

Aligned for Whom? A Brazilian Research Agenda for Culturally Grounded AI Safety Evaluation

Livy Real*

Kunumi Institute

Universidade Federal do Amazonas
livy@kunumi.com

Roberta Viola*

Kunumi Institute

Universidade Federal de Minas Gerais
roberta@kunumi.com

Abstract

Frontier models were constructed within a few specific cultural and linguistic contexts and validated against a particular set of values. We argue this limitation operates at three levels: technical (safety evaluations fail silently in non-English languages), theoretical (alignment frameworks inherit an implicit universalism about human values that obscures cross-cultural variation), and governance (the power to declare a model safe remains concentrated in private laboratories unaccountable to the populations they affect). Brazilian Portuguese, spoken by over 215 million people, sits at the intersection of these three failures. We propose a Brazilian research agenda structured around three fronts: (i) culturally-grounded dangerous capability benchmarks constructed participatorily in Brazilian Portuguese; (ii) adaptive evaluation methodologies that keep evaluation responsive to evolving models and locally defined risks; and (iii) a distributed safety evaluation infrastructure in which nationally accountable institutions assess frontier models before large-scale deployment.

1 Introduction

Measuring the social impact of AI presupposes agreement on what counts as harm, benefit, and risk. We argue that this agreement presupposition is false and that its falseness is most visible from the periphery of the global AI ecosystem. The benchmarks used to evaluate dangerous capabilities in frontier models were constructed within a specific cultural and linguistic context, optimized for a specific population, and validated against a specific set of values. That context is, in general, Anglophone and North American.

This paper argues that this is not merely an oversight but a structural flaw, politically motivated, with compounding consequences at three distinct

levels. At the **technical level**, dangerous capability evaluations fail silently in non-English languages (Papadimitriou et al., 2023). At the **theoretical level**, superalignment frameworks inherit an implicit universalism about human values that obscures deep cross-cultural variation (Arzberger et al., 2024). And at the **governance level**, the power to declare a frontier model safe for deployment remains concentrated in a handful of private laboratories (Kennedy et al., 2025), none of which are accountable to the populations they serve.

Brazilian Portuguese, spoken by over 215 million people and the most widely spoken language in the Southern Hemisphere, sits at the intersection of all three failures. This positioning is especially significant given Brazil’s growing centrality in global AI adoption: recent studies place the country among the top one or two global users of generative AI¹, while other reports describe it as the world’s most “AI-addicted” nation, with ChatGPT traffic increasing by more than 14 times in 2025².

The technical problem is now well-documented: safety alignment is brittle and highly optimized for English, fading rapidly as linguistic distance increases (Wang et al., 2024). More critically, investigations of sycophancy, manipulation, and unsafe behavior in non-English contexts remain overwhelmingly Anglocentric, limiting our understanding of whether Reinforcement Learning from Human Feedback (RLHF) uniformly protects against harm across the global linguistic spectrum (Levy et al., 2025). The problem runs deeper than align-

¹<https://www.meioemensagem.com.br/marketing/brasil-e-o-segundo-pais-que-mais-usa-ia-generativa-no-mundo>;

<https://analise.com/noticias/brasil-e-o-pais-que-mais-usa-ia-nas-empresas-aponta-pesquisa-da-pwc>

²<https://www.ibtimes.com/ai-addiction-grows-check-point-ranks-brazil-no1-workers-handling-tasks-bots-3794837>

*Equal contribution.

ment degradation. Persuasion, manipulation, and strategic deception are not language-neutral concepts; they are embedded in social norms, political histories, and communicative repertoires that vary substantially across cultures. No existing evaluation suite captures this distinction.

This technical gap reflects a deeper theoretical problem. The superalignment literature operates under an implicit assumption that there exists a coherent set of human values toward which models can be aligned (Christiano et al., 2017; Leike et al., 2018). Yet perspectives from the Global South, considering technological accessibility, equitable development, and data sovereignty, remain largely absent from mainstream alignment frameworks (Ayana et al., 2025). This is not a problem that better translation can solve. It requires reconceiving alignment itself as a pluralist project, one in which values are treated as distributed and culturally situated rather than as a fixed target to be optimized.

The governance consequences are direct. Most AI safety frameworks are written in Washington, Brussels, and Beijing, with priorities set without participation from those who will bear their effects (Kaur, 2025). The 2025 BRICS Leaders’ Statement on AI Governance, adopted in Rio de Janeiro, explicitly affirmed that all states retain the right to design domestic regulatory regimes aligned with their own priorities (BRICS, 2025), a normative commitment that currently has no technical infrastructure to support it.

Although the alignment problem is global (Deng et al., 2026), its normative foundations cannot be treated as universally stable, since even concepts such as safety, privacy, and sovereignty are culturally shaped (Wong, 2020). For this reason, we propose not a global solution, but a Brazilian research agenda that may offer comparative value to other nations facing analogous historical and institutional conditions, particularly across Latin America.

We propose a three-part research agenda to close this gap: (i) *culturally-grounded dangerous capability benchmarks* in Brazilian Portuguese, constructed participatorily rather than translated from English originals; (ii) an *adaptive evaluation methodology* capable of translating revisable, locally defined risk categories into evaluation outputs as models and social conditions evolve; and (iii) a *safety evaluation infrastructure* in which frontier models undergo independent assessment by nation-

ally accountable institutions before large-scale deployment. Together, these constitute the foundation for a sovereign Brazilian contribution to one of the most consequential research agendas of our time.

2 Related Work

Research on AI ethics and safety has focused on evaluating harmful behaviors, alignment failures, and moral reasoning in LLMs. The present work proposes a culturally grounded, collaborative benchmarking framework to be built and used by the public sectors interested in in-group safety, so we briefly introduce previous works that inspire us.

In terms of moral measurement, Simmons (2023) introduces *moral mimicry*, showing that LLMs adapt their moral reasoning to align with political identities, particularly along the North American liberal-conservative divide. The study demonstrates that models shift their moral language to align with human moral profiles, but its moral constructs, datasets, and evaluation are rooted in North American and Anglophone contexts. Although the experimental design is innovative, the work remains centered on Western moral frameworks.

On culturally adapted benchmarks, Sahoo et al. (2024) introduce *IndiBias*, a benchmark for evaluating social biases in multilingual language models within the Indian context, incorporating locally relevant social dimensions such as caste, religion, and stereotypes. The benchmark grounds evaluation in context-specific social structures, recognizing that categories of harm are not universal. However, it relies on a predefined and static set of categories, with cultural grounding limited to benchmark construction and annotator participation, lacking broader collaborative or community-driven evaluation processes. Our proposal shares the goal of contextualized evaluation while extending it toward more collaborative and dynamic approaches.

A related direction moves evaluation beyond private laboratories toward collaborative and public approaches. Kennedy et al. (2025) propose cooperative public red teaming, in which governments, civil society, and multidisciplinary participants contribute to evaluation efforts, addressing the limitations of in-house evaluations based on restricted perspectives. However, the framework still relies on predefined risk categories and transferable methodologies, with participation focused on assessing existing harms rather than defining context-specific harms and values. Our proposal

similarly advocates independent and publicly accountable evaluation while incorporating culturally situated values and harms as part of the evaluation process itself.

3 The Informational Power Structure Behind AI Deployment

The emergence of large-scale AI systems does not occur in a social vacuum. It unfolds within a specific logic of accumulation that Zuboff (2015) has theorized as *surveillance capitalism*: an economic order in which human experience is claimed as free raw material for behavioral data extraction, prediction, and modification. Under this logic, populations occupy a structurally ambiguous position. Zuboff's original account emphasizes the company's formal indifference to its users: Silicon Valley's hyperscale companies have broken the Fordist reciprocity between capital and labor, no longer depending on populations as employees or as a consuming mass. Google employs tens of thousands where General Motors once employed millions; finally, its customers are advertisers, not users.

The deployment of AI at scale, however, introduces a complication that Zuboff's framework anticipates but does not fully develop. When AI systems move from commercial platforms into the infrastructure of public life — into courts, hospitals, transit systems, and welfare allocation through large contracts directly signed with governments — populations become not merely data sources but also **captive** consumers of the resulting models. This double position is not, however, symmetrical across populations. The behavioral data that trains frontier models is drawn disproportionately from Anglophone, high-connectivity, and high-income populations; the models built on that surplus are then deployed globally, including to populations whose linguistic and cultural practices contributed minimally to the training corpus. In this configuration, surveillance capitalism does not simply require populations twice over, it requires different populations for each role: some as the primary suppliers of behavioral surplus that shapes model behavior, others as the downstream subjects upon whom those models are imposed, often without meaningful consent, transparency, or recourse. The extractive circuit is global, but its benefits and its risks are distributed unequally along lines that closely follow existing geopolitical hierarchies.

This dynamic is no longer speculative. Palantir Technologies, whose platforms integrate large-scale data analysis into state decision-making, derives approximately 55% of its revenue from government contracts globally, including deployments across defense, immigration enforcement, and public health agencies (Palantir Technologies Inc., 2025). In Brazil, the pattern is equally visible: Google's infrastructure is present in several Brazilian Justice Courts³, powers traffic management systems in several cities⁴, and even in public efforts to combat racism and hate speech⁵. The National Telecommunications Agency (Anatel), for example, integrated a generative-AI assistant into its Electronic Information System (SEI) in 2024, using, at that time, GPT-4o through Microsoft Azure OpenAI Service to analyze, summarize, translate, review, and generate institutional documents⁶. Other AI companies are also present in the Brazilian governmental sector, such as in the Unified Health System⁷ (SUS), without transparency about the used models. These are not peripheral experiments, but represent the gradual privatization of the computational backbone of public administration.

In the Brazilian judiciary, this situation is even more concerning. Different courts have adopted systems based on different proprietary AI models⁸, creating a fragmented landscape of AI adoption across the judicial system. The National Council of Justice (CNJ), for example, reports partnerships between Brazilian courts and multiple major AI providers, including Microsoft/Copilot, Google/Gemini, and OpenAI/ChatGPT (Conselho Nacional de Justiça, 2025). In some cases, courts are also testing or making multiple AI tools available to judicial staff for overlapping tasks. Yet the existence of governance requirements does not

³<https://cloud.google.com/customers/intl/pt-br/tjgo>

⁴<https://www.cartacapital.com.br/toquetec/google-leva-ia-aos-semaforos-de-sao-paulo-como-o-green-light-pode-reduzir-transito-poluicao-e-tem-po-perdido-no-volante/>

⁵<https://www.gov.br/secom/en/latest-news/2023/brazil-and-google-establish-a-partnership-in-support-of-institutions-that-research-online-racism>

⁶<https://convergenciadigital.com.br/governo/solucao-criada-pela-anatel-e-primeira-inteligencia-artificial-de-facil-distribuicao-no-poder-publico/>

⁷<https://www1.folha.uol.com.br/tec/2026/06/suas-aposta-em-ia-para-cobrar-reembolso-de-plano-s-de-saude.shtml>

⁸<https://minutaia.tjsp.jus.br/login>

necessarily translate into systematic, transparent comparison of competing proprietary systems for specific judicial tasks. Decisions about which AI systems are appropriate for judicial work may therefore be made without publicly available, context-specific evidence comparing their performance, limitations, safety, reliability, and fitness for purpose. In a public institution responsible for decisions with direct consequences for citizens, such choices should not be driven solely by the availability of commercial systems or vendor offerings.

Against this backdrop, the regulatory response has been structurally mismatched. The European Union’s General Data Protection Regulation (GDPR) and the EU AI Act establish baseline protections around data handling and high-risk system classification. Brazil has developed its own regulatory apparatus, including the *Lei Geral de Proteção de Dados* (LGPD), a series of rulings by the Brazilian Supreme Federal Court (STF) on algorithmic decision-making, and guidelines issued by the Coordination for the Improvement of Higher Education Personnel (CAPES) and the National Council for Scientific and Technological Development (CNPq) for the use of AI in federally funded research. Yet none of these frameworks directly addresses the question at the center of this paper: how to determine whether a given frontier model is safe for deployment to a specific population, in a specific language, within a specific institutional context. They regulate data flows and human-in-the-loop needs; they do not evaluate model behavior. The gap between data governance and model safety evaluation remains largely unaddressed.

This gap is widened, rather than closed, by the ethical frameworks that have proliferated at the international level. The UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2022), the OECD Principles on AI⁹, Anthropic’s model specification, and OpenAI’s usage policies all converge on a shared vocabulary: fairness, transparency, accountability, human oversight, non-maleficence. The principles are not wrong, but they are strikingly generic, abstract enough to command universal assent precisely because they defer the hard questions of operationalization. Who determines what counts as harm in a given cultural context? Which population’s values anchor the benchmark? By what institutional

process is a model declared safe for deployment? These questions do not appear in the documents, and their absence is not accidental. The universalist framing of AI ethics functions, in practice, to legitimate deployment decisions made within a narrow technological and cultural elite while projecting the appearance of global consensus.

It is within this conjuncture that the need for a culturally grounded, participatorily constructed safety evaluation framework becomes not merely desirable but urgent. The urgency of this framework is compounded by a more fundamental problem: before asking whether a model is safe for a given population, one must ask what *alignment* means.

4 Alignment

The term *alignment* in AI research lacks a single definition, or, at least, its discussion. Across the literature, alignment has been used to refer to multiple objectives, including ensuring that systems pursue human goals, follow user intentions, satisfy safety requirements, reflect human preferences, or represent societal and cultural values (Shankar et al., 2026). Recent work further treats alignment as representativeness, measuring the degree to which model outputs correspond to the opinions and norms of specific populations (Santurkar et al., 2023). However, these interpretations originate from different research traditions and emphasize distinct aspects of model behavior. Importantly, these objectives may not coincide and can even conflict. A system aligned with user preferences may not satisfy safety constraints, while one aligned with dominant social norms may fail to represent minority perspectives. Moreover, what counts as aligned behavior may vary across cultures, stakeholders, languages, and evaluation methodologies. As a result, *alignment* should not be understood as a singular or self-evident concept, but rather as a multidimensional and contested construct.

Among these dimensions, *moral alignment* occupies a particularly important position in AI safety because it concerns the principles that determine which behaviors are considered acceptable, harmful, or desirable. Since the central argument of this paper concerns who defines those negotiable and principles, we begin by examining the dominant frameworks for moral alignment before turning to the challenges of measuring and operationalizing them.

⁹<https://www.oecd.org/en/topics/ai-principles.html>

4.1 Moral Alignment

The question of what it means for an LLM to be morally aligned has attracted much attention from both industry and academia. Two proposals merit particular examination: the Helpful, Honest, and Harmless (HHH) framework (Askell et al., 2021) and the Flourishing–Justice–Autonomy (FJA) framework (Naseem et al., 2026). The HHH framework, introduced by Anthropic, defines alignment in terms of three criteria that are ‘simple and memorable’ and intended to ‘capture the majority of what we want from an aligned AI’. Helpfulness requires the model to attempt tasks in the user’s best interest; honesty requires calibrated, non-deceptive communication; harmlessness requires the model to avoid offensive, discriminatory, or dangerous outputs. The framework acknowledges that these criteria are ‘at least somewhat subjective’ and that ‘what behaviors are considered harmful and to what degree will vary across people and cultures’, a candid admission that the framework’s own authors recognize its cultural situatedness. Nevertheless, in practice, HHH has operated as a static, post-training set of constraints implemented through RLHF, with annotator pools that do not reflect global linguistic or cultural diversity. The result is a tendency toward overcautious refusals, suppressed reasoning, and diminished cultural sensitivity, the ‘alignment tax’.

The FJA framework proposed by Naseem et al. (2026) represents a more ambitious reconceptualization. Drawing on Nussbaum’s capabilities approach (Nussbaum, 2011) and Rawlsian justice (Rawls, 1971), it replaces the static HHH triad with three dynamic pillars: Flourishing (enabling holistic well-being and long-term capability development), Justice (ensuring equitable procedures that amplify marginalized voices), and Autonomy (safeguarding user agency and informed choice). FJA proposes inference-time alignment with participatory constitutions that users can challenge and amend, rather than fixed post-training constraints. This is a substantive advance since it takes the plurality of human values and the need for dynamic, context-sensitive governance of model behavior seriously.

Yet both frameworks share a structural limitation that their own authors acknowledge but do not fully resolve: neither is culturally grounded in any operationally meaningful sense. HHH treats harm as a universal concept while conceding it is not.

FJA invokes justice and flourishing as organizing principles without specifying whose conception of justice or flourishing anchors the constitution when cultures conflict. The participatory mechanisms FJA proposes, user amendments to constitutions, community deliberation, presuppose institutional infrastructures, digital literacy levels, and political conditions that are unevenly distributed across the globe. A framework designed in Singapore or London, however pluralistic in aspiration, cannot automatically serve a community in Recife or an indigenous population in Amazonas. The principles travel; the populations they were calibrated against do not. This is not a problem that more inclusive language can solve. It requires a different starting point, one in which the values to be aligned are not imported from existing frameworks and then extended to new populations, but are elicited from those populations participatorily and treated as constitutive inputs to the alignment process itself.

4.2 Challenges in Alignment Measurement

If alignment frameworks struggle to specify *whose* values anchor the system, measuring whether a model is aligned inherits the same indeterminacy, now compounded by the practical mechanics of evaluation. Current measurement practice treats alignment as a property to be verified against a fixed, pre-specified set of behaviors: a benchmark presents a prompt, a classifier or human rater judges the response against criteria decided in advance, and a single aggregate score is reported. This pipeline presupposes precisely what the previous section showed to be unresolved: a stable, transferable definition of harm.

The consequences are threefold. First, benchmarks built around one cultural and linguistic context are tacitly assumed to generalize when translated or extended to others, even though what counts as offensive, manipulative, or dangerous is not preserved under translation. A behavior rated as harmful by an Anglophone annotator pool may be miscalibrated for a Brazilian context, and no mechanism in the standard evaluation pipeline surfaces this mismatch. Second, static benchmarks cannot capture frameworks that are explicitly designed to be dynamic and amendable at inference time; a fixed test set measures compliance with a snapshot of a constitution, not the legitimacy or responsiveness of the amendment process itself. Third, and most fundamentally, measurement requires an answer to the question the field has de-

ferred: aligned *according to whom*. Reporting a single harmlessness score implies a single relevant evaluator, collapsing exactly the plurality.

These are not flaws to be patched within the existing measurement paradigm. They follow directly from attempting to measure a culturally and politically contested construct using instruments built as if it were settled. Closing this gap requires an evaluation infrastructure in which the population subject to deployment participates in defining what is measured, not only in interpreting the result.

4.3 Challenges in Alignment Failures Mitigation

Mitigation strategies inherit the same indeterminacy as measurement, displaced from evaluation to intervention. RLHF, constitutional AI, and FJA’s proposed inference-time amendments all act on the model *after* some set of values has already been specified — by annotators, by a constitution’s drafters, by a developer’s policy team. The open question is not how to optimize a model against a given target, but how that target comes to be legitimate for the population it will govern. Mitigation techniques are agnostic to this question: a constitutional clause can be enforced with equal technical fidelity regardless of whether the population it binds had any role in drafting it. Without a resolved account of legitimate elicitation, mitigation risks becoming more efficient at imposing values rather than more just in selecting them.

5 Proposal

We do not attempt to resolve these challenges here; as a position paper, our aim is to indicate the shape a resolution should take. Our central claim is procedural: evaluating AI’s social impact in Brazil requires first determining *what* should be measured: which risks and challenges are materially salient to Brazilian society, before asking *how* to measure it well. Importing benchmarks calibrated to a different material reality answers a question we have not asked, however precisely it answers it. We propose an evaluation framework organized around *national and group sovereignty*, the principle that populations subject to a model’s deployment hold a substantive, institutionally recognized role in defining and verifying its safety. This is not a single deliverable but a conjoint and parallel research agenda along three interdependent fronts.

Culturally-grounded dangerous capability

benchmarks. Determining what to measure means identifying which harms are materially salient to Brazilian society (e.g., the integrity of judicial and welfare decision-making, or vulnerability to manipulation in a media ecosystem with Brazil’s specific patterns of misinformation spread) rather than inheriting a risk taxonomy built around a different country’s material conditions. Only once these categories are established does the question of measurement instruments: datasets, classifiers, annotation protocols, become well-posed; harm and risk categories should be elicited directly from the populations a model will affect, with local institutions, not external annotator pools, responsible for defining what counts as dangerous capability in their own context.

Adaptive evaluation methodology. A fixed benchmark, however well-grounded in Brazilian material conditions, would answer only for the capabilities and harms it was built to test; it degrades as models change and as the risks salient to Brazilian society shift faster than any static dataset can be re-annotated. What is needed instead is an evaluative methodology encompassing automated red-teaming procedures, model-graded or algorithmic risk classifiers, and update protocols, that generalizes beyond the specific cases enumerated in a benchmark and remains responsive to emergent capabilities and conditions as they arise, rather than requiring a wholesale benchmark redesign each time. Determining what to measure, then, is necessary but not sufficient: it must be paired with a mechanism for measuring it that does not become stale the moment a new model or a new social pattern of harm appears. The specific form this mechanism takes: heuristic scoring rules, learned classifiers, or fully algorithmic pipelines etc., is itself an open design question, but its function is fixed: translating revisable risk categories into evaluation outputs without each revision requiring a new static benchmark from scratch.

Infrastructure for safety evaluation. The first two fronts cannot remain dynamic, nor can the determination of what to measure stay responsive to Brazil’s material conditions as they change, without a standing institutional infrastructure. This includes, at minimum: continuously revisable (not one-shot) benchmark governance, capable of updating risk categories as material conditions shift; formal channels through which national and community bodies can contest a deployment decision; and accountability mechanisms linking evaluation

outcomes to actual deployment authority, closing the gap this paper identifies between data governance and the absence of any binding model-safety evaluation requirement. We treat this infrastructure as structurally prior to the other two fronts: without it, the dangerous capability benchmarks and the adaptive evaluation methodology described above both remain advisory rather than binding.

6 Discussion and Challenges

The research agenda proposed here faces challenges that are themselves diagnostic of the problem it addresses, the most immediate is institutional. Participatory benchmark construction requires sustained collaboration between government, AI researchers, legal scholars, civil society organizations, and community representatives.

The second challenge is technical. Culturally-grounded benchmarks are harder to automate than translated ones. Annotation protocols that treat harm as locally defined rather than universally fixed require annotator pools that reflect Brazil’s regional, racial, and linguistic diversity, and that diversity is itself unevenly distributed across the country’s digital and institutional infrastructure.

The third is political. Requiring frontier model owners to meet locally-defined safety standards as a condition of deployment implies regulatory authority that Brazil currently lacks the technical infrastructure to exercise, however strong its legal basis. Evaluation without enforcement is advisory; the accountability mechanisms proposed are therefore not supplementary to the benchmarks and methodology, but structurally prior to their having any binding force.

Finally, there is a risk internal to the agenda itself: that participatory processes, however well-designed, reproduce existing domestic inequalities rather than correcting them. Designing against this risk is not a technical problem, it is political, and it must be treated as such from the outset.

7 Conclusion

This paper argued that measuring AI’s social impact presupposes a consensus on harm, benefit, and risk that does not exist and that this absence is most visible from the periphery of the global AI ecosystem. We showed that this structural flaw operates simultaneously at the technical, theoretical, and governance levels, and that neither existing alignment framework resolves the central question

of who defines the values anchoring the system. The logic we propose is not without precedent in other safety-critical technologies. Automobiles are not certified safe once, globally, and then shipped everywhere unchanged: a car sold in Brazil must meet crash-test, suspension, emissions, and road-worthiness standards set by Brazilian regulators, calibrated to Brazilian roads, traffic patterns, and fuel, because safety is relational to context, and the vehicle is expected to adapt to the road, not the other way around. Frontier models are currently held to no analogous standard: a system validated against Anglophone, Global North benchmarks is deployed into Brazilian courts, hospitals, and welfare systems with no equivalent certification that it is safe for that road. Our proposal is not a universal solution but a procedural starting point: determining what should be measured, with the participation of affected populations, and then measuring it effectively, fairly, and well. The Brazilian research agenda outlined here, culturally-grounded benchmarks, adaptive evaluation methodology, and institutional safety-evaluation infrastructure, may offer a replicable model for other Global South nations facing analogous historical and institutional conditions, particularly across Latin America, and a step toward treating model safety with locally-accountable rigor. Once a population can reliably identify and evaluate the risks specific to its own context, it is also positioned to require that the owners of frontier models meet those locally-defined standards as a condition of deployment, rather than leaving compliance to their discretion.

References

- Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 2024. [Nothing comes without its world – practical challenges of aligning llms to situated human values through rlhf](#). In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 61–73.
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, and 3 others. 2021. [A general language assistant as a laboratory for alignment](#). *Preprint*, arXiv:2112.00861.
- Yibekal Ayana and 1 others. 2025. Seeking human security consensus: A unified value scale for generative AI value safety. *arXiv preprint arXiv:2601.09112*.

- BRICS. 2025. BRICS leaders’ statement on the global governance of AI. Adopted at the BRICS Summit, Rio de Janeiro, Brazil.
- Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martić, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.
- Conselho Nacional de Justiça. 2025. [Pesquisa inteligência artificial no judiciário 2024](#).
- Tiffany Deng, Yumeng Sun, Xinyu Zhu, Nanyang Li, Xinrui Huang, Qingqing Du, Liyuhuan Peng, Kaiping Peng, and Xiaomeng Hu. 2026. [Whose values are artificial intelligence models aligning with? how culture shapes people’s normative expectations of artificial intelligence value](#). *Well-Being Sciences Review*, 2(1):29–36.
- Anjali Kaur. 2025. [From divide to delivery: How AI can serve the global south](#). *Center for Strategic and International Studies*.
- Wm. Matthew Kennedy, Cigdem Patlak, Jayraj Dave, Blake Chambers, Aauysh Dhanotiya, Darshini Ramiah, Reva Schwartz, Jack Hagen, Akash Kundu, Mouni Pendharkar, Liam Baisley, Rumman Chowdhury, and Theodora Skeadas. 2025. [Ask what your country can do for you: Towards a public red teaming model](#). *Preprint*, arXiv:2510.20061.
- Jan Leike, David Krueger, Tom Everitt, Miljan Martić, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.
- Dor Levy and 1 others. 2025. Sycophancy as a multilingual alignment failure: How safety degrades across languages, topics, and models. *arXiv preprint arXiv:2606.08451*.
- Usman Naseem, Tanmoy Chakraborty, Kai-Wei Chang, Mark Dras, Preslav Nakov, Nanyun Peng, and Soujanya Poria. 2026. [LLM alignment should go beyond harmlessness–helpfulness and incorporate human agency](#). *Cognitive Computation*, 18(1):26.
- Martha C. Nussbaum. 2011. *Creating Capabilities: The Human Development Approach*. Harvard University Press.
- Palantir Technologies Inc. 2025. [Palantir reports Q4 2024 revenue growth of 36% Y/Y, U.S. revenue growth of 52% Y/Y; issues FY 2025 revenue guidance of 31% Y/Y growth](#). Press release. Accessed: 2025.
- Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. [Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models](#). In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 143–146, Dubrovnik, Croatia. Association for Computational Linguistics.
- John Rawls. 1971. *A Theory of Justice*. Harvard University Press, Cambridge, MA.
- Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, and Pushpak Bhattacharyya. 2024. Indibias: A benchmark dataset to measure social biases in language models for indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Hari Shankar, Vedanta S P, Sriharini Margapuri, Debjani Mazumder, Ponnurangam Kumaraguru, and Abhijnan Chakraborty. 2026. [Mind the gap: Pitfalls of llm alignment with asian public opinion](#). *arXiv preprint arXiv:2603.06264*.
- Gabriel Simmons. 2023. [Moral mimicry: Large language models produce moral rationalizations tailored to political identity](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297, Toronto, Canada. Association for Computational Linguistics.
- UNESCO. 2022. Recomendação sobre a Ética da inteligência artificial. Programme and Meeting Document SHS/BIO/PI/2021/1, UNESCO, Paris, France. Adopted on 23 November 2021. Portuguese edition. Licensed under CC BY-NC-SA.
- Yihong Wang and 1 others. 2024. All languages are not created (tokenized) equal. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. XSafety benchmark; documents safety drift across languages.
- Pak-Hang Wong. 2020. [Cultural differences as excuses? human rights and cultural values in global ethics and governance of ai](#). *Philosophy & Technology*, 33(4):705–715.
- Shoshana Zuboff. 2015. [Big other: Surveillance capitalism and the prospects of an information civilization](#). *Journal of Information Technology*, 30:75–89.