge for ensuring that AI systems act as genuine extensions of human will (the extended mind paradigm[31]) rather than autonomous optimizers that may inadvertently undermine human welfare.[32] More recently, theorists including T. LaCroix[33] and M. Suleyman[34] have emphasized that alignment cannot be solved through technical means alone but requires integrating cultural, legal, and philosophical frameworks. LaCroix argues that alignment is ultimately a value interpretation problem requiring contextual sensitivity to diverse human normative frameworks. He identifies multiple "axes of value alignment" that must be simultaneously considered: the temporal axis (how values evolve over time), the cultural axis (how values differ across societies), the individual-collective axis (tensions between personal autonomy and social good), and the explicit-implicit axis (the gap between stated and revealed preferences). LaCroix emphasizes that AI systems must navigate what he terms "normative pluralism" – the reality that equally valid but potentially conflicting value systems coexist within and across human communities.[35] Meanwhile Suleyman frames alignment as a governance challenge requiring new institutions and cross-cultural coordination mechanisms.[36] What is more, he reflects on the related problem to alignment, that is the containment problem. Containment, as Suleyman articulates it, represents the challenge of controlling the proliferation and impact of AI technologies once they are developed. The paradox is that as AI becomes cheaper, more powerful, and more accessible, traditional containment mechanisms (export controls, regulatory frameworks, technical safeguards) become increasingly ineffective. Unlike nuclear technology, which requires specialized materials and infrastructure, AI can be replicated, modified, and deployed with minimal resources once the underlying knowledge exists. This creates what Suleyman calls an "impossible dilemma": aggressive containment risks creating techno-authoritarian surveillance states that stifle innovation and

[30] B. Christian, *The Alignment Problem: Machine Learning and Human Values*, New York 2020, pp. 291–320.

[31] A. Clark, D.J. Chalmers, *The Extended Mind*, "Analysis" 1998, vol. 58(1), pp. 7–19.

[32] B. Christian, *op. cit.*, pp. 287–295.

[33] T. LaCroix, *Artificial Intelligence and the Value Alignment Problem: A Philosophical Introduction*, Peterborough 2025.

[34] See M. Suleyman, M. Bhaskar, *The Coming Wave: AI, Power and the Twenty-First Century's Greatest Dilemma*, London 2023. It should be mentioned that Suleyman can be biased in his views because of his actual (as for August 2025) work as CEO of Chief of AI in Microsoft (before he worked with Hassabis in Google Deepmind), albeit his book was written when he wasn't working for the biggest tech companies.

[35] T. LaCroix, *op. cit.*, part 2 (*Axes of Value Alignment*).

[36] M. Suleyman, M. Bhaskar, *op. cit.*, pp. 35–50.

human freedom, while open development risks catastrophic misuse by malicious actors.[37] Thus, the containment problem emphasises that even perfectly aligned AI systems could destabilize society if we cannot control who accesses them and how they are deployed. That's the dilemma.

What unites these perspectives is a recognition that as AI systems increasingly mediate human experience and decision-making across diverse cultural contexts, ensuring their alignment with human values requires not just technical safeguards but also interpretive frameworks that can bridge the gap between algorithmic optimization and human sense-making across diverse cultural and legal traditions. The escalating trajectory from ANI (artificial narrow intelligence) to AGI/ASI (artificial general/super intelligence) intensifies these challenges exponentially.[38] While, e.g., I. Sutskever envisions artificial superintelligence as an inevitable progression that will fundamentally transform civilization, G. Hinton warns of existential risks from systems that could soon surpass human cognitive capabilities. Contrasting perspective emerges from researchers like A. Narayanan and S. Kapoor. They argue that framing AI as a path to superintelligence obscures more pressing concerns, suggesting we should instead understand AI as "normal technology" subject to

[37] *Ibidem*, p. XIII.

[38] The labels defining this trajectory, such as AGI, are themselves a subject of intense debate, often carrying more weight in market and narrative contexts than in strict scientific ones. These terms have become a perceived "layer" of progress, strategically employed by figures like Altman to frame the technological frontier and generate anticipation, for instance, around upcoming releases like GPT-5. This raises concerns about the goals behind using such fluid terminology. For a critical perspective on this phenomenon within OpenAI, see K. Hao, *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*, New York 2025. Meanwhile, setting aside the hype, one of the substantive technical avenues toward more advanced AI capabilities involves fundamental shifts in machine reasoning. Current systems rely heavily on techniques like Chain-of-Thought (CoT), but new approaches are emerging. For a recent promising attempt to move beyond current limitations, see G. Wang, J. Li, Y. Sun, X. Chen, C. Liu, Y. Wu, M. Lu, S. Song, Y.A. Yadkori, *Hierarchical Reasoning Model*, 4.8.2025, https://arxiv.org/abs/2506.21734 (access: 20.7.2025). However, even though it's a label, it refers to milestones, which can be achieved by AI systems in the future. The underlying pursuit of general intelligence can be understood more rigorously through scientific and mathematical frameworks rather than corporate milestones. A more sufficient way to assess progress is to refer to formal theories, such as the work of computer scientist M. Hutter on Universal AI. His theory provides a mathematical blueprint for a "perfectly rational" agent (named AIXI) capable of learning to solve any computable problem. Unlike a commercial product, this theoretical model serves as a stable, scientific benchmark for what true general intelligence could be, offering a way to measure real-world systems that is independent of corporate roadmaps and product releases. Also worth seeing is an article from 2018 with indications towards AGI Safety – when it finally appears (if it hasn't already). See T. Everitt, G. Lea, M. Hutter, *AGI Safety Literature Review*, 21.5.2018, https://arxiv.org/abs/1805.01109 (access: 20.5.2025); R. Hutter, M. Hutter, *Chances and Risks of Artificial Intelligence – a Concept of Developing and Exploiting Machine Intelligence for Future Societies*, "Applied System Innovation" 2021, vol. 4(2).

typical engineering constraints, social impacts, and regulatory needs.[39] Whether we conceptualize AI as an exceptional, potentially transcendent technology shapes how we approach questions of control, interpretation, and human agency.

Moreover, knowledge extraction from AI systems is a critical yet underexplored dimension of the alignment problem – one that intersects with questions about how technical systems embody and perpetuate particular modes of understanding. K. Crawford's reveals extraction as a foundational logic governing contemporary AI development: from the mining of lithium for data centres to the harvesting of human labour for data annotation, and the appropriation of creative works for training datasets.[40] Extractive paradigm extends to knowledge itself. Artificial intelligence systems do not contain information; they actively transform human knowledge into computational forms, raising profound questions about whose knowledge gets preserved, whose gets erased, and how cultural and contextual meanings become flattened into statistical patterns. As Crawford emphasised, this is key to understand anatomy of AI.[41] Also S. Zuboff's concept of surveillance capitalism provides another lens for understanding knowledge extraction in AI alignment.[42] Just as surveillance capitalism created unprecedented asymmetries of knowledge – where platforms know individuals better than they know themselves – AI systems create asymmetries where models may encode patterns and relationships that humans cannot access or comprehend. Thus, alignment problem is not about ensuring A systems pursue human goals, but maintaining meaningful human agency in systems that increasingly extract, process, and act upon knowledge in ways that exceed human understanding (e.g. technics used to understand text by AI systems as embeddings[43]). As well G. Marcus's advocacy for transparency in AI development reflects growing recognition that these systems' opacity perpetuates and amplifies existing power imbalances.[44] The convergence of knowledge extraction, social inequality, and institutional capture suggests that alignment cannot be achieved

[39] A. Narayanan, S. Kapoor, *AI as Normal Technology: An Alternative to the Vision of AI as a Potential Superintelligence*, 15.4.2025, https://knightcolumbia.org/content/ai-as-normal-technology (access: 15.5.2025). See eidem, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*, Princeton 2024.

[40] K. Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven 2021, passim.

[41] K. Crawford, V. Joler, *Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources*, https://anatomyof.ai (access: 13.8.2025).

[42] See S. Zuboff, *The Age of Surveillance Capitalism*, New York 2019.

[43] Embeddings are a key element of natural language processing in artificial intelligence. See R. Jha, C. Zhang, V. Shmatikov, J.X. Morris, *Harnessing the Universal Geometry of Embeddings*, 18.5.2025, https://arxiv.org/abs/2505.12540 (access: 25.6.2025).

[44] G.F. Marcus, *Taming Silicon Valley: How We Can Ensure That AI Works for Us*, Cambridge 2024. Cf. A. Becker, *More Everything Forever: AI Overlords, Space Empires, and Silicon Valley's Crusade to Control the Fate of Humanity*, New York 2025.

through technical transparency alone. Requires structural reforms that address how AI development and deployment reinforce existing asymmetries of power.[45]

Analysed insights suggest that interpretability research, as advocated by Amodei, may be necessary but insufficient for addressing alignment challenges. True alignment might require not just the ability to peer inside AI systems (the "MRI for AI") but a critical rethinking of how knowledge is extracted, processed, and redeployed. Then how do we ensure that the process of extracting knowledge to, through and from AI systems doesn't reproduce the extractive logics that Crawford identify as central to contemporary digital capitalism?[46]

As so, I reckon that the key to problems of alignment lays in culture and its normative dimensions.

CULTURAL AND LEGAL CHALLENGE

Large language models trained on vast datasets to generate natural language have revolutionized how we access information through AI assistants like ChatGPT. While these systems excel at tasks from text summarization to question answering, their behaviour varies dramatically based on design, training data, and implementation. Variations that extend far beyond technical capabilities into the realm of cultural perspectives and embedded values.

Dimension of cultural challenges emerges from three interconnected factors: the algorithmic monoculture dominating today's LLM landscape, the specific datasets feeding these models, and the post-training refinement processes that shape their responses. The result? AI assistants that inadvertently embody the cultural norms and biases of their creators while amplifying the dominant perspectives found in their training corpora. Nevertheless, the traditional response to cultural challenges – embedding predetermined cultural values from static databases – represents a flawed, one-directional approach. More promising methodologies treat cultural alignment as an ongoing, bidirectional dialogue between human values and AI behaviour.[47] It provides the question how cultural values manifest

[45] G. Marcus, E. Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust*, New York 2019.

[46] Cf. D. Acemoglu, *The Simple Macroeconomics of AI*, "Economic Policy" 2025, vol. 40(121), pp. 13–58.

[47] A. Glaese et al., *Improving Alignment of Dialogue Agents via Targeted Human Judgements*, 28.8.2022, https://arxiv.org/abs/2209.14375 (access: 20.6.2025). Although differences reflect broader global attitudes toward AI but are complicated by risks such as human over-reliance on AI systems. See F. dell'Acqua, *Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters*, 2023, https://www.almendron.com/tribuna/wp-content/uploads/2023/09/falling-asleep-at-the-whee.pdf (access: 15.5.2025); F. Dell'Acqua, C. Ayoubi, H. Lifshitz, R. Sadun, E. Mollick,

in real-world AI interactions and how user patterns actively reshape system responses over time. It gains importance in the light of recent studies of this issue, where we can explicitly see how generative models reflect the ideology of their creators and possibly impacts info-creation[48] and worldviews of the users. In the age of unprecedented changes,[49] this one could be very impactful for contemporary politics[50] and future of democracy.[51] Furthermore, the alignment problem is fundamentally social and political, requiring principles that can earn widespread public trust and legitimacy. An opaque system, whose reasoning is inscrutable to its users and overseers, can never achieve this. Explainability, the ability of an AI to justify its decisions in human-understandable terms, is the critical bridge to securing this social contract. For AI to be integrated safely into high-stakes domains like law, medicine, or governance, it must be accountable. This accountability is impossible without clear explanations. It brings the field of Explainable AI (XAI) into focus as a complementary perspective. XAI, along with the closely related goal of interpretability, seeks to open the "black box" of complex models to make their decision-making processes transparent and understandable to humans. It is a crucial tool for addressing the socio-cultural issues outlined above.[52] Moreover, what constitutes a "good" explanation is itself culturally dependent, requiring that

L. Mollick, Y. Han, J. Goldman, H. Nair, S. Taub, K. Lakhani, *The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise*, "Harvard Business School Strategy Unit Working Paper" 2025, no. 25-043.

[48] A. Kostikova, Z. Wang, D. Bajri, O. Pütz, B. Paaßen, S. Eger, *LLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models*, 25.5.2025, https://arxiv.org/abs/2505.19240 (access: 20.8.2025); S. Vijay, A. Priyanshu, A.R. KhudaBukhsh, *When Neutral Summaries Are Not That Neutral: Quantifying Political Neutrality in LLM-Generated News Summaries*, 13.10.2024, https://arxiv.org/abs/2410.09978 (access: 2.8.2025).

[49] I refer here to Z.B. Simon and the concept of "unprecedented change", which we are witnessing and experiencing nowadays. See Z.B. Simon, *History in Times of Unprecedented Change: A Theory for the 21st Century*, London 2019; idem, *The Epochal Event: Transformations in the Entangled Human, Technological and Natural Worlds*, Cham 2020.

[50] Especially for possibilities of political persuasion and creating misinformation. On political persuasion, see K. Hackenburg, B.M. Tappin, L. Hewitt, E. Saunders, S. Black, H. Lin, C. Fist, H. Margetts, D.G. Rand, C. Summerfield, *The Levers of Political Persuasion with Conversational AI*, 18.7.2025, https://arxiv.org/abs/2507.13919 (access: 21.7.2025). Also in this text there is interesting finding about emerging trade-off between persuasiveness and factual accuracy in AI models reveals a troubling paradox at the heart of advanced language model development. On misinformation and dynamics of "alternative facts" (or in case of AI so called 'hallucinations'), see C. O'Connor, J.O. Weatherall, *The Misinformation Age: How False Beliefs Spread*, New Haven 2020, pp. 147–186.

[51] "AI, as it is currently developed and used, risks undermining the fundamental principles and knowledge basis on which our democracies are built and does not contribute to the common good" (M. Coeckelbergh, *Why AI Undermines Democracy and What to Do About It*, Cambridge 2024, p. 120).

[52] Also here we should be aware of culture bias in XAI research. See U. Peters, M. Carman, *Cultural Bias in Explainable AI Research: A Systematic Analysis*, "Journal of Artificial Intelligence Research" 2024, vol. 79, pp. 971–1000.

XAI methods be sensitive to the cognitive and cultural contexts of their users to be truly effective.[53] From a legal and democratic standpoint, this transparency is a prerequisite for accountability. It enables meaningful regulatory oversight and aligning algorithmic behaviour with legal frameworks that demand fairness and non-discrimination, such as the "right to explanation" in the EU's GDPR.[54]

Realisation of this vision demands radical transparency in AI development: detailed demographic reporting of RLHF (Reinforcement Learning from Human Feedback) evaluators, open documentation of training methodologies, and inclusive participation in model creation.[55] Only through such openness can the industry move beyond the current paradigm where ostensibly global AI systems reflect remarkably narrow cultural perspectives. Creating tools that serve humanity's diversity rather than homogenizing it. Thus, cultural diversity as well as legal pluralism create complex challenges for AI alignment beyond technical solutions. These challenges emerge in how AI systems interpret and operationalize human values across different cultural contexts, and how regulatory frameworks govern these interpretations. Recent research highlights profound cultural biases in AI systems, particularly large language models (LLMs). These systems inevitably encode cultural, political, and moral perspectives of their developers, training data, and fine-tuning processes.

Relationships between AI systems and cultural values can be systematically analysed through established anthropological frameworks. For example, Masoud and his team provide compelling evidence that LLMs exhibit measurable biases across Hofstede's six cultural dimensions: power distance, individualism/collectivism, uncertainty avoidance, masculinity/femininity, long-term orientation, and indulgence/restraint. Their analysis demonstrates that leading AI systems consistently favour low power distance, high individualism, low uncertainty avoidance,

[53] On the need for culturally-aware explanations, see D. Saha, A. Chattopadhyay, A.K. Singh, P.P. Talukdar, *Towards Culturally-Aware and Explainable AI: A Survey*, [in:] *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 2024, pp. 985–997. Also important to AI Alignment and XAI is improvement in prompt engineering technics, which enhances capabilities of generative models. Prompt literacy seems to be a key skill to align AI, e.g. fine tune it, to human goals. See E. Jahani, B.S. Manning, J. Zhang, H.-Y. TuYe, M. Alsobay, C. Nicolaides, S. Suri, D. Holtz, *As Generative Models Improve, People Adapt Their Prompts*, 19.7.2024, https://arxiv.org/abs/2407.14333v1 (access: 30.7.2025).

[54] For an analysis of the legal demand for explainability, particularly in the European context, see B. Goodman, S. Flaxman, *European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'*, "AI Magazine" 2017, vol. 38(3), pp. 50–57.

[55] At the moment canonical for the alignment problem is CIRL framework. CIRL means cooperative inverse reinforcement learning and it is a partial-information game with two agents, human and robot, were both are rewarded according to the human's reward function. It addresses value alignment through, as optimal CIRL solutions produce behaviours like active learning and teaching, as well as communicative actions. It makes alignment more likely to be successful. See D. Hadfield-Menell, A. Dragan, P. Abbeel, S. Russell, *Cooperative Inverse Reinforcement Learning*, 9.6.2016, https://arxiv.org/abs/1606.03137 (access: 15.8.2025).

and medium-term orientation – cultural preferences associated with Western, particularly Anglo-American, societies.[56] When operating in high power distance, collectivist societies with different approaches to uncertainty and time orientation, fundamental alignment failures occur despite technical accuracy.

Meanwhile cultural biases create particularly acute challenges in legal contexts, where normative frameworks vary substantially across jurisdictions. AI systems are trained primarily on English-language texts (as well legal), thus LLMs might demonstrate systematic biases toward common law reasoning patterns even when operating in civil law jurisdictions.[57] When analysing identical legal scenarios, AI systems demonstrate tendency to apply common law principles of precedent even within strict civil law jurisdictions where statutory interpretation should predominate.[58] Legal challenges extend beyond jurisdictional differences into deeper questions of how different legal traditions conceptualize foundational principles like justice, rights, and responsibility. AI systems tend to operationalize Western conceptions of individual rights even when deployed in cultural contexts that prioritize collective harmony or family obligations over individual freedoms.[59] Cultural alignment in LLMs could create particularly problematic issues in domains like family law, where cultural and legal frameworks are deeply interconnected. Regulatory frameworks attempting to address these challenges face their own cultural limitations. Comparative analysis of EU[60] and Korean[61] AI Acts reveals that regulatory approaches embed cultural assumptions about risk, responsibility, and appropriate governance mechanisms. Generally individualistic focus in European regulations contrasts with more collective, harmony-oriented Asian approaches, creating meta-regulatory alignment challenges for global AI governance.[62]

---

[56] Due to cultural distance embedded in the dataset based on English texts. See R. Masoud, Z. Liu, M. Ferianc, P.C. Treleaven, M.R. Rodrigues, *Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions*, [in:] *Proceedings of the 31st International Conference on Computational Linguistics*, eds. O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B.D. Eugenio, S. Schockaert, Abu Dhabi 2025, pp. 8474–8503.

[57] Due to cultural distance embedded in the dataset based on English texts. Thus language-specific fine-tunning significantly affects cultural response patterns. See *ibidem.*

[58] F. Ariai, G. Demartini, *Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges*, 25.10.2024, https://arxiv.org/abs/2410.21306 (access: 5.8.2025).

[59] Y. Tao, O. Viberg, R.S. Baker, R.F. Kizilcec, *Cultural Bias and Cultural Alignment of Large Language Models*, "PNAS Nexus" 2023, vol. 3(9), p. 346.

[60] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (OJ L 2024/1689, 19.6.2024).

[61] Basic Act on the Development of Artificial Intelligence and Establishment of Trust https://cset.georgetown.edu/publication/south-korea-ai-law-2025 (access: 6.8.2025).

[62] All legal analyses are based on the legal position as of 10 August 2025.

Nevertheless, regulating under uncertainty has become the defining challenge of AI governance, as F. G'sell demonstrates in her comprehensive analysis of global regulatory approaches.[63] The exponential acceleration in AI development since ChatGPT's release in late November 2022 has created a fundamental temporal mismatch: governments must craft regulations with incomplete information about technologies whose impacts remain largely unknown, yet waiting for perfect knowledge may prove catastrophically late. This uncertainty is compounded by the dual nature of AI's promise – from revolutionary breakthroughs (economical, medical, educational, etc.) to existential risks (extinction, etc.) – making it impossible to predict whether today's regulatory decisions will enable innovation or prevent disaster. Most critically, the emergence of general-purpose AI models defies traditional sector-specific regulation, as these systems can be applied across countless unforeseen contexts. Each potentially carrying its own cultural interpretations of harm, benefit, and acceptable risk.[64]

Law, as a normative system, is based on the assumption of shared, intersubjective understanding of concepts such as intent, causality, and responsibility. Thus, "cultural sense-making practices" constitute a challenge for AI law and regulation. The actions of AI systems, arising from statistical correlations, shatter communities of meaning. As a result, attempts to regulate AI (e.g. in the AI Act) and enforce law (e.g. regarding liability for harm) encounter an interpretive barrier: How can we apply law created for human actions to the "acts" of machines whose logic is alien to us?

AI safety framework and community within it can be perceived as an answer – a concept and commune of explaining and addressing properly challenges of "thinking machines". Enumerating is not possible, we need general injunction in case of actions with any kind of large impact of AI.[65] What is more, theoretical work of the AI safety community is now colliding with the practical demands of

[63] F. G'sell, *Regulating under Uncertainty: Governance Options for Generative AI*, 2024, p. 10.

[64] What is more, legal theory itself needs urgency of interpretability and reference to concepts, which refer to this issue both in interpretive/hermeneutical and practical way (e.g. R. Dworkin, J. Rawls, L. Petrażycki, L. Nowak). Turn to interpretability needs adequate thought framework in each dimension. I mean, that this requires drawing on humanistic legal theories that prioritize interpretation and social context over mere rule application. Key concepts would include Dworkin's "law as integrity", Rawls's framework of "justice as fairness" for auditing bias, and the Polish school of legal theory represented by Petrażycki (psychological dimensions of law) and Nowak (social systems modelling). Other relevant thinkers from different traditions include Germany's J. Habermas (communicative action and law's legitimacy) and R. Alexy (law as practical argumentation), W. Fikentscher (anthropology of law), and Portuguese A. Castanheira Neves (methodological problems of legal interpretation). Even the positivist theory of H.L.A. Hart, particularly his concept of the "internal point of view", poses a challenge to whether a non-human agent can truly participate in a legal system.

[65] Worth reading is the paper foundational for this approach – S. Armstrong, B. Kevinstein, *Low Impact Artificial Intelligences*, 30.5.2017, https://arxiv.org/abs/1705.10720 (access: 13.8.2025).

legal governance.[66] Regulations transforms AI alignment from an abstract technical problem into a concrete legal compliance requirement. Regulatory frameworks might effectively transform the alignment problem from a philosophical and technical challenge into a legal compliance requirement. Developers must now demonstrate not just that their systems work, but that it works correctly according to legally mandated definitions of human values and preferences – definitions that vary significantly across jurisdictions. Thus, technical safety research is no longer an isolated academic pursuit. Yet, it is 2025 and while I'm writing these words, we still cannot explicitly express, how to find pragmatic and adequate answer, how to align AI to human values.

However, the legal context adds another layer of complexity to alignment challenges. The EU AI Act, as the world's first comprehensive AI regulation, establishes a risk-based approach that categorizes AI systems into different tiers, with "high-risk applications" facing the strictest requirements. These high-risk systems, including those used in e.g. critical infrastructure, employment decisions, law enforcement, jurisprudence and healthcare, must demonstrate not only technical safety but also alignment with fundamental rights. The AI Act mandates that such systems undergo rigorous conformity assessments, maintain comprehensive documentation, and provide explanations for their decision-making processes. Nevertheless, it creates a practical paradox: how can developers ensure compliance when, as Amodei notes, "we have no idea, at a specific or precise level, why [AI systems – K.K.] make the choices it does – why it chooses certain words over others, or why it occasionally makes a mistake despite usually being accurate"?[67]

A broad international consensus has emerged on the necessity of aligning AI with human values and societal goals through regulations. In September 2021, the United Nations Secretary-General called for AI regulation to ensure alignment with "shared global values".[68] That same month, the People's Republic of China published ethical guidelines requiring AI to respect shared human values and remain under human control.[69] Similarly, a March 2021 report from the U.S. National Security Commission on Artificial Intelligence stated that AI systems

[66] It is precisely this need for formal assurance that elevates the importance of technical research like that of V. Krakovna at DeepMind. Her work on methods for penalizing unintended side effects represents a tangible approach to translating abstract legal prohibitions into computable, verifiable constraints on an AI's behavior. See V. Krakovna, L. Orseau, R. Kumar, M. Martic, S. Legg, *Penalizing Side Effects Using Stepwise Relative Reachability*, 4.6.2018, https://arxiv.org/abs/1806.01186 (access: 14.8.2025).

[67] D. Amodei, *op. cit*.

[68] United Nations Secretary-General, *Our Common Agenda*, 2021.

[69] National New Generation Artificial Intelligence Governance Specialist Committee, Ethical Norms for New Generation Artificial Intelligence, 21.10.2021.

must align with national goals and values, including safety and trustworthiness.[70] Furthermore within the European Union, this principle has been legally codified, as AI systems must align with the doctrine of substantive equality to comply with non-discrimination law.[71]

Therefore, the EU AI Act represents an attempt to codify these challenges through "regulation by proxy" – instead of directly regulating the internal state of model alignment. It seeks to apply verifiable requirements regarding data, risk management, and human oversight. Article 14 of this Act, requiring effective human oversight of high-risk systems is a direct response to concerns about autonomous, misaligned AI behaviours in critical social domains. Specifically, Article 9 mandates continuous, iterative risk identification and mitigation processes throughout the system's lifecycle, while Article 10 addresses bias directly by requiring training datasets to be "relevant, representative, free of errors and complete" with explicit obligations to examine data for potential biases.

Analysis of the legal culture domain must be holistic. In common law systems, where precedent and case-by-case reasoning dominate, AI supporting judicial decisions must provide particularized reasoning that engages with the specifics of the case at hand. In common law systems like the United Kingdom, where precedent and case-specific reasoning are paramount, the focus is on contestable, particularized explanations.[72] This principle is reflected in the 2023 guidance on AI for the judiciary from the Lord Chief Justice of England and Wales, which emphasizes that judges retain ultimate responsibility and any AI-assisted analysis must be intelligible and reviewable.[73] In contrast, civil law systems such as France prioritize fidelity to statutory requirements. A study by the French Conseil d'État stressed that for AI to be lawful, it must primarily demonstrate consistency with established legal codes and principles, focusing on systematic compliance rather than bespoke, case-specific justifications.[74]

---

[70] National Security Commission on Artificial Intelligence, *Final Report*, 2021, https://www.dwt.com/-/media/files/blogs/artificial-intelligence-law-advisor/2021/03/nscai-final-report--2021.pdf (access: 15.8.2025).

[71] M. De Vos, *The European Court of Justice and the March Towards Substantive Equality in European Union Anti-discrimination Law*, "International Journal of Discrimination and the Law" 2020, vol. 20(1), pp. 62–87; R.L. Poe, *Why Fair Automated Hiring Systems Breach EU Non-Discrimination Law*, 7.11.2023, https://arxiv.org/abs/2311.03900 (access: 25.7.2025).

[72] The United Kingdom's National AI Strategy, also from September 2021, explicitly acknowledges the long-term risks of non-aligned Artificial General Intelligence.

[73] Courts and Tribunals Judiciary, *Artificial Intelligence (AI): Guidance for Judicial Office Holders*, 12.12.2023, https://www.judiciary.uk/wp-content/uploads/2023/12/AI-Judicial-Guidance.pdf (access: 19.10.2025).

[74] Conseil d'État, *Artificial Intelligence and Public Action: Building Trust, Serving Performance*, Paris 2022. A French summary is available at https://www.conseil-etat.fr/publications-colloques/

Need for transparency and accountability is underscored by landmark European court rulings. For example, a Dutch court's 2020 decision to outlaw the SyRI (System Risk Indication)[75] welfare fraud detection system was not due to its technical failings. But because its opaque, risk-scoring mechanism was deemed a violation of the European Convention on Human Rights (ECHR), making its logic incomprehensible and its outcomes unaccountable.[76] This principle is a cornerstone of the EU AI Act. Specifically, Article 52 of the AI Act imposes transparency obligations on deployers of high-risk systems, requiring that affected individuals be provided with clear and adequate information. However, the interpretation of what constitutes "sufficiently transparent" information will inevitably be shaped by local legal norms, proving that even with harmonized law, cultural contingency remains a key factor in the practical governance of AI.

In contrast, South Korea's framework with its "high-impact AI systems" takes a more substantive approach, requiring actual demonstration of value compatibility with specific cultural and social norms through "algorithmic auditing" and "value alignment certification". While the EU focuses on preventing discrimination and ensuring fairness through technical safeguards and human oversight, the Korean framework goes further by demanding positive proof that AI systems embody culturally specific values – a distinction that highlights the tension between universal human rights (EU approach) and culturally relative interpretations of ethical behaviour (Korean approach). Korean regulations explicitly address the alignment problem. It's accomplished by requiring that high-impact systems demonstrate compatibility with Korean cultural values and social norms.[77]

National ambition is powerfully demonstrated by South Korea's sovereign AI initiative, a state-led project aiming to rival the U.S. and China by 2027. Even though "sovereign AI" is the recent path, which is developing in many parts of the world. The global landscape of artificial intelligence is increasingly defined by AI nationalism, a phenomenon where nations strategically leverage AI for geopolitical, economic, and cultural advantage. This trend has given rise to the pursuit of "sovereign AI" sovereignty, as countries seek to avoid dependency on foreign tech-

---

etudes/intelligence-artificielle-et-action-publique-construire-la-confiance-servir-la-performance (access: 12.7.2025).

[75] Judgment of the District Court of The Hague of 5 February 2020 in the case of *System Risk Indication (SyRI)*, C/09/550982/HA ZA 18-388.

[76] See Human Rights Watch, *Netherlands: Landmark Court Ruling Against Welfare Fraud Detection System*, 5.2.2020, https://www.hrw.org/news/2020/02/05/netherlands-landmark-court-ruling-against-welfare-fraud-detection-system (access: 12.7.2025). The case, brought by a coalition of NGOs, successfully argued that the SyRI system violated Article 8 ECHR.

[77] See D.H. Park, E. Cho, Y. Lim, *A Tough Balancing Act: The Evolving AI Governance in Korea*, "East Asian Science, Technology and Society: An International Journal" 2024, vol. 18(2), pp. 135–154.

nology and align AI development with their own national interests and values. As of August 2025, the world is fracturing into distinct regulatory and strategic blocs, moving far beyond a one-size-fits-all approach to AI governance.[78] AI sovereignty is determined by its access to critical technologies like advanced semiconductors. The U.S. formalised this hierarchy through its three-tier AI chip export policy, which stratifies nations based on their access to critical hardware. This policy is a clear exercise in managing a geopolitical chokepoint – the highly concentrated supply chain for AI chips. As strategist E. Fishman argues, by controlling access to essential technologies from firms like Nvidia, Washington can "cajole foreign governments and businesses into embracing standards for the responsible use of AI", while transforming chip access into a primary instrument of foreign policy and technological containment.[79]

However, the U.S. approach is characterized by a deep commitment to market-led innovation, viewing AI as a critical engine for economic growth and national security. The government's AI Action Plan prioritizes investment and public-private partnerships over heavy-handed regulation.[80] This philosophy can be described as "discontainment" – a strategy focused on unleashing domestic innovation while simultaneously using economic leverage to contain rivals. Case study of the "discontainment" was the issue of "One Big Beautiful Bill Act", where a 10-year moratorium on state-level AI regulations was proposed, which finally was defeated in Senate, but in the end put a ground for AI Action Plan. Ideologically, the plan mandates that AI systems be purged of "bias" by revising the NIST AI Risk Management Framework to eliminate references to concepts like disinformation and DEI. Externally, the strategy is a new technological cold war: aggressively exporting the full "American AI stack" to allies while using strengthened export controls to cut off rivals like China from advanced technology.

Nevertheless, China's model is the antithesis of the American one. It is a top-down, state-centric approach where AI development is tightly controlled and explicitly directed to serve national strategic goals, from social governance to military modernization. Beijing has implemented a comprehensive suite of regulations that require AI service providers to obtain licenses, undergo security reviews, and ensure their models' outputs align with socialist values and do not challenge state

[78] K. Payne, *The Geopolitics of AI*, "The RUSI Journal" 2024, vol. 169(5), pp. 54–55.

[79] "To cajole foreign governments and businesses into embracing standards for the responsible use of AI, Washington could ban Nvidia and other U.S. tech firms from transacting with anyone that refuses to adopt these standards" (E. Fishman, *Chokepoints: American Power in the Age of Economic Warfare*, New York 2025, p. 422).

[80] This market-centric view also runs into complex constitutional questions, particularly regarding the regulation of AI-generated content and its intersection with free speech under the First Amendment. See C.R. Sunstein, *Artificial Intelligence and First Amendment*, "George Washington Law Review" 2024, vol. 92(6).

narratives.[81] Yet, in a sophisticated strategic play, China also promotes open-source models as a tool of soft power. The case of DeepSeek-V2, a powerful open-source reasoning model, illustrates this dual strategy: by providing a high-performance alternative to Western models, China builds global dependency and establishes its technology as a viable standard, complicating the simple narrative of a closed vs open AI ecosystem.

The global trajectory of AI regulation is now transitioning from abstract ethical pronouncements to the establishment of concrete legal frameworks, though evolution is far from uniform. We are witnessing the clear emergence of distinct regulatory philosophies, led by the EU's rights and risk-based AI Act, the United States' innovation-focused, sector-specific strategy, and China's state-driven, control-oriented model. Caught between these spheres of influence, other nations (e.g. Japan, India, Brazil, Singapore) are forging hybrid approaches that reflect their unique strategic priorities and legal traditions. In general, it signals the end to the era of AI's unregulated "Wild West".[82]

Law, as a "cultural technology", must evolve alongside the systems it regulates. Create frameworks capable of adaptation in the face of emergent AI behaviours while preserving cultural diversity in the interpretation of human values. The withdrawn AI Liability Directive (AILD) would have addressed the "black box" problem through presumptions of causality and rights of access to evidence, but its failure signals political reluctance to impose the radical transparency necessary to resolve accountability gaps in AI systems. Product Liability Directive (PLD) extends the definition of "product" to explicitly include software and AI systems, incorporating them into strict liability frameworks where the injured party need not prove fault, only that the product was "defective" and caused harm. Article 15 of the AI Act further mandates appropriate levels of accuracy, robustness, and cybersecurity, with robustness against adversarial attacks representing a direct legal response to known technical pathways leading to intentional misalignment.

The EU AI Act explicitly establishes that GDPR takes precedence in cases of regulatory collision, recognizing data protection as a fundamental human right while positioning the AI Act primarily as product safety legislation. Article 10 (5) of the AI Act permits the processing of "special categories of personal data" (such as racial/ethnic origin or health data) for monitoring, detecting, and correcting

[81] N. Karpiuk-Wawryszuk, K. Kasprowicz, *Legal Cultures and Strategies for Implementing Artificial Intelligence Regulations: Case Studies of the United States, People's Republic of China and European Union*, "Teka Prawnicza" 2025, vol. 18(1), pp. 131–147.

[82] See Global AI Regulation Tracker, https://www.techieray.com/GlobalAIRegulationTracker (access: 20.6.2025). Worth tracking are also texts by L. Jarovsky on her newsletter: https://www.luizasnewsletter.com/?utm_campaign=profile_chips (access: 10.8.2025).

biases in high-risk AI systems, provided that appropriate GDPR legal bases (such as explicit consent) and safeguards are met.

Nevertheless, the EU AI Act's defines an AI system as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Article 3).

And it is, from a behavioural economics perspective, emphasis on systems that "exhibit adaptiveness after deployment" and generate outputs that "influence physical or virtual environments" captures virtually all meaningful open source AI projects, from foundation models to specialized tools. This broad scope triggers compliance obligations that behavioural economics predicts will create a "regulatory chill effect", where the cognitive burden of compliance, combined with liability concerns, disincentivizes participation in open source AI development. Thus, the AI Act's approach to open source represents a profound misunderstanding of innovation incentives and collaborative dynamics. The Act's requirements, set to take effect in August 2025, impose obligations on open source AI providers that misalign with the decentralized, iterative nature of open source development.

Recent behavioural economics research emphasizes that AI biases are "highly context-dependent" presenting significant challenges for traditional liability frameworks.[83] If bias mitigation strategies that work in financial decision-making fail to transfer to employment decisions, then holding AI model creators accountable becomes less effective. Instead, attention shifts to deploying companies that implement models in specific use cases, requiring them to conduct rigorous bias audits, maintain transparency regarding AI utilization and ensure compliance with anti-discrimination regulations. Bias manifestation and mitigation are highly dependent on real-world application contexts. The Open Source Initiative has highlighted that the AI Act's requirements – including detailed documentation, risk assessments, and conformity procedures – impose costs that volunteer maintainers and small organizations cannot bear.[84] The AI Act's attempt to apply product liability frameworks to collectively-developed, continuously-evolving open source models represents a category error that behavioural economics would predict – it will possibly lead to strategic withdrawal from the European market by key open source projects

[83] M. Schreiber, *Bias in Large Language Models – and Who Should Be Held Accountable*, 13.2.2025, https://law.stanford.edu/press/bias-in-large-language-models-and-who-should-be-held-accountable (access: 10.8.2025).

[84] GitHub, *Supporting Open Source and Open Science in the EU AI Act*, https://github.blog/wp-content/uploads/2023/07/Supporting-Open-Source-and-Open-Science-in-the-EU-AI-Act.pdf (access: 10.8.2025).

(e.g. Polish *Bielik* or French *Mixtral*). It might create innovation dead zones and increasing market concentration among large commercial providers who can afford compliance costs. Consequently, it would lead to erosion of open source culture in AI development in Europe.

Regulatory landscape emerges as even more urgent, as it may be the only viable path toward meeting these legal requirements for high-risk AI systems while maintaining the technological capabilities that make AI valuable. And here lies the major problem. In spite of open source, collaborative AI research, the market is dominated by big technological companies run by "cyberlords". The "cyberlords" as critics aptly describe them, now gatekeep both the technology and the discourse around its regulation. OpenAI serves here as a good example – what began as democratized innovation has crystallized into oligopolistic control. Once committed to open research – have pivoted toward closed, commercial models. It undermines promise of accessible AI development. Big Tech's approach to the EU AI Act reveals calculated strategic positioning. As of 2 August 2025, the pattern is telling: Mistral, OpenAI, Anthropic, and Microsoft have indicated intention to sign the EU AI Act's Code of Practice for general-purpose AI, while Meta conspicuously abstains.[85] Selective participation demonstrates how voluntary frameworks enable regulatory arbitrage, because companies sign when convenient, abstain when costly.[86]

As the AI Act's Preamble articulates aspirations: "The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (the 'Charter'), including democracy, the rule of law and environmental protection, to protect against the harmful effects of AI systems in the Union, and to support innovation".

Yet Big Tech's actual practices systematically contradict these principles. Firstly, optimize for market dominance, not democratic values. Then prioritize shareholder returns over fundamental rights and extract value while minimizing tax obligations through complex international structures.[87] Furthermore, AI alignment faces fundamental limitations when confronting entities with resources exceeding

[85] Code of Practice for AI, https://code-of-practice.ai/?section=safety-security (access: 10.8.2025).

[86] See more about "digital colonialism" made by Big Tech companies: A. Becker, *op. cit.*; S. Czubkowska, *Bógtechy. Jak wielkie firmy technologiczne przejmują władzę nad Polską i światem*, Kraków 2025, pp. 16–35.

[87] See G. Zucman, *The Hidden Wealth of Nations: The Scourge of Tax Havens*, Chicago 2015.

many nation-states. Big Tech companies do not circumvent regulations; they shape the regulatory environment itself. They fund research institutions, employ former regulators, and influence policy through sophisticated lobbying operations. The revolving door between Silicon Valley and Brussels ensures that regulations arrive pre-compromised. Big Tech companies simultaneously advocate for AI safety while consolidating market power that makes meaningful regulation impossible. Amazon, Google, Microsoft, and Meta control the cloud infrastructure essential for AI development. They acquire potential competitors before they pose threats. They lobby against regulations that would limit their data collection practices. Tax avoidance strategies further expose the hypocrisy. These companies benefit from public infrastructure, educated workforces, and legal systems while contributing minimally to public coffers. Google's "Double Irish with a Dutch Sandwich" structure, Amazon's Luxembourg arrangements, and Microsoft's Puerto Rico subsidiaries exemplify systematic avoidance of social obligations.

The pursuit of trustworthy AI thus becomes a contradiction. How can systems be trustworthy when controlled by entities that systematically evade accountability? How can AI serve democratic values when its development concentrates power in anti-democratic structures? The gap between the AI Act's aspirations and Big Tech's practices reveals not regulatory failure. It creates regulatory impossibility under current power arrangements. Big Tech companies dominate AI safety research, defining what "trustworthy AI" means ensuring definitions align with their business models. They champion interpretability research that maintains their competitive advantages while resisting transparency requirements that would expose their practices.

The most pressing challenges center on privacy violations and copyright infringement – areas where Big Tech's practices most egregiously violate stated principles. Personal data is being harvested at unprecedented scales, converting privacy invasion into profit centers. Companies train AI systems on copyrighted content without permission, claiming fair use while building commercial empires on others' creative work. The Code of Practice becomes, in this context, a fig leaf for systemic non-compliance. It creates an illusion of self-regulation while enabling continued extraction. Companies that sign gain reputational benefits without meaningful constraints. Those that refuse, like Meta, signal their unwillingness to accept even voluntary limitations.

The gulf between Big Tech's self-serving interpretations of "fair use" and creators' rights has become untenable. Courts must now determine whether the massive appropriation of creative works can hide behind the shield of "transformation" or whether the scale and commercial nature of this extraction demands a reconsideration of how AI systems acquire their training data. The landmark

cases like *The New York Times v. OpenAI*,[88] the publishing industry's legal action against Anthropic[89] and the music industry's lawsuits against Suno[90] are forcing a re-examination of the fair use doctrine. This legal principle allows for the limited use of copyrighted material without permission under a four-factor test, with the most critical factor in these cases being whether the use is "transformative" – that is, whether it repurposes the original work for a new, different objective rather than merely superseding it. AI companies argue that training a model is inherently transformative, as the goal is not to reproduce the original works but to teach an AI to recognize patterns. Conversely, rights holders contend that when an AI's output directly competes with their work, for instance, by generating summaries of news articles or creating music in a specific artist's style, the use is substitutive, not transformative, and thus constitutes infringement. While good practices are rare, some companies are already moving in this direction; for example, ElevenLabs has proactively secured licensing deals for the music used to train its AI, providing a potential blueprint for a future where innovation and copyright compliance are not mutually exclusive.[91]

The conclusions reached by American courts in these initial cases will have profound and far-reaching consequences. Should the judiciary broadly accept the "transformative use" defence for AI training, it would solidify a market-driven, innovation-first approach, significantly lowering the barrier to entry for AI development by reducing the need for costly and complex licensing agreements. It alters economic calculus for AI companies and also champions a model where creators' rights and consent are central. However, if courts rule in favour of rights holders, as the early tide of judicial opinion suggests may happen – particularly given the growing judicial engagement with AI's capabilities and limitations, as exemplified by Judge Kevin Newsom's thoughtful exploration of LLMs' interpretive potential in *Snell v. United Specialty Insurance* – it would compel a systemic change towards a licensing-first model.[92] Newsom has emerged as a notable judicial voice in understanding AI technology, conducting what he calls "mini-experiments" with ChatGPT and other

[88] United States District Court for the Southern District of New York, *The New York Times Co. v. Microsoft Corp., et al.*, No. 1:2023cv11195, 27 December 2023.

[89] See M. Hiltzik, *Here's the Number That Could Halt the AI Revolution in Its Tracks*, 25.7.2025, https://www.latimes.com/business/story/2025-07-25/heres-the-number-that-could-halt-the-ai-revolution-in-its-tracks (access: 10.8.2025).

[90] United States District Court for the District of Massachusetts, *Universal Music Group et al. v. Suno Inc.*, No. 1:24-cv-10893, 24 June 2024.

[91] Eleven Music, see https://x.com/elevenlabsio/status/1952754097976721737 (access: 5.8.2025).

[92] Judgment of the United States Court of Appeals for the Eleventh Circuit of 22 May 2024, *Snell v. United Specialty Insurance Co.*, No. 22-12581 (11th Cir. 2024; J. Newsom, concurring; exploring the potential use of large language models like ChatGPT in legal interpretation and acknowledging that LLMs train on "mind-bogglingly enormous amount of raw data taken from the internet").

generative AI programs to help interpret legal terms. In *United States v. Deleon*, he queried multiple AI models about the ordinary meaning of "physically restrained" observing that the programs produced slight variations in their answers. And these variations "accurately reflects real people's everyday speech patterns" demonstrating the models' ability to predict ordinary meaning. As he claims, LLMs "may well serve a valuable auxiliary role as we aim to triangulate ordinary meaning".[93]

However, the legal domain of AI alignment emerges as a nexus where philosophical principles of distributive justice intersect with technological governance. It can be exemplified by the application of Lockean property theory to contemporary AI development. When tech companies appropriate digital commons – publicly available data, open-source code, and collective human knowledge – through computational processing to create proprietary AI systems, they engage in a modern form of labour-mixing that parallels Locke's original formulation of property acquisition, yet this appropriation raises fundamental alignment concerns when viewed through the lens of the Lockean Proviso's requirement that "enough and as good" remain for others.[94] The alignment problem thus transcends technical specifications of goal preservation and value loading to encompass broader societal impacts: whether AI systems concentrate power asymmetrically, degrade the quality of the information commons through synthetic content proliferation, or create barriers to entry that prevent equitable access to AI capabilities. Legal frameworks such as the Korean AI Act or EU AI Act, the proposed AI Liability Directive, and emerging national AI strategies represent institutional attempts to operationalize the Proviso's normative constraints, establishing ex ante requirements for transparency, risk assessment, and fundamental rights impact assessments that effectively mandate consideration of whether AI development leaves sufficient opportunity for others to benefit from the digital commons.[95] Thus, it can be socio-legal imperative to preserve the commons from which these systems derive their capabilities, thereby preventing the emergence of what might be termed "alignment enclosure" – where technically aligned systems nonetheless violate broader principles of distributive justice by exhausting or degrading shared resources upon which future innovation and societal flourishing depend.

These finds a practical application in emerging domains of rule-making: property rights controversies (evidenced by lawsuits against ChatGPT, Suno, and Midjourney over training data appropriation), responsibility allocation (spurring startups like Armilla to develop AI insurance frameworks), and knowledge extraction dis-

[93] Judgment of the United States Court of Appeals for the Eleventh Circuit of 21 June 2024, *United States v. Deleon*, No. 23-10478 (J. Newsom, concurring).

[94] See J. Locke, *Second Treatise of Government*, (1689). Locke outlines his labour theory of property in chapter V *Of Property*. The specific proviso requiring that "enough, and as good" be left for others is articulated in section 27.

[95] Cf. P. Dolniak, T. Kuźma, A. Ludwiński, K. Wasik, *Sztuczna inteligencja w wymiarze sprawiedliwości. Między prawem a algorytmami*, Warszawa 2024.

putes. The latter category has become particularly contentious as copyright lawsuits against OpenAI highlight the tensions between existing intellectual property regimes and AI systems that extract, transform, and regenerate human knowledge at unprecedented scale. There are profound property rights controversies, evidenced by the wave of lawsuits against generative AI companies like OpenAI (ChatGPT), Suno, and Midjourney. These cases center on the unauthorized appropriation of copyrighted text, music, and images for training data, challenging the very foundation of digital ownership. Second is the critical domain of responsibility allocation. The inherent unpredictability of AI systems has spurred the creation of novel solutions, with startups developing AI insurance and warranty frameworks designed to distribute liability when these complex systems inevitably fail. Finally, the domain of knowledge extraction has become particularly contentious. Copyright lawsuits, such as the prominent case against Meta for training its LLaMA models on their books, underscore the acute tensions between existing intellectual property regimes and AI systems that extract, transform, and regenerate human knowledge at an unprecedented scale.[96] These practical battlegrounds – in courtrooms and boardrooms – are where the abstract challenges of pluralistic AI alignment and the rule of law, as described by N. Caputo and K. Frazer, are becoming concrete realities.[97]

But do these legal battles over static outputs truly prepare us for the imminent challenge of future development (e.g. autonomous AI agents, robotics)? If we struggle to assign liability for a single piece of generated content, how can our legal system hope to trace causation back through a complex chain of an agent's independent, probabilistic actions in the real world? When an agent acts on a vague user prompt and causes financial or physical harm, does the culpability lie with the user who gave the command, the corporation that deployed the system, or does the agent's very autonomy create an accountability vacuum our current laws cannot fill? Is this the shift from regulating content to governing conduct not the true frontier of the alignment problem, demanding a legal and ethical paradigm for which we are profoundly unprepared?

DISCUSSION AND CONCLUSIONS

Cultural sense-making practices represent the collective, unwritten norms, expectations, and interpretive frameworks shaped by a community's history, language,

[96] See J. Horwitz, *Meta's AI Rules Have Let Bots Hold 'Sensual' Chats with Kids, Offer False Medical Info*, 14.8.2025, https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines (access: 15.8.2025).

[97] N. Caputo, *Rules, Cases, and Reasoning: Positivist Legal Theory as a Framework for Pluralistic AI Alignment*, 28.10.2024, https://arxiv.org/abs/2410.17271 (access: 3.4.2025).

and value systems. In the context of this analysis, AI alignment is not a universal technical problem – it depends on the local, cultural "software of the mind" (Hofstede) that determines whether AI actions in education or the job market will be perceived as comprehensible, legitimate, and trustworthy. Studying these practices shows why attempts to create universal ethical regulations for AI must account for the pluralism of human ways of understanding the world. The alignment problem emerges here as a possibility to build bridges between these two worlds of meaning in such crucial spheres as work, education, and law.

Addressing cultural and legal challenges requires moving beyond technical solutions to develop what I term "cultural alignment infrastructures" – frameworks that systematically adapt AI systems' explanations to diverse contexts while maintaining core functionality. Rather than seeking universal alignment principles, these infrastructures acknowledge necessary cultural adaptation while providing mechanisms to identify and address fundamental value conflicts when they arise.[98]

I presume that this approach effectively reframes the alignment challenge as a task of computational cultural modelling. The goal is to equip AI systems not with a single set of universal values, but with a library of context-specific models representing different cultural logics for reasoning and communication. This task, however, presents several core analytical problems. First is the problem of representation: how to model a dynamic culture without reducing it to a static, and potentially harmful, caricature (in other words: how to create adequate models of the world)? Second is the problem of bidirectionality: the model must account for the co-evolutionary process where the AI not only adapts to a culture but also actively reshapes it. Finally, there is the problem of value incommensurability. For instance, consider an AI agent moderating content depicting a caricature of a religious figure. A cultural model adapted for French law, grounded in the principle of *laïcité* (secularism), would classify this as protected free speech, as blasphemy is not a crime. In contrast, a model adapted for Polish law would have to consider Article 196 of the Criminal Code, which criminalizes "offending religious feelings". The very same content would be flagged as a potential criminal offense. Here, the values of secular free expression and the protection of religious sentiment are incommensurable; they cannot be resolved by a common metric. The system's infrastructure must therefore recognize when these models lead to irreconcilable ethical commands and flag this conflict for human intervention. Rather than attempt to resolve it algorithmically.

[98] A project similar to idea presented in M. Bravansky, F. Trhlik, F. Barez, *Rethinking AI Cultural Alignment*, 7.3.2025, https://arxiv.org/abs/2501.07751 (access: 10.6.2025), where authors tried to challenge perspective, that cultural alignment is one-directional, therefore it should be perceived as bidirectional process with understanding the specific context of AI systems.

Future of human development is deeply rooted in the answer to the challenges of alignment problem.[99] Thus, bridging AI's computational logic with human meaning-making systems requires us to examine the sophisticated mechanisms through which human societies have historically coordinated complex collective behaviour and resolved conflicts between different interpretive frameworks. Among these mechanisms, legal systems represent perhaps humanity's most elaborate attempt to codify shared understanding and enable coordinated action across diverse communities.[100] Unlike informal social norms that vary fluidly across contexts, law provides structured frameworks for translating between different cultural logics while maintaining operational coherence. Within this broader social context, law emerges as a particularly refined tool for fostering cooperation among agents and facilitating joint actions.

As far as we know, culture, and especially law, cannot be conceived just in one dimension: technological one. Nowadays AI systems have knowledge but can only imitate experience. To be human is to experience, thus, an AI that only mimics this can never fully grasp the human context of the rules it is asked to follow. We can paraphrase famous article, which is often regarded as breakthrough in machine learning – alignment is not all we need.[101] Effective AI alignment requires a new, transdisciplinary approach integrating technical, cultural, social and legal dimensions. From legal practice point of view, the core problem is not just whether AI will become "singularity", but how it refers to legal cultures and the fundamental inefficiency of traditional regulation when applied to dynamic learning systems.[102] This challenge is perfectly encapsulated by Goodhart's Law, famously known as: "When a measure becomes a target, it ceases to be a good measure". When applied to law, it predicts a cycle where the very act of regulation undermines its own goals. In the context of AI regulation and the alignment problem, it would provide the following steps (from governance perspective):

1. We identify a desirable end goal (safe and beneficial AI systems that do not cause harm).
2. We can't ultimately control or directly measure this end goal, so we pick proxy metrics that seem correlated: compliance checkboxes, safety testing

[99] See United Nations Development Programme, *Human Development Report 2025: A Matter of Choice: People and Possibilities in the Age of AI*, 2025, https://hdr.undp.org/system/files/documents/global-report-document/hdr2025reporten.pdf (access: 10.8.2025).

[100] H.R. Kirk, I. Gabriel, C. Summerfield, B. Vidgen, S.A. Hale, *Why Human–AI Relationships Need Socioaffective Alignment*, "Humanities and Social Sciences Communications" 2025, vol. 12(728).

[101] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, *Attention is All You Need*, 2.8.2023, https://arxiv.org/abs/1706.03762 (access: 20.7.2025).

[102] Cf. K. Frazer, *A Different Alignment Problem: AI, the Rule of Law, and Outdated Legal Institutions and Practices*, "Journal of Business & Technology Law" 2023, vol. 19.

benchmarks, or documentation requirements. This is based on a simple assumption: more compliance metrics passed → safer AI systems.
3. We tell AI companies about these regulations and penalize them for non-compliance (or reward them with market access for compliance).
4. Building genuinely aligned AI systems is really hard. But passing specific benchmark tests, producing required documentation, and checking regulatory boxes is comparatively easy.
5. AI developers optimize for passing the regulatory metrics while potentially missing the deeper safety issues. They might train models specifically to ace safety benchmarks, produce extensive but meaningless documentation, or implement superficial safety features that look good to regulators but do not address fundamental alignment problems. The AI systems appear "safe" on paper but the actual alignment problem remains unsolved (not that the company minds – they have achieved regulatory approval and gained market access).

What then? One of the possibilities – AI Alignment Benchmarking. However, the use of standardized tests to evaluate models, becomes fraught with complexity when applied to the domain of ethics and alignment. The core challenge is that ethics is not a solved problem with quantifiable answers; any attempt to create a universal "ethics benchmark" inevitably embeds a specific metaethical viewpoint and risks falling into the trap of Goodhart's Law as well, where the benchmark score becomes a gamed target rather than a true measure of alignment.[103] This has spurred a search for more foundational approaches. One path, exemplified by F. Chollet's ARC-AGI (Abstract Reasoning Corpus), is to test for genuine fluid intelligence with novel puzzles that resist rote memorization.[104] An alternative, pursued by researchers like R. Rzepka, is to design AI with an inherent "top-down" moral architecture based on normative ethical theories, rather than just testing external behaviour.[105]

Even though these advanced general benchmarks are insufficient for high-stakes, specialized domains like law, which requires more than abstract reasoning. This necessitates the development of domain-specific alignment benchmarks that test

[103] T. LaCroix, A. Luccioni, *Metaethical Perspectives on 'Benchmarking' AI Ethics*, "AI and Ethics" 2025, vol. 5.

[104] GitHub, *Abstraction and Reasoning Corpus for Artificial General Intelligence v1 (ARC-AGI-1)*, https://github.com/fchollet/ARC-AGI (access: 10.8.2025). Abstraction and Reasoning Corpus is a benchmark designed by F. Chollet to measure an AI's fluid intelligence, distinguishing it from rote memorization. It consists of unique, abstract visual reasoning puzzles that a model has never seen before. To solve them, an AI must infer the underlying logic from a few examples and apply it, a task that is easy for humans but has proven exceptionally difficult for LLMs. See ArcPrize, https://arcprize.org/leaderboard (access: 2.7.2025); F. Chollet, M. Knoop, G. Kamradt, B. Landers, *ARC Prize 2024: Technical Report*, 5.12.2024, https://arxiv.org/abs/2412.04604 (access: 2.7.2025).

[105] T. Masashi, R. Rzepka, A. Kenji, *Towards Theory-based Moral AI: Moral AI with Aggregating Models Based on Normative Ethical Theory*, 20.6.2023, https://arxiv.org/abs/2306.11432 (access: 5.8.2025).

for performance within concrete professional and cultural contexts. A leading example is the misalignment classifier, LLMs designed to classify transcripts that represent intentionally misaligned behaviour.[106] However, as researchers from the UK AI Safety Institute have argued, evaluating such classifiers is notoriously difficult because "intentional misalignment" is a fuzzy, psychological concept rather than a crisp, observable outcome. Optimizing against such a classifier often leads to ambiguous edge cases rather than clear-cut failures, making robust adversarial evaluation nearly impossible. It underscores why domain-specific benchmarks are so crucial; by grounding evaluation in concrete professional rules and outputs, they can sidestep the intractable problem of judging abstract intent – especially in the case of autonomous actions (e.g. by AI agents).

In case of law, we already have *CaseLaw Benchmark* from the legal tech startup Gaius, which evaluates an AI's ability to interpret statutes, adhere to judicial precedent, and construct sound legal arguments. By focusing on established legal reasoning rather than abstract morality, such tools provide a more meaningful and practical way to assess an AI's safety and reliability for real-world application.[107] The future of effective alignment benchmarking thus lies in a dual approach: combining robust, general reasoning tests with a suite of highly specialized, domain-specific evaluations.

As for now methods for evaluating artificial intelligence are insufficient for legal and social alignment, a challenge best understood through the legal-theoretic framework of L. Lessig's "code is law" dictum. It means that law regulates technology, but the architecture of that technology becomes a de facto legal system, enforcing norms and shaping social structures.[108] Perhaps most crucially, Lessig's insight that "code is law" reveals the bidirectional relationship between technological architecture and legal governance – suggesting that alignment requires not just regulating AI through law. Recognition how AI systems themselves encode and enforce normative frameworks, potentially can reshape the very legal structures designed to govern them. It also reveals why classic evaluation metrics like the Turing Test[109] and Winograd Schema Challenge,[110] or technical fixes for issues like "hallucinations" are inadequate. They assess surface-level performance or

[106] LessWrong, *Misalignment Classifiers: Why They're Hard to Evaluate Adversarially, and Why We're Studying Them Anyway*, 15.8.2025, https://www.lesswrong.com/posts/jzHhJJq2cFmisRKB2/misalignment-classifiers-why-they-re-hard-to-evaluate (access: 16.8.2025).

[107] *The Case Law Benchmark*, developed by the legal tech startup Gaius, is available at https://www.vals.ai/benchmarks/case_law-02-03-2025 (access: 13.7.2025).

[108] L. Lessig, *Code Is Law: On Liberty in Cyberspace*, "Harvard Magazine" 2000, vol. 102(3); idem, *Code and Other Laws of Cyberspace*, New York 1999.

[109] A. Turing, *Computing Machinery and Intelligence*, "Mind" 1950, vol. 59(236), pp. 433–460.

[110] Designed by H. Levesque to improve Turing Test. The Winograd Schema Challenge requires a machine to resolve an ambiguous pronoun in a sentence that has a nearly identical twin; changing a single "special" word in the sentence alters the correct answer. See H.J. Levesque, *On Our Best Behaviour*, "Artificial Intelligence" 2014, vol. 212, pp. 27–35.

symptoms, failing to address the deeper challenge articulated by philosophical critiques like Searle's Chinese Room argument – that an AI's ability to manipulate syntax does not equate to the semantic understanding necessary for genuine legal interpretation and application. True alignment, therefore, requires benchmarks that assess fidelity to the principles of justice, not just convincing imitation.

Furthermore, we need not traditional governance by bureaucracy – we need constant scientific observation and critical analysis in the unprecedented technological advancements. We have arrived at a moment where the slow march of traditional bureaucracy can no longer keep pace with the exponential leap of machine intelligence. The governance we need is not one of static rules and delayed oversight. There is necessity for constant scientific observation and critical analysis fit for an age of unprecedented change.

Urgency of answering the issues raised by alignment challenge needs to be correctly addressed. As H.-G. Gadamer's dictum states – a truth that legal philosophers like R. Dworkin and C. Neves would recognize as the heart of their own work – all interpretation is, in the end, application.[111] When viewed through this lens, AI ceases to be a passive tool for processing information and becomes an active agent of application. The implication is that embedding our laws, ethics, and values into an AI system is not a neutral act of data entry; it is the inherent pre-configuration of that system's real-world actions. Therefore, every dataset, rule, and objective we provide is fundamentally a blueprint for enactment. It transforms abstract principles into tangible consequences with a speed and scale that challenge our very understanding of cause and effect. Thus, our collective effort to align AI with human values must ultimately grapple with challenges that are deeply human – rooted in the complex, culturally-contingent, and normative nature of our own behaviour (e.g. "collaborative AI" framework[112]).

Nevertheless, alignment, interpretability, and explainability research are not just a technical necessity. It became the democratic imperative. In the world of the code dependent societies,[113] assuring democratic values is the cornerstone of alignment.[114] Artificial intelligence has the potential to reboot our reality, particularly the way

[111] See H.-G. Gadamer, *Truth and Method*, London 1989.

[112] E. Mollick, *Co-Intelligence: Living and Working with AI*, New York 2024; A. Przegalinska, T. Triantoro, *Converging Minds: The Creative Potential of Collaborative AI*, Boca Raton 2024. See also before mentioned CIRL solutions to value alignment.

[113] Reference to classical book of G. Lakoff and M. Johnson in title of interesting book: B. Christian, T. Griffiths, *Algorithms to Live By: The Computer Science of Human Decisions*, Dublin 2017.

[114] Otherwise, we can awake in the chaotic space of misinformation and likely dictatorship. See Ł. Olejnik, *Propaganda: From Disinformation and Influence to Operations and Information Warfare*, New York 2025.

we live our every day life,[115] not less than industrial revolution. Consequently, the very legitimacy of our future institutions will depend on our ability to render these systems to be transparent and accountable to the societies they would serve.[116]

REFERENCES

Literature

Acemoglu D., *The Simple Macroeconomics of AI*, "Economic Policy" 2025, vol. 40(121), **DOI: https://doi.org/10.1093/epolic/eiae042**.

Asimov I., *Runaround*, "Astounding Science Fiction" 1942, no. 3.

Becker A., *More Everything Forever: AI Overlords, Space Empires, and Silicon Valley's Crusade to Control the Fate of Humanity*, New York 2025.

Bennett M., *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs That Made Our Brains*, New York–Boston 2023.

Bengio Y., *Government Interventions to Avert Future Catastrophic AI Risks*, "Harvard Data Science Review" 2024, no. 5 (Special Issue), **DOI: https://doi.org/10.1162/99608f92.d949f941**.

Biagini G., *Towards an AI–Literate Future: A Systematic Literature Review Exploring Education, Ethics, and Applications*, "International Journal of Artificial Intelligence in Education" 2025, **DOI: https://doi.org/10.1007/s40593-025-00466-w**.

Bostrom N., *Deep Utopia: Life and Meaning in a Solved World*, 2024.

Bostrom N., *Superintelligence: Paths, Dangers, Strategies*, Oxford 2014.

Christian B., *The Alignment Problem: Machine Learning and Human Values*, New York 2020.

Christian B., Griffiths T., *Algorithms to Live By: The Computer Science of Human Decisions*, Dublin 2017.

Clark A., Chalmers D.J., *The Extended Mind*, "Analysis" 1998, vol. 58(1), **DOI: https://doi.org/10.1093/analys/58.1.7**.

Coeckelbergh M., *Why AI Undermines Democracy and What to Do About It*, Cambridge 2024.

Conseil d'État, *Artificial Intelligence and Public Action: Building Trust, Serving Performance*, Paris 2022.

Crawford K., *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven 2021, **DOI: https://doi.org/10.12987/9780300252392**.

Czubkowska S., *Bógtechy. Jak wielkie firmy technologiczne przejmują władzę nad Polską i światem*, Kraków 2025.

De Vos M., *The European Court of Justice and the March Towards Substantive Equality in European Union Anti-discrimination Law*, "International Journal of Discrimination and the Law" 2020, vol. 20(1), **DOI: https://doi.org/10.1177/1358229120927947**.

Dell'Acqua F., Ayoubi C., Lifshitz H., Sadun R., Mollick E., Mollick L., Han Y., Goldman J., Nair H., Taub S., Lakhani K., *The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise*, "Harvard Business School Strategy Unit Working Paper" 2025, no. 25-043, **DOI: https://doi.org/10.3386/w33641**.

[115] For example, see M. Murgia, *Code Dependent: How AI Is Changing Our Lives*, London 2024; A. Elliott, *The Culture of AI: Everyday Life and the Digital Revolution*, London 2019.

[116] See. G. Biagini, *Towards an AI–Literate Future: A Systematic Literature Review Exploring Education, Ethics, and Applications*, "International Journal of Artificial Intelligence in Education" 2025.

Dolniak P., Kuźma T., Ludwiński A., Wasik K., *Sztuczna inteligencja w wymiarze sprawiedliwości. Między prawem a algorytmami*, Warszawa 2024.

Elliott A., *Making Sense of AI: Our Algorithmic World*, Cambridge 2022.

Elliott A., *The Culture of AI: Everyday Life and the Digital Revolution*, London 2019, **DOI: https://doi.org/10.4324/9781315387185**.

Fishman E., *Chokepoints: American Power in the Age of Economic Warfare*, New York 2025.

Frazer K., *A Different Alignment Problem: AI, the Rule of Law, and Outdated Legal Institutions and Practices*, "Journal of Business & Technology Law" 2023, vol. 19, **DOI: https://doi.org/10.2139/ssrn.4849429**.

G'sell F., *Regulating under Uncertainty: Governance Options for Generative AI*, 2024, **DOI: https://doi.org/10.2139/ssrn.4918704**.

Gadamer H.-G., *Truth and Method*, London 1989.

Goodman B., Flaxman S., *European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'*, "AI Magazine" 2017, vol. 38(3), **DOI: https://doi.org/10.1609/aimag.v38i3.2741**.

Haidt J., *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, New York 2012.

Hao K., *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*, New York 2025.

Hofstede G., Hofstede G.J., Minkov M., *Cultures and Organizations: Software of the Mind*, New York 2010.

Horn E., *The Future as Catastrophe: Imagining Disaster in the Modern Age*, New York 2018, **DOI: https://doi.org/10.7312/horn18862**.

Hutter R., Hutter M., *Chances and Risks of Artificial Intelligence – a Concept of Developing and Exploiting Machine Intelligence for Future Societies*, "Applied System Innovation" 2021, vol. 4(2), **DOI: https://doi.org/10.3390/asi4020037**.

Karpiuk-Wawryszuk N., Kasprowicz K., *Legal Cultures and Strategies for Implementing Artificial Intelligence Regulations: Case Studies of the United States, People's Republic of China and European Union*, "Teka Prawnicza" 2025, vol. 18(1), **DOI: https://doi.org/10.32084/tkp.9619**.

Kirk H.R., Gabriel I., Summerfield C., Vidgen B., Hale S.A., *Why Human–AI Relationships Need Socioaffective Alignment*, "Humanities and Social Sciences Communications" 2025, vol. 12(728), **DOI: https://doi.org/10.1057/s41599-025-04532-5**.

Korbak T. et al., *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*, 15.7.2025, https://arxiv.org/abs/2507.11473 (access: 20.6.2025).

LaCroix T., *Artificial Intelligence and the Value Alignment Problem: A Philosophical Introduction*, Peterborough 2025.

LaCroix T., Luccioni A., *Metaethical Perspectives on 'Benchmarking' AI Ethics*, "AI and Ethics" 2025, vol. 5, **DOI: https://doi.org/10.1007/s43681-025-00703-x**.

Lem S., *Golem XIV*, Kraków 1981.

Lessig L., *Code and Other Laws of Cyberspace*, New York 1999.

Lessig L., *Code Is Law: On Liberty in Cyberspace*, "Harvard Magazine" 2000, vol. 102(3).

Levesque H.J., *On Our Best Behaviour*, "Artificial Intelligence" 2014, vol. 212, **DOI: https://doi.org/10.1016/j.artint.2014.03.007**.

Locke J., *Second Treatise of Government*, (1689).

Mamak K., *Whether to Save a Robot or a Human: On the Ethical and Legal Limits of Protections for Robots*, "Frontiers in Robotics and AI" 2021, vol. 8, **DOI: https://doi.org/10.3389/frobt.2021.712427**.

Marcus G., *Taming Silicon Valley: How We Can Ensure That AI Works for Us*, Cambridge 2024.

Marcus G., Davis E., *Rebooting AI: Building Artificial Intelligence We Can Trust*, New York 2019.

Masoud R., Liu Z., Ferianc M., Treleaven P.C., Rodrigues M.R., *Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions*, [in:] *Pro-*

ceedings of the 31st International Conference on Computational Linguistics, eds. O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B.D. Eugenio, S. Schockaert, Abu Dhabi 2025.

Mollick E., *Co-Intelligence: Living and Working with AI*, New York 2024.

Murgia M., *Code Dependent: How AI Is Changing Our Lives*, London 2024.

Narayanan A., Kapoor S., *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*, Princeton 2024, **DOI: https://doi.org/10.1515/9780691249643**.

O'Connor C., Weatherall J.O., *The Misinformation Age: How False Beliefs Spread*, New Haven 2020, **DOI: https://doi.org/10.2307/j.ctv8jp0hk**.

Olejnik Ł., *Propaganda: From Disinformation and Influence to Operations and Information Warfare*, New York 2025, **DOI: https://doi.org/10.1201/9781003499497**.

Park D.H., Cho E., Lim Y., *A Tough Balancing Act: The Evolving AI Governance in Korea*, "East Asian Science, Technology and Society: An International Journal" 2024, vol. 18(2), **DOI: https://doi.org/10.1080/18752160.2024.2348307**.

Pasquinelli M., *The Eye of the Master: A Social History of Artificial Intelligence*, London–New York 2023.

Payne K., *The Geopolitics of AI*, "The RUSI Journal" 2024, vol. 169(5), **DOI: https://doi.org/10.1080/03071847.2024.2413274**.

Peters U., Carman M., *Cultural Bias in Explainable AI Research: A Systematic Analysis*, "Journal of Artificial Intelligence Research" 2024, vol. 79, **DOI: https://doi.org/10.1613/jair.1.14888**.

Przegalinska A., Triantoro T., *Converging Minds: The Creative Potential of Collaborative AI*, Boca Raton 2024, **DOI: https://doi.org/10.1201/9781032656618**.

Russel S.J., *Human Compatible: Artificial Intelligence and the Problem of Control*, New York 2019.

Russell S.J., Norvig P., *Artificial Intelligence: A Modern Approach*, New Jersey 2010.

Saha D., Chattopadhyay A., Singh A.K., Talukdar P.P., *Towards Culturally-Aware and Explainable AI: A Survey*, [in:] *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 2024.

Sautoy M. de, *The Creativity Code: Art and Innovation in the Age of AI*, Cambridge 2020, **DOI: https://doi.org/10.4159/9780674240407**.

Simon Z.B., *History in Times of Unprecedented Change: A Theory for the 21st Century*, London 2019, **DOI: https://doi.org/10.5040/9781350095083**.

Simon Z.B., *The Epochal Event: Transformations in the Entangled Human, Technological and Natural Worlds*, Cham 2020, **DOI: https://doi.org/10.1007/978-3-030-47805-6**.

Suleyman M., Bhaskar M., *The Coming Wave: AI, Power and the Twenty-First Century's Greatest Dilemma*, London 2023, **DOI: https://doi.org/10.17104/9783406814143**.

Sunstein C.R., *Artificial Intelligence and First Amendment*, "George Washington Law Review" 2024, vol. 92(6).

Tao Y., Viberg O., Baker R.S., Kizilcec R.F., *Cultural Bias and Cultural Alignment of Large Language Models*, "PNAS Nexus" 2023, vol. 3(9), **DOI: https://doi.org/10.1093/pnasnexus/pgae346**.

Torres E.P., *Human Extinction: A History of the Science and Ethics of Annihilation*, New York 2024, **DOI: https://doi.org/10.4324/9781003246251**.

Turing A., *Computing Machinery and Intelligence*, "Mind" 1950, vol. 59(236), **DOI: https://doi.org/10.1093/mind/LIX.236.433**.

United Nations Secretary-General, *Our Common Agenda*, 2021.

Whorf B., *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, ed. J.B. Carroll, Cambridge 1956.

Zajdel J., *Limes Inferior*, Warszawa 1982.

Załuski W., *Game Theory in Jurisprudence*, Kraków 2014.

Zuboff S., *The Age of Surveillance Capitalism*, New York 2019.

Zucman G., *The Hidden Wealth of Nations: The Scourge of Tax Havens*, Chicago 2015, **DOI: https://doi.org/10.7208/chicago/9780226245560.001.001**.

Online sources

AI Security Institute, *The Alignment Project*, https://alignmentproject.aisi.gov.uk (access: 20.7.2025).

Amodei D., *The Urgency of Interpretability*, 2025, https://www.darioamodei.com/post/the-urgency-of-interpretability (access: 14.5.2025).

Ariai F., Demartini G., *Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges*, 25.10.2024, https://arxiv.org/abs/2410.21306 (access: 5.8.2025).

Armstrong S., Kevinstein B., *Low Impact Artificial Intelligences*, 30.5.2017, https://arxiv.org/abs/1705.10720 (access: 13.8.2025).

ArcPrize, https://arcprize.org/leaderboard (access: 2.7.2025).

Bengio Y. (ed.), *International AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, January 2025, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf (access: 17.10.2025).

Bravansky M., Trhlik F., Barez F., *Rethinking AI Cultural Alignment*, 7.3.2025, https://arxiv.org/abs/2501.07751 (access: 10.6.2025).

Caputo N., *Rules, Cases, and Reasoning: Positivist Legal Theory as a Framework for Pluralistic AI Alignment*, 28.10.2024, https://arxiv.org/abs/2410.17271 (access: 3.4.2025).

Chen R., Arditi A., Sleight H., Evans O., Lindsey J., *Persona Vectors: Monitoring and Controlling Character Traits in Language Models*, 5.8.2025, https://arxiv.org/abs/2507.21509 (access: 19.10.2025).

Chollet F., Knoop M., Kamradt G., Landers B., *ARC Prize 2024: Technical Report*, 5.12.2024, https://arxiv.org/abs/2412.04604 (access: 2.7.2025).

Code of Practice for AI, https://code-of-practice.ai/?section=safety-security (access: 10.8.2025).

Conseil d'État, *Intelligence artificielle et action publique : construire la confiance, servir la performance*, 31.8.2022, https://www.conseil-etat.fr/publications-colloques/etudes/intelligence-artificielle-et-action-publique-construire-la-confiance-servir-la-performance (access: 12.7.2025).

Courts and Tribunals Judiciary, *Artificial Intelligence (AI): Guidance for Judicial Office Holders*, 12.12.2023, https://www.judiciary.uk/wp-content/uploads/2023/12/AI-Judicial-Guidance.pdf (access: 19.10.2025).

Crawford K., Joler V., *Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources*, https://anatomyof.ai (access: 13.8.2025).

Dell'Acqua F., *Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters*, 2023, https://www.almendron.com/tribuna/wp-content/uploads/2023/09/falling-asleep-at-the-whee.pdf (access: 15.5.2025).

Everitt T., Lea G., Hutter M., *AGI Safety Literature Review*, 21.5.2018, https://arxiv.org/abs/1805.01109 (access: 20.5.2025).

Future of Life Institute, *Pause Giant AI Experiments: An Open Letter*, 22.3.2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments (access: 20.7.2025).

GitHub, *Abstraction and Reasoning Corpus for Artificial General Intelligence v1 (ARC-AGI-1)*, https://github.com/fchollet/ARC-AGI (access: 10.8.2025).

GitHub, *Supporting Open Source and Open Science in the EU AI Act*, https://github.blog/wp-content/uploads/2023/07/Supporting-Open-Source-and-Open-Science-in-the-EU-AI-Act.pdf (access: 10.8.2025).

Glaese A. et al., *Improving Alignment of Dialogue Agents via Targeted Human Judgements*, 28.8.2022, https://arxiv.org/abs/2209.14375 (access: 20.6.2025).

Global AI Regulation Tracker, https://www.techieray.com/GlobalAIRegulationTracker (access: 20.6.2025).

Hackenburg K., Tappin B.M., Hewitt L., Saunders E., Black S., Lin H., Fist C., Margetts H., Rand D.G., Summerfield C., *The Levers of Political Persuasion with Conversational AI*, 18.7.2025, https://arxiv.org/abs/2507.13919 (access: 21.7.2025).

Hadfield-Menell D., Dragan A., Abbeel P., Russell S., *Cooperative Inverse Reinforcement Learning*, 9.6.2016, https://arxiv.org/abs/1606.03137 (access: 15.8.2025).

Hastings-Woodhouse S., Kokotajło D., *We Should Not Allow Powerful AI to Be Trained in Secret: The Case for Increased Public Transparency*, 27.5.2025, https://www.aipolicybulletin.org/articles/we-should-not-allow-powerful-ai-to-be-trained-in-secret-the-case-for-increased-public-transparency (access: 20.6.2025).

Hiltzik M., *Here's the Number That Could Halt the AI Revolution in Its Tracks*, 25.7.2025, https://www.latimes.com/business/story/2025-07-25/heres-the-number-that-could-halt-the-ai-revolution-in-its-tracks (access: 10.8.2025).

Horwitz J., *Meta's AI Rules Have Let Bots Hold 'Sensual' Chats with Kids, Offer False Medical Info*, 14.8.2025, https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines (access: 15.8.2025).

Human Rights Watch, *Netherlands: Landmark Court Ruling Against Welfare Fraud Detection System*, 5.2.2020, https://www.hrw.org/news/2020/02/05/netherlands-landmark-court-ruling-against-welfare-fraud-detection-system (access: 12.7.2025).

Jahani E., Manning B.S., Zhang J., TuYe H.-Y., Alsobay M., Nicolaides C., Suri S., Holtz D., *As Generative Models Improve, People Adapt Their Prompts*, 19.7.2024, https://arxiv.org/abs/2407.14333v1 (access: 30.7.2025).

Jarovsky L., *Luiza's Newsletter*, https://www.luizasnewsletter.com/?utm_campaign=profile_chips (access: 10.8.2025).

Jha R., Zhang C., Shmatikov V., Morris J.X., *Harnessing the Universal Geometry of Embeddings*, 18.5.2025, https://arxiv.org/abs/2505.12540 (access: 25.6.2025).

Kokotajło D., Alexander S., Larsen T., Lifland E., Dean R., *AI 2027*, 3.4.2025, https://ai-2027.com (access: 19.10.2025).

Kostikova A., Wang Z., Bajri D., Pütz O., Paaßen B., Eger S., *LLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models*, 25.5.2025, https://arxiv.org/abs/2505.19240 (access: 20.8.2025).

Krakovna V., Orseau L., Kumar R., Martic M., Legg S., *Penalizing Side Effects Using Stepwise Relative Reachability*, 4.6.2018, https://arxiv.org/abs/1806.01186 (access: 14.8.2025).

LessWrong, *Misalignment Classifiers: Why They're Hard to Evaluate Adversarially, and Why We're Studying Them Anyway*, 15.8.2025, https://www.lesswrong.com/posts/jzHhJJq2cFmisRKB2/misalignment-classifiers-why-they-re-hard-to-evaluate (access: 16.8.2025).

Masashi T., Rzepka R., Kenji A., *Towards Theory-based Moral AI: Moral AI with Aggregating Models Based on Normative Ethical Theory*, 20.6.2023, https://arxiv.org/abs/2306.11432 (access: 5.8.2025).

Mishra R., Varshney G., *Exploiting Jailbreaking Vulnerabilities in Generative AI to Bypass Ethical Safeguards for Facilitating Phishing Attacks*, 16.7.2025, https://arxiv.org/abs/2507.12185 (access: 2.8.2025).

Moral Machine Platform, *MIT Media Lab*, https://www.moralmachine.net (access: 5.8.2025).

Narayanan A., Kapoor S., *AI as Normal Technology: An Alternative to the Vision of AI as a Potential Superintelligence*, 15.4.2025, https://knightcolumbia.org/content/ai-as-normal-technology (access: 15.5.2025).

Natale S., Biggio F., Arora P., Downey J., Fassone R., Grohmann R., Guzman A., Keightley E., Ji D., Obia V., Przegalinska A., Raman U., Ricaurte P., Villanueva-Mansilla E., *Global AI Cultures: How a Cultural Focus Can Empower Generative Artificial Intelligence*, 8.8.2025, https://cacm.acm.org/opinion/global-ai-cultures (access: 15.8.2025).

National Security Commission on Artificial Intelligence, *Final Report*, 2021, https://www.dwt.com/-/media/files/blogs/artificial-intelligence-law-advisor/2021/03/nscai-final-report--2021.pdf (access: 15.8.2025).

Poe R.L., *Why Fair Automated Hiring Systems Breach EU Non-Discrimination Law*, 7.11.2023, https://arxiv.org/abs/2311.03900 (access: 25.7.2025).

RenAIssance Foundation, *The Rome Call for AI Ethics*, 28.2.2020, https://www.romecall.org/the-call (access: 20.6.2025).

Schreiber M., *Bias in Large Language Models – and Who Should Be Held Accountable*, 13.2.2025, https://law.stanford.edu/press/bias-in-large-language-models-and-who-should-be-held-accountable (access: 10.8.2025).

United Nations Development Programme, *Human Development Report 2025: A Matter of Choice: People and Possibilities in the Age of AI*, 2025, https://hdr.undp.org/system/files/documents/global-report-document/hdr2025reporten.pdf (access: 10.8.2025).

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I., *Attention is All You Need*, 2.8.2023, https://arxiv.org/abs/1706.03762 (access: 20.7.2025).

Vijay S., Priyanshu A., KhudaBukhsh A.R., *When Neutral Summaries Are Not That Neutral: Quantifying Political Neutrality in LLM-Generated News Summaries*, 13.10.2024, https://arxiv.org/abs/2410.09978 (access: 2.8.2025).

Wang G., Li J., Sun Y., Chen X., Liu C., Wu Y., Lu M., Song S., Yadkori Y.A., *Hierarchical Reasoning Model*, 4.8.2025, https://arxiv.org/abs/2506.21734 (access: 20.7.2025).

Legal acts

Basic Act on the Development of Artificial Intelligence and Establishment of Trust (Korean AI Act).

National New Generation Artificial Intelligence Governance Specialist Committee, Ethical Norms for New Generation Artificial Intelligence, 21.10.2021.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (OJ L 2024/1689, 19.6.2024).

United Kingdom, National AI Strategy, September 2021.

Case law

Judgment of the District Court of The Hague of 5 February 2020 in the case of *System Risk Indication (SyRI)*, C/09/550982/HA ZA 18-388.

Judgment of the United States Court of Appeals for the Eleventh Circuit of 22 May 2024, *Snell v. United Specialty Insurance Co.*, No. 22-12581.

Judgment of the United States Court of Appeals for the Eleventh Circuit of 21 June 2024, *United States v. Deleon*, No. 23-10478.

United States District Court for the District of Massachusetts, *Universal Music Group et al. v. Suno Inc.*, No. 1:24-cv-10893, 24 June 2024.

United States District Court for the Southern District of New York, *The New York Times Co. v. Microsoft Corp., et al.*, No. 1:2023cv11195, 27 December 2023.

ABSTRAKT

W artykule analizie poddano problem dostosowania sztucznej inteligencji (*AI alignment problem*), stanowiący fundamentalne wyzwanie w zakresie międzykulturowej komunikacji między ludzkimi ramami interpretacyjnymi a algorytmiczną optymalizacją. Autor wskazuje, że skuteczne dostosowanie AI wymaga integracji praktyk kulturowego nadawania sensu oraz ram prawnych, które różnią się w poszczególnych społeczeństwach. Analiza prowadzi do wniosku, że obecne próby regulacyjne, w tym rozporządzenie o sztucznej inteligencji Unii Europejskiej (EU AI Act) oraz krajowe strategie dotyczące AI, napotykają trzy powiązane ze sobą wyzwania: zapewnienie interpretowalności decyzji algorytmicznych, zarządzanie indeterminizmem właściwym systemom AI oraz rozwiązywanie kontrowersji związanych z pozyskiwaniem wiedzy. Poprzez analizę nowych zjawisk, takich jak agenci AI, a także zjawiska „przechwytywania" regulacji przez globalne korporacje technologiczne (Big Tech) oraz wzrostu tzw. nacjonalizmu AI, autor dowodzi, że niepowodzenia w procesie dostosowania wynikają nie tylko z ograniczeń technicznych, lecz również z niewystarczającego uwzględnienia różnorodnych kulturowych logik interpretacyjnych. Autor proponuje ramy umożliwiające adaptację systemów AI do odmiennych kontekstów przy jednoczesnym zachowaniu ich podstawowej funkcjonalności. W konkluzji wskazuje, że rozwiązanie problemu dostosowania wymaga zastosowania obliczeniowego modelowania kulturowego, zdolnego do nawigowania w warunkach pluralizmu wartości. Autor ostrzega, że bez integracji technicznych mechanizmów bezpieczeństwa z kulturowymi ramami społeczeństw systemy AI mogą stać się narzędziami eksploatacji i kontroli, a nie partnerami służącymi dobru społecznemu.

Słowa kluczowe: dostosowanie; sztuczna inteligencja; interpretowalność; regulacje; nadawanie sensu; kultura

Ministerstwo Nauki i Szkolnictwa Wyższego

Doskonała Nauka II

Projekt dofinansowany ze środków budżetu państwa, przyznanych przez Ministra Edukacji i Nauki w ramach Programu „Doskonała Nauka II"