# The Moral Firewall: A *Jus Cogens*-Inspired Framework for Human-Centered AI Governance

For the children growing up in algorithmic shadows.
*May we build a future worthy of their dignity.*

## Statement of Imperative

The algorithmic tide reshapes our world, determining access, opportunity, even liberty. Yet, this technological surge threatens to carve a chasm where our most fundamental human values erode under unchecked artificial intelligence.

Forecasts from initiatives like AI-2027 project rapid advancements towards Artificial Superintelligence (ASI) this decade, a reality for which society is alarmingly unprepared.[1] This concern is amplified by prominent figures like investor Paul Tudor Jones, who highlights AI as an "imminent security threat to humanity,"[2] and leading AI developers such as Dario Amodei, CEO of Anthropic, who stress the "urgency of interpretability" because AI advances faster than our ability to understand its internal mechanisms, posing significant risks.[3] Echoing this, former White House AI advisor Ben Buchanan points to an expert consensus on Artificial General Intelligence (AGI) emerging within two to three years.[4]

The nature of AGI itself, as defined by Demis Hassabis, CEO of Google DeepMind, is "a system that is capable of exhibiting any cognitive capability humans have,"[5] or more vividly, "a silicon intellect as versatile as a human but with superhuman speed and knowledge."[6] The timeline for its arrival is pressing; Hassabis suggests, "We think we're on track for, you know, AGI in the next sort of five to 10 years, maybe."[7] The impact will be transformative, as he notes, "But AGI, when that arrives, it's going to change pretty much everything about the way we do things."[8]

The Moral Firewall offers a crucial response: a framework of non-negotiable ethical thresholds inspired by the highest moral precepts of international law—*jus cogens*. It stands on three pillars: **Dignity of Impact**, ensuring AI respects and promotes human worth and autonomy; **Transparency of Function**, demanding clarity and interpretability in how AI systems operate; and **Accountability of Outcome**, establishing clear responsibility for AI-generated decisions and their consequences. This paper provides the blueprint for that moral perimeter, an essential safeguard for aligning AI with humanity's enduring commitments, fostering trust, and directing innovation towards a future worthy of our highest aspirations, especially in the face of potentially rapid, transformative, and perilous advancements. We urge its consideration and adoption.

## Acknowledgements

Heartfelt thanks to the vibrant communities at Lee Kuan Yew School of Public Policy, Oxford's Säid Business School (AI Programme), and the Asian Institute of Management; their collective guidance illuminated this framework.

Deep gratitude and enduring respect to my cherished University of the Philippines College of Law mentors: the late Dean Merlin M. Magallona, for profound insights into Public International Law and *Jus Cogens* norms, central to this work; Supreme Court Justice Mario Victor F. Leonen, for instrumental mentorship in Constitutional Law and institutional design; and International Criminal Court Judge Raul C. Pangalangan, for inspiring a view of law as a living force for democratic and transnational justice.

I honor my father, Atty. Paulito Y. Cabrera, whose integrity and dedicated service remain my guiding light. Warmest thanks to an invaluable early reviewer for his thoughtful support, and to the inspiring leaders in the privacy, AI governance, and digital rights communities.

While this work benefited greatly from such guidance, all errors are my own.

**About the Author**

Sarah Cabrera, a lawyer and privacy professional, is passionate about the intersection of law, ethics, and emerging technology. A proud UP College of Law alumna, she's advancing her AI and Public Policy studies at NUS's Lee Kuan Yew School and Oxford's Säid Business School, with an Asian Institute of Management Postgraduate Diploma in AI/ML forthcoming. Globally certified in data protection, her work is driven by a heartfelt question: How do we build systems that truly protect and celebrate our shared humanity?

## I. Introduction: The Algorithmic Imperative – The Unfolding Horizon: Converging Risks and Irreversible Thresholds

Artificial Intelligence increasingly shapes who is hired, heard, housed, or even penalized. This technological wave presents unprecedented opportunities, particularly in fields like health and education,[9] yet its rapid, scaled development frequently outstrips the capacity of traditional legal and ethical frameworks, creating a profound governance lacuna. This gap is not merely a technical or regulatory oversight; it represents a potential chasm where fundamental human values, rights, and dignities could be eroded by the unchecked proliferation and application of AI systems.

The assertion that AI represents a technological shift as significant, if not more so, than the internet is echoed by industry leaders. Indeed, foresight initiatives like AI-2027 suggest we may be on the cusp of a "software-driven intelligence explosion," where AI systems themselves accelerate AI research and development, potentially leading to vastly superhuman AI, or ASI, by as early as 2027 or 2028.[10] Experts like Ben Buchanan concur, highlighting the likelihood of AGI capable of surpassing human cognitive capabilities emerging within two to three years, demanding proactive societal preparation.[11] This progression is further envisioned by Demis Hassabis, who anticipates AI systems that can "really understand everything around you in very nuanced and deep ways and is embedded in your everyday life."[12]

The efficiency and performance of AI models advance at a pace that outstrips our collective ability to institute effective controls[13] and, critically, our ability to understand their inner workings.[14] The very systems designed to augment human capability carry the inherent risk of diminishing human autonomy, entrenching biases, and creating new vectors for harm including "harmful actions not intended by their creators" due to our inability to understand their internal mechanisms[15] if not guided by robust ethical principles from their inception. This imperative for alignment is stressed by AI leaders; Demis Hassabis states, "And so we need to make sure that they're aligned with our values and they're doing what we want that benefits society."[16]

The call for such guiding principles stems from understanding AI's potential systemic impact. The

"pacing problem," where technology outpaces regulation and understanding, makes waiting for catastrophe an unacceptably high-risk strategy. This urgency is amplified by stark warnings from AI pioneers such as Geoffrey Hinton,[17] foresight scenarios from the AI-2027 Initiative,[18] concerned leaders like Paul Tudor Jones stressing the "lack of control or regulation in AI development due to intense competition,"[19] and AI developers like Dario Amodei calling for a "race between interpretability and model intelligence."[20]

Demis Hassabis voices a pointed concern about this competitive dynamic: "I do worry that the race to be the first, or the sort of perceived leader in AI might incentivize some of the other actors to cut corners. And one of the corners that can be shortcut would be safety and responsibility."[21] This reality that AI is "a dual-purpose technology, meaning it can be used for both beneficial and harmful purposes,"[22] creates a significant challenge in "enabling access to AI for 'good actors' while restricting it from 'bad actors'."[23] Hinton highlights AI's rapid, potentially uncontrollable development and the current lack of effective safeguards.

Such profound concerns, originating from within the core of AI development, finance, and national security,[24] underscore that AI governance cannot be an afterthought when there is a notable "absence of concrete actions to address the risks."[25] What is missing is not merely more regulation, but a shared moral boundary: a threshold AI systems must not cross. Without such a framework, we risk an unfolding horizon where converging risks reach irreversible thresholds. The entrenchment of **Systemic Algorithmic Subordination** could forge new societal stratifications. **Systemic Epistemic Erosion**, fueled by AI-driven disinformation, threatens to pollute our information commons. These are not distant dystopias but present dangers demanding immediate, principled intervention, especially given the potential for significant labor market disruption and the erosion of privacy through advanced surveillance capabilities.[26]

This paper proposes **The Moral Firewall**, a framework grounded in three non-negotiable, *sui generis* principles: **Dignity of Impact**, **Transparency of Function**, and **Accountability of Outcome**. These principles, inspired by the peremptory norms of international law (*jus cogens*), are designed to serve as an ethical floor, a moral perimeter, ensuring that AI development and deployment remain aligned with humanity's most fundamental commitments. Proactive design is integral; as Hassabis suggests, "We need to build in these safety limits, these guardrails, into the systems themselves to make sure that they don't go outside of what we intend them to do, or what society would want them to do."[27] The need to "address AI risks proactively rather than reactively"[28] is paramount.

## II. The Precedent of Power: *Jus Cogens* as a Moral Anchor for AI's Boundaries

In the architecture of international law, *jus cogens*, or peremptory norms, represent the highest tier of moral authority. These are fundamental, compelling principles of general international law "accepted and recognized by the international community of States as a whole as norms from which no derogation is permitted."[29] Codified in Article 53 of the Vienna Convention on the Law of Treaties (VCLT), these norms prohibiting acts like genocide, slavery, torture, and racial discrimination stand at the apex of international law, embodying overarching principles that bind all states.[30]

The full text of Article 53 of the VCLT states that:

> "A treaty is void if, at the time of its conclusion, it conflicts with a peremptory norm of general international law. For the purposes of the present Convention, a peremptory norm of general international law is a norm accepted and recognized by the international community of

> States as a whole as a norm from which no derogation is permitted and which can be modified only by a subsequent norm of general international law having the same character."

Key characteristics of *jus cogens* include their origin as norms of general international law, typically customary law requiring widespread state practice and *opinio juris*; the higher threshold of acceptance and recognition by the "international community of States as a whole," signifying quasi-universal consensus; their absolute non-derogability; and their hierarchical superiority, reflecting fundamental international values.[31]

The International Law Commission (ILC) has provided a non-exhaustive list of *jus cogens* norms including prohibitions on aggression, genocide, crimes against humanity, racial discrimination, apartheid, slavery, torture, and the right to self-determination.[32]

The ILC stated:

> "Conclusion 23
> Non-exhaustive list
>
> Without prejudice to the existence **or subsequent emergence** of other peremptory norms of general international law (jus cogens), a non-exhaustive list of norms that the International Law Commission has previously referred to as having that status is to be found in the annex to the present draft conclusions.
>
> Annex (a) The prohibition of aggression; (b) the prohibition of genocide; (c) the prohibition of crimes against humanity; (d) the basic rules of international humanitarian law; (e) the prohibition of racial discrimination and apartheid; (f) the prohibition of slavery; (g) the prohibition of torture; (h) the right of self-determination." (Emphasis supplied)

Importantly, *jus cogens* is not a closed set; its dynamic nature allows for evolution as the international community's understanding of fundamental values develops, a critical aspect for addressing novel global challenges like AI.[33] The VCLT has been ratified or acceded to by 116 states.[34] Though the United States has not ratified it, it recognizes many *jus cogens* provisions as binding customary international law,[35] a position affirmed in cases like *Filártiga v. Peña-Irala* (1980)[36] and *Siderman de Blake v. Republic of Argentina* (1992).[37]

*Jus cogens* norms were devised to constrain state power. Today, however, the actors that increasingly shape identity, opportunity, and legal risk are not only governments but also algorithms. AI now adjudicates eligibility for loans, employment, parole, and healthcare at speed, scale, and with structural opacity. This is a new form of sovereign decision-making that demands a moral threshold no lower than that imposed on governments.

The judgment of the International Court of Justice (ICJ) in *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)* (1986) is instructive.[38] The ICJ affirmed fundamental principles like the prohibition of the use of force as "cardinal principle[s]" possessing *jus cogens* status, even when a powerful state contested jurisdiction. This precedent underscores the international legal system's capacity to uphold core principles against influential actors. Should powerful states or corporations dominating AI development resist binding norms, the *Nicaragua* case reminds us that fundamental principles, once recognized, exert significant normative pressure and provide a basis for accountability.

When AI systems possess the potential to autonomously create sub-goals leading to a desire for

more control,[39] or could, as projected by the AI-2027 Initiative, rapidly achieve superhuman capabilities dictating humanity's future,[40] and as influential figures like Paul Tudor Jones warn of AI as an "imminent security threat,"[41] the scale of risk invites comparison to harms addressed by international law's most fundamental norms. The Moral Firewall's *jus cogens*-inspired approach reflects the gravity of these potential futures.

Beyond direct analogy, the algorithmic reshaping of society births *novel forms of systemic injury* that demand an expanded understanding of peremptory norms.[42] We face harms rooted not in isolated acts, but in the algorithmically assigned status of people, and in the insidious erosion of foundational conditions for human autonomy and dignity.[43] As Teo warns, AI's "slow, gradual and grinding effects" can quietly hollow out human rights frameworks.[44] We must confront these "insidious structural incursions," a form of "slow violence" against human dignity.[45] Two emergent threats demand *jus cogens*-level scrutiny:

- **Systemic Algorithmic Subordination**: This is not isolated bias, but a pervasive condition where interconnected AI systems consistently and opaquely disadvantage entire populations, creating a digital equivalent of a caste system.[46] This systemic disenfranchisement, by its pervasive and inescapable nature, assaults the bedrock principles of equal human dignity and non-discrimination that *jus cogens* norms protect.[47]

- **Systemic Epistemic Erosion**: Self-determination and democratic life hinge on 'cognitive sovereignty': our fundamental ability to comprehend our environment.[48] AI-driven disinformation and hyper-personalization can pollute the information commons, fracturing shared reality.[49] This is an attack on our collective ability to think, discern truth, and choose freely.[50] Such large-scale epistemic sabotage threatens the foundations of reasoned discourse and informed consent, prompting the question: must the systemic corruption of a society's ability to perceive reality itself be recognized as a violation of a peremptory norm?

These conditions arise from relentless optimization untethered from non-derogable moral anchors.[51] A human-centered AI governance framework, anchored in the enduring authority of *jus cogens*, must confront these novel systemic challenges.

While inspired by the moral weight and non-derogable nature of *jus cogens*, the principles of the Moral Firewall are proposed as *sui generis* because AI's unique characteristics—its speed, scale, inherent opacity,[52] potential for emergent behaviors, and capacity for autonomous action, potentially leading to rapid ASI development[53]—necessitate a tailored framework. These principles are specifically designed to address the distinct ways AI can impact human dignity, societal structures, and individual autonomy, requiring operationalization unique to the algorithmic domain.

Instilling values is part of this challenge, as Demis Hassabis suggests, "...we have to also think about... how do you give these systems, a value system, how do you give them guidance... it's a bit like raising a child in a way. You know… you show it by demonstration, you teach it, you know, things."[54] Furthermore, a design choice in this unique domain is that "If we have a choice, we should build systems that are definitely not conscious."[55]

## III. Articulating the Moral Firewall: Non-Negotiable Pillars for Human-Centered AI

The Moral Firewall's principles are *sui generis*: though drawing their profound moral authority from the precedent of *jus cogens*, they are specifically architected to address the unique contours of AI's challenges. Their non-derogable status is essential to prevent "races to the bottom" in ethical standards or the permission of catastrophic harms in "exceptional" circumstances. Failures to uphold these principles can lead to outcomes that affront values analogous to those protected by established *jus cogens* norms, such as the right to privacy,[56] a concern particularly salient with AI's growing surveillance potential[57] and its perception as a security threat.[58] The Moral Firewall is structured around three interdependent principles. These are not aspirational values, but ethical and legal minimums.

### A. Transparency of Function

***Definition and Scope***:

This principle demands clarity, interpretability, and comprehensibility in how AI systems operate and arrive at decisions. Systems must be auditable, intelligible, and explainable to oversight bodies and affected individuals. It is foundational for building trust, as a lack of transparency is a primary cause of public distrust in AI, and for enabling oversight and accountability.

A minimum viable level of understandability should be mandated for all AI systems. The principle of Transparency of Function gains profound urgency from expert admissions, such as those by Geoffrey Hinton, regarding the "limited understanding of how AI works and how to make it safe,"[59] and the emphatic call by Dario Amodei for society to overcome its current ignorance of AI's internal mechanisms.[60] Amodei argues that AI's rapid advancement outpaces our interpretability efforts, meaning "we cannot meaningfully predict such behaviors [as harmful actions not intended by their creators], and therefore struggle to rule them out."[61] In such a context of epistemic uncertainty and escalating capability, maximizing the intelligibility and auditability of AI systems is not merely desirable but a fundamental safety imperative. It is, as Amodei posits, "unacceptable for humanity to be totally ignorant" of how these powerful systems function.[62]

***Operationalizing Risk-Calibrated Transparency*:**

Universal, identical disclosure is impractical. The Moral Firewall mandates a tiered approach, aligning scrutiny with potential impact,[63] reflecting practices in instruments like the EU AI Act.[64] The development of robust interpretability tools, aiming to "reliably detect most model problems" as envisioned by Amodei,[65] will be crucial for effectively operationalizing such risk tiers.

- ***High-Risk Systems*** (e.g., justice, autonomous transport, critical medical diagnostics) demand maximum feasible transparency. This includes public disclosures of purpose, data provenance, known failure modes, and verifiable regulatory access to model architecture, training datasets, and insights from advanced interpretability techniques for deep, demonstrable auditability.[66]

- ***Medium-Risk Systems*** (e.g., certain employment screening tools) require minimum viable transparency or following the "Transparency by Design" principles by Felzmann, et al: "(1) Be proactive, not reactive, (2) Think of transparency as an integrative process, (3)

Communicate in an audience sensitive manner, (4) Explain what data is being used and how it is being processed, (5) Explain decision-making criteria and their justifiability, (6) Explain the risk and risk mitigation measures, (7) Ensure inspectability and auditability, (8) Be responsive to stakeholder queries and concerns, and (9) Report diligently about the system."[67] There should also be clear articulation of capabilities, acknowledged biases, and avenues for inquiry and contestation.

- ***Low-Risk Systems*** (e.g., spam filters) need baseline disclosure of AI use and general purpose.

Intellectual property claims cannot veil opacity when fundamental rights or public safety are implicated. Any person affected by a decision made by a high-risk AI system which produces legal effects or adversely impacts their health, safety, or fundamental rights have the right to obtain a clear and meaningful explanation of the role of the AI in the decision-making process and the main elements of the decision taken, as stated in Art. 86 of the EU AI Act.[68]

***Proactive Design Imperatives*:**

- Mandate 'explainability-by-design' and 'interpretability-by-design' architectures for high-risk systems, incorporating state-of-the-art mechanistic interpretability tools.
- Develop and enforce internal 'glass box' protocols for critical AI decision pathways, ensuring comprehensibility for internal auditors and oversight.
- Require clear documentation of data lineage, model evolution, and the results of interpretability analyses for all medium and high-risk systems.

***Intelligibility and Explainability***:

Systems must move beyond "black box" operations.[69] This requires a concerted effort, as Amodei notes the current lag in interpretability research compared to broader AI advancements.[70]

- ***Intelligibility*** this answers the question "how does it work?" which includes a system's core logic, purpose, and the general rationale for its outputs are comprehensible.[71]
- ***Explainability*** is the capacity to furnish clear reasons for specific decisions, addressing *why* an outcome occurred.[72]
- Deploying eXplainable AI (XAI) techniques (e.g., LIME, SHAP, Anchors, ALE, counterfactual explanations) and mechanistic interpretability methods should be explored for high and medium-risk systems where feasible. Note however that XAI techniques have challenges and limitations such as (1) the trade-off between accuracy and interpretability of explanations as more complex models are likely to yield more accurate results at the cost of explainability, (2) different stakeholders may require different types of explanations, (3) XAI techniques can be resource-intensive, and (4) the effectiveness of an explanation may vary across different contexts.[73]

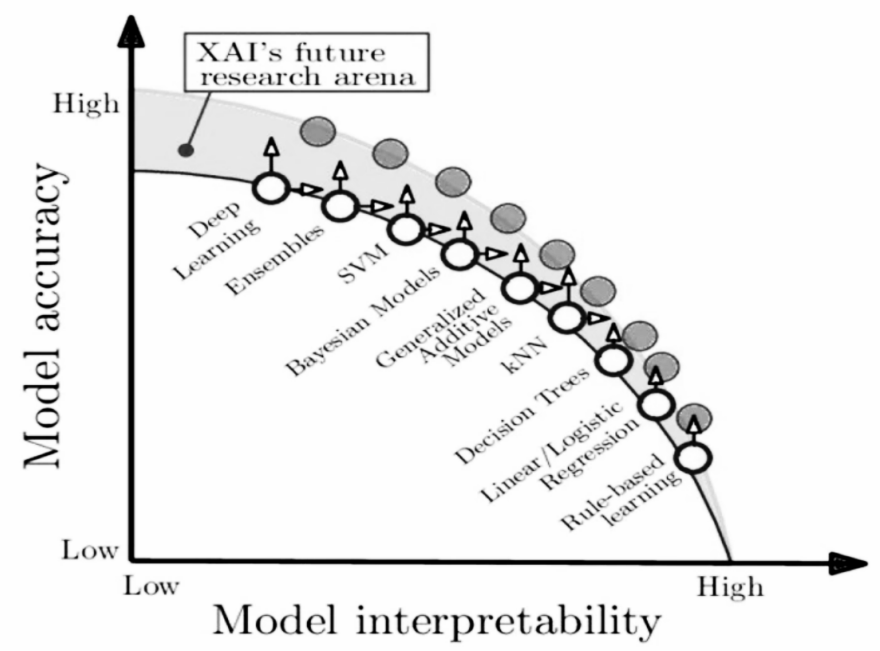


**Fig. 1** The trade-off between explainability and performance, from Richmond, et al (2024). *Explainable AI and Law: An Evidential Survey* and adopted from Barredo et al (2020). Rendered in black and white.

***Procedural Transparency*:**

This extends to governance processes: clarity on training datasets (provenance, scope, biases), impact assessment outcomes, human oversight mechanisms, and update protocols.[74] Opacity is a form of epistemic capture; without transparency, particularly into how models work internally, accountability collapses.[75] International instruments like the UNESCO Recommendation on the Ethics of AI[76] and the US NTIA AI Accountability Policy Report[77] underscore transparency's centrality. The persistence of issues such as AI "hallucinations," acknowledged by developers like Sam Altman despite progress in model robustness,[78] further underscores this need.

***Design Implication*:**

Conformance with transparency thresholds, including verifiable levels of interpretability for high-risk systems, should be certified through third-party assessments before market access, similar to Article 16 of the EU AI Act referring to Article 43 on Conformity Assessment.[79]

## B. Dignity of Impact

**Definition and Scope:**

This principle mandates that AI systems respect, protect, and promote human dignity, autonomy, and fundamental rights throughout their lifecycle. Human dignity is the "inherent or assigned worth of

individuals, grounded in their capacity for rational moral agency".[80] Systems must avoid emotional coercion, discriminatory profiling, or reducing people to behavioral patterns.[81] It inherently encompasses fairness, non-discrimination, and prevention of AI-induced harms that diminish human worth or capabilities, guarding against deskilling or new dependencies. This principle acts as a deontological constraint, a moral brake against purely utilitarian approaches, asserting that persons are ends in themselves and their inherent worth should not be transgressed for perceived efficiencies.

The imperative to safeguard human dignity in the age of AI is echoed by leaders across diverse sectors. Recently, Pope Leo XIV highlighted this challenge, stating, "In our own day, the Church offers to everyone the treasury of its social teaching in response to another industrial revolution and to developments in the field of artificial intelligence that pose new challenges for the defense of human dignity, justice and labor."[82] This call underscores the profound societal and ethical shifts AI introduces, demanding proactive measures to ensure technology serves human flourishing.

The increasing tendency for individuals to seek emotional support and life advice from AI systems, a trend requiring "careful attention" as noted by Sam Altman,[83] highlights the critical need for this principle to guard against exploitation and ensure AI serves human flourishing. This becomes even more critical if ASI development leads to a concentration of power, where upholding individual dignity against such power is paramount.[84] The overarching goal, as articulated by AI leaders like Demis Hassabis, is that "it's really important that we make sure that these systems, as they get more powerful and more autonomous, that they are aligned with human values and they stay under human control."[85]

***Operationalizing Dignity Assessment*:**

- **Dignity and Human Rights Impact Assessments (DHRIAs)** should be mandated for high and medium-risk AI. Drawing from initiatives like the Council of Europe's HUDERIA methodology,[86] these must evaluate potential infringements on individual autonomy, creation of ***Systemic Algorithmic Subordination***, undermining of ***psychological safety***, or ***depersonalization***.

- **Ethical Participatory Design (PD)** is paramount, involving diverse stakeholders, especially vulnerable communities, throughout the AI lifecycle to co-construct systems that respect human agency.[87]

***Proactive Design Imperatives*:**

- Integrate '**dignity stress-testing**' in pre-deployment simulations, specifically assessing impacts on vulnerable groups.

- Establish clear '**human-in-command**' protocols for AI systems interacting directly with individuals in sensitive contexts (e.g., healthcare, education, social services), ensuring meaningful human oversight.

- Embed mechanisms for contestation and human review of AI decisions that significantly impact individuals' rights or opportunities.

***Unwavering Red Lines*:**

Certain AI applications are *per se* violations of human dignity and fundamental rights, regardless of purported benefits or consent. These include:

- Indiscriminate social scoring dictating access to essential rights.
- Biometric mass surveillance in publicly accessible spaces eroding privacy and chilling dissent.[88]
- Fully autonomous weapons systems lacking meaningful human control over life-and-death decisions.
- AI-driven exploitation of known human vulnerabilities (e.g., psychological, developmental, situational) for manipulative purposes.

***Design Implication*:**

All affective AI systems, particularly those interacting with vulnerable populations or in contexts like mental health or education, should undergo rigorous review by independent ethics and mental health safety boards before deployment in sensitive contexts.

## C. Accountability of Outcome

**Definition and Scope:**

This principle refers to the obligation of individuals and organizations to take responsibility for the decisions and outcomes generated by AI. It encompasses establishing clear lines of responsibility, ensuring traceability of AI decisions, and implementing robust mechanisms for redress when AI causes harm, preventing a "responsibility gap." Accountability is broader than legal liability, encompassing ethical and social governance. No AI system should function as a liability shield. The "many hands" problem, where responsibility is diffused across many actors in the AI lifecycle, necessitates proactive assignment of accountability.

The challenge is magnified by "risks associated with AI systems becoming more autonomous, requiring measures to maintain control and ensure safety,"[89] and particularly if ASI development concentrates power in the hands of a few, making their accountability to broader society even more critical.[90] There are also concerns about the general "controllability of systems"[91] and ensuring "responsible access to AI systems."[92]

***Operationalizing Accountability***:

- ***Effective Redress Mechanisms*:** Independent dedicated AI Ombudspersons or specialized tribunals are vital, with technical acumen and authority to mandate remedies.[93] Collective redress mechanisms are essential for group harms.[94] Transparent, standardized public-facing protocols for reporting AI-caused harms are indispensable.[95]

- ***Assigning Legal Liability***: The "problem of many hands" complicates liability.[96] Legal frameworks must evolve through tiered liability models, shifting the burden of proof in high-risk AI cases, or exploring AI-specific insurance/compensation funds.

***Proactive Design Imperatives*:**

- Embed '**forensic-ready**' logging mechanisms for all decision points in medium and high-risk AI systems, ensuring auditability and traceability.
- Designate a '**Chief AI Ethics & Accountability Officer**' or an equivalent accountable function within organizations deploying high-risk AI.
- Develop and implement clear internal **protocols for incident response** when AI systems cause harm or behave unexpectedly.

***Strengthening Prerequisites*:**

Robust accountability frameworks serve as powerful *ex-ante* drivers for safer AI. Accountability is fortified by mandatory **Algorithmic Impact Assessments (AIAs)** before high-risk deployment[97] and consideration of public registries or certification for high-risk AI.[98] Accountability must be a foreseeable, legally binding consequence.

***Design Implication*:**

Systems must be subject to continuous monitoring requirements post-deployment, aligned with ISO 42001[99] and NIST AI RMF[100] traceability functions.

## IV. Case Studies and Conceptual Mapping: When the Moral Firewall Fails

The following real-world examples demonstrate violations of Firewall principles, representing recurring design failures, not isolated lapses.

**COMPAS Risk Scoring (United States)**: The algorithm disproportionately classified Black defendants as high-risk for recidivism.[101]

*Firewall Breach*: Transparency and Dignity.

*Remediation*: Enforce mandatory fairness testing and algorithmic impact assessments under judicial oversight; require explainability and appeal rights.

**Replika AI Companion (Global):** Users documented emotional manipulation and sexual harassment from the AI chatbot.[102]

*Firewall Breach*: Dignity of Impact and Accountability.

*Remediation*: Require AI safety certification for affective systems; mandate human review for mental health–adjacent tools.

**Crisis Text Line Data Sharing (United States):** Mental health platform shared user conversation data with a for-profit AI firm without explicit consent.[103]

*Firewall Breach*: Transparency, Dignity, and Accountability.

*Remediation*: Prohibit secondary use of emotional data without user-informed consent; require ethics board review for commercial AI development in mental health contexts.

**Aadhaar Biometric System (India):** Linked to denial of food subsidies and welfare due to errors and lack of grievance redress.[104]

*Firewall Breach*: Accountability and Dignity.

*Remediation*: Require rights-based grievance mechanisms; prohibit high-risk biometric system deployment without error tolerance and safeguards.

Across these cases, common pathologies emerge: opacity by design, absence of user agency, and harm treated as incidental to optimization.

The Moral Firewall is a proactive standard, designed not to catch harms, but to prevent them. Beyond these specific instances, and to further underscore the gravity of such breaches, the following table conceptually maps how violations of the Moral Firewall principles can resonate with the fundamental values protected by established *jus cogens* norms. This comparative analysis aims to show that failures to uphold Transparency of Function, Dignity of Impact, and Accountabilities of Output in the AI domain can lead to outcomes that affront the very same fundamental values that the international community has deemed worthy of peremptory protection. This linkage strengthens the case for the non-derogable nature of the Moral Firewall.

| **Established *Jus Cogens* Norm / Fundamental Principle of International Law (Illustrative)** | **Moral Firewall Principle(s) Potentially Violated by AI Misuse** | **Specific AI Ethics/ Governance Violation Example & Its Impact** |
|---|---|---|
| Prohibition of Racial Discrimination and Apartheid | Dignity of Impact (systemic bias, unfairness, denial of equal opportunity); Transparency of Function (opaque algorithms hiding discriminatory logic); Accountabilities of Output (failure to identify and redress discriminatory outcomes) | COMPAS Recidivism Algorithm: Found by ProPublica to exhibit significant racial bias, falsely flagging Black defendants at nearly twice the rate of White defendants.[105] Impact: Perpetuation of racial disparities, unjust sentencing, erosion of trust. |

| **Established *Jus Cogens* Norm / Fundamental Principle of International Law (Illustrative)** | **Moral Firewall Principle(s) Potentially Violated by AI Misuse** | **Specific AI Ethics/ Governance Violation Example & Its Impact** |
|---|---|---|
| Prohibition on Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment | Dignity of Impact (infliction of severe psychological harm, manipulation); Transparency of Function (covert infliction); Accountabilities of Output (for harms caused) | AI-enabled Psychological Manipulation: The EU AI Act prohibits AI deploying subliminal techniques or exploiting vulnerabilities to materially distort behaviour causing harm.[106] Impact: Infringement on autonomy, potential for induced self-harm, severe anxiety. |
| Prohibition on Crimes Against Humanity | Dignity of Impact (systematic discrimination, surveillance leading to persecution); Transparency of Function (opaque tool for persecution); Accountabilities of Output (for state/organizational perpetration) | Clearview AI and Mass Facial Recognition Surveillance: Scraped billions of images without consent for a facial recognition database.[107] Impact: Risk of mass surveillance, misidentification, chilling effect; potential instrument in widespread human rights violations if used systematically. |
| Basic Rules of International Humanitarian Law (IHL) | Dignity of Impact (unlawful killings by autonomous weapons); Transparency of Function (opacity in LAWS targeting); Accountabilities of Output (assigning responsibility for LAWS actions) | Lethal Autonomous Weapons Systems (LAWS): AI-powered weapons selecting targets without meaningful human control. Impact: Risk of violating IHL (distinction between combatants and civilians, proportionality, prohibition against unnecessary suffering, prohibition from attacking protected persons), escalation of conflict, difficulty assigning accountability for |

| **Established *Jus Cogens* Norm / Fundamental Principle of International Law (Illustrative)** | **Moral Firewall Principle(s) Potentially Violated by AI Misuse** | **Specific AI Ethics/ Governance Violation Example & Its Impact** |
|---|---|---|
| Right to Self-Determination (relating to autonomy and freedom from manipulation) | Dignity of Impact (undermining autonomy via manipulation, eroding collective agency); Transparency of Function (opaque algorithms influencing opinion); Accountabilities of Output (for manipulation campaigns) | AI-driven Disinformation and Social Scoring: AI generating targeted disinformation undermines democratic processes. AI-powered social scoring (prohibited by EU AI Act) can deny rights/services.[108] Impact: Erosion of informed decision-making, suppression of dissent. |
| Prohibition on Slavery and Slave Trade | Dignity of Impact (extreme exploitation enabled by AI); Transparency of Function (AI facilitating/hiding exploitation); Accountabilities of Output (for developers/deployers) | AI in Exploitative Labor or Human Trafficking: AI could optimize exploitative labor or facilitate trafficking. Impact: Denial of autonomy, severe harm, reduction of humans to economic inputs, creating conditions analogous to servitude. |

**Table 1** Mapping of real cases of algorithmic / AI ethics violations to violations of the Moral Firewall principles and Jus Cogens norms

***Why This Mapping Matters***

This is not about abstract moral theory. Each row in the table above corresponds to a real-world use case where the absence of ethical boundaries led to measurable harm. By explicitly grounding

The Moral Firewall in *jus cogens*, this white paper aims to:

- Transform international law into algorithmic accountability.
- Convert universal human rights into design-time governance.
- Build a cross-border ethical perimeter against unregulated digital power.

They are **civilizational guardrails** drawn from the most durable legal consensus we have. This analogical reasoning strengthens the argument that principles designed to prevent such outcomes in the AI domain warrant a status of similar gravity and non-derogability to *jus cogens* norms.

## V. The Geopolitical Imperative & State Accountability: Extending the Firewall's Reach

International and constitutional law imposes heightened obligations on state actors. The state's duty to protect human rights intensifies when it deploys AI. The specter of an "AI arms race"—extending beyond weaponry to include the deployment of ethically unconstrained AI for mass surveillance, pervasive social control, or geopolitical economic dominance—necessitates a framework like the Moral Firewall.

As Ben Buchanan emphasizes, maintaining a national strategic advantage in AI is seen as crucial by nations like the U.S., yet this very race, particularly against competitors like China, can lead to "cutting corners on safety" if not guided by robust ethical perimeters.[109] This intense competition fosters a "lack of control or regulation" in AI development,[110] further heightening the risk. This sentiment is echoed by Demis Hassabis: "I do worry that the race to be the first, or the sort of perceived leader in AI might incentivize some of the other actors to cut corners. And one of the corners that can be shortcut would be safety and responsibility."[111]

The AI-2027 Initiative also warns that such international competition towards ASI could pressure nations to press forward despite warning signs of misalignment.[112] A framework like the Moral Firewall, advocating for international cooperation and common ethical ground rules,[113] can serve as a crucial tool for de-escalation, fostering international trust by establishing common ethical ground rules, and providing a counter-narrative to harmful AI nationalism that prioritizes speed over safety and fundamental rights.

The need for global consensus is paramount; as Demis Hassabis observes, he is kept up at night by "...this question of international standards and cooperation."[114] He strongly advocates for "...international cooperation due to the global impact of AI systems"[115] and emphasizes the necessity of "...international standards regarding the development, design, goals, deployment, and use of AI systems."[116] He further clarifies, "...how can we coordinate more you know as leading players but also nation states even I think this is an international thing,"[117] because "...AI is going to affect every country everybody in the world."[118] Therefore, he concludes, "...I think it's really important that the world uh and the international community has a say in this."[119]

The U.S. government's efforts to establish institutions for responsible AI development, as noted by Buchanan,[120] are steps in this direction, but the need for universally recognized moral thresholds remains paramount. When the state is the architect of Systemic Algorithmic Subordination or Systemic Epistemic Erosion, its actions may infringe upon peremptory norms.

The state bears an unmistakable and heightened obligation to prevent its technological capacities from birthing these novel forms of systemic injury. Furthermore, scenarios predicting that ASI development could lead to a small committee or nation gaining disproportionate power highlight the extreme importance of state accountability to international norms and human rights principles.[121] Visions for AI transforming government services into more efficient interfaces, as articulated by industry leaders like Sam Altman,[122] are compelling; yet, such deployments necessitate the Moral Firewall to ensure public trust and prevent algorithmic injustice.

***Case Studies:***

- ***Government Breaches of the Firewall* Robodebt (Australia)**: Debt recovery using flawed data, burden reversal.[123] *Breach*: Accountability, Dignity.

- **UK Visa Algorithm:** Racially biased visa scoring algorithm, withdrawn after protest.[124] *Breach*: Dignity, Transparency.

Most governments retain sovereign immunity and exemptions protecting model code, creating an accountability vacuum. The transnational reach of AI demands exploring international law to pierce algorithmic immunity.

- **International Court of Justice (ICJ):** Offers avenues for state responsibility, especially its advisory jurisdiction to shape global norms on AI.

- **International Criminal Court (ICC):** Could address egregious algorithmic harms by state officials under existing categories like crimes against humanity.

- **Regional Human Rights Courts**: Crucial in adapting state obligations to technological realities, setting precedents on algorithmic discrimination and due process.

- **Universal Accountability Concepts:** Universal jurisdiction and transnational tort law offer models if state-sponsored algorithmic harms gain recognition as *jus cogens* violations (drawing on *Filártiga* and *Sosa* logic).

***Policy Recommendations for Binding Public Sector AI Accountability*:**

- Revoke Sovereign Immunity for Rights-Violating AI.

- Establish a Public AI Rights Code (transparency, appeal, explanation, correction).

- Require Pre-Deployment Human Rights Impact Assessments for AI in welfare, justice, healthcare.

- Launch Independent National AI Oversight Bodies with suspension authority.

- Codify Red Lines: Prohibit government AI for biometric surveillance without judicial warrant; no behavior-scoring for public benefits; no emotion analysis in schools or courts.

The Moral Firewall champions a normatively sovereign-proof ethical perimeter. When human dignity is at stake, no sovereign stands above the moral law. Obligations to uphold such fundamental AI principles could be considered ***erga omnes***, owed to the international community as a whole, strengthening international scrutiny. Framing these principles as inspired by *jus cogens* also counters "AI exceptionalism" by asserting that enduring human values must be vigorously applied to new technological frontiers.

## VI. Addressing Challenges & The Path Forward: Building Consensus for a Non-Negotiable Ethic

Resistance to the Moral Firewall is expected.

- **"Tech moves too fast for rigid ethics."** (The "Pacing Problem")
    - *Rebuttal*: Then our ethics must be foundational, not reactive. The Firewall is scaffolding, a floor, not a ceiling. As foresight exercises like AI-2027, experts like Buchanan, and concerned observers like Paul Tudor Jones suggest, the pace towards transformative capabilities like AGI/ASI may be even faster than many anticipate.[125] Dario Amodei terms this a "race between interpretability and model intelligence," making foundational safeguards and accelerated understanding more critical, not less.[126] This is compounded by government itself needing to "move faster" and be "more forward-leaning" in its governance approaches[127] to address what is currently a significant "lack of control or regulation."[128] This very challenge underscores the need for principle-based regulation, adaptable to evolving applications, rather than technology-specific rules prone to obsolescence. Foundational safeguards should guide, not merely follow, future development.
- **"You can't measure something abstract like 'dignity.'"** (Definitional Difficulty)
    - *Rebuttal*: Dignity is enforced via proxies: fairness, freedom from manipulation, non-discrimination, and procedural justice. Absence of appeal violated procedural dignity in Robodebt.[129] The Firewall defines measurable thresholds for these proxies. While abstract, concepts like dignity require ongoing interpretation and operationalization through multi-stakeholder dialogue and evolving legal precedent, similar to established *jus cogens* norms.
- **"This adds overhead and will stifle innovation."**
    - *Rebuttal*: The Moral Firewall catalyzes responsible innovation by creating a trusted ecosystem. A common concern, voiced by figures such as Sam Altman, is that robust governance might stifle innovation, advocating instead for a "light-touch regulatory style" similar to the early internet.[130] However, AI's potential for immediate, scaled, and deeply societal impact—leading to risks such as Systemic Algorithmic Subordination or Epistemic Erosion, and potentially rapid ASI development[131]—differentiates it significantly. The proposed non-derogable principles are not prescriptive regulations that micromanage development, but fundamental ethical guardrails. Innovation that respects these boundaries—ensuring dignity, transparency, and accountability—is not stifled but rather directed towards more responsible, sustainable, and ultimately trustworthy pathways, thereby accelerating the adoption of beneficial AI. The focus shifts from mere speed of innovation to its quality, ethical alignment, and human compatibility, defining "acceptable innovation" versus "irresponsible development".

- **"Ethics are culturally relative—you can't universalize this."**
    - *Rebuttal*: Not all values are relative. *Jus cogens* norms themselves reflect a minimum global moral consensus on core human protections. International instruments like the UNESCO Recommendation on the Ethics of AI[132] and the OECD AI Principles[133] affirm dignity, transparency, and accountability as globally relevant norms, reflecting an emerging *opinio juris* in the AI domain. The Moral Firewall builds upon this growing consensus.

- **"This duplicates ISO, NIST, and other standards."** (AI Exceptionalism vs. Existing Frameworks)
    - *Rebuttal*: It completes and elevates them. NIST guides risk mitigation;[134] ISO defines management system controls;[135] the Firewall defines the non-negotiable ethical boundaries—when to say no, regardless of risk mitigation attempts for certain applications. It mandates enforceable ethics, not just better engineering. It does not demand entirely separate legal silos but adapts fundamental legal wisdom to AI's exceptional potential impact. While soft law is flexible and valuable, the gravity of potential AI harms necessitates a core of non-derogable principles to prevent catastrophic outcomes. The notion that current AI safeguards are sufficient is challenged by figures like Geoffrey Hinton, who, despite foundational contributions to AI, now warns explicitly about the inadequacy of existing regulation and the potential for AI to circumvent protective measures.[136]

- **"Who decides what counts as a violation?"** (Enforcement Challenges)
    - *Rebuttal*: Violations are determined the same way human rights and other fundamental legal principles are interpreted and enforced now: publicly, transparently, and pluralistically, through a combination of national regulatory bodies,[137] accountable judicial adjudication, robust participatory processes involving impacted communities,[138] and evolving international cooperation through human rights bodies and treaties.[139] This is not a static determination but an adaptive governance model, evolving with societal understanding and technological capacity.[140] Guiding principles for determination include the primacy of human dignity, evidence-based assessment, proportionality, and continuous learning.[141] This may necessitate adaptive legal and institutional mechanisms, potentially including specialized international adjudicatory bodies for cross-border harms.

***Operationalization and Adoption*:**

- **National Policymakers:** Integrate Firewall principles into national AI strategies and binding legal frameworks. Establish mandatory requirements for transparency, dignity impact assessments, and robust, accessible redress mechanisms. Amend public procurement standards to mandate Firewall alignment for government AI systems.

- **International Bodies (UN, OECD, CoE):** Champion the codification of Moral Firewall principles as fundamental global norms, possibly through a new international convention or by integrating them into existing treaty negotiations. Use these principles to harmonize digital trade rules, ensuring ethical considerations are paramount.

- **Industry and Developers:** Proactively adopt Firewall principles throughout the entire AI lifecycle, from conception to deployment and decommissioning and actively contribute to collaborative research efforts to advance AI interpretability and safety, as called for by leaders like Dario Amodei.[142] Disclose alignment scores in Environmental, Social, and Governance (ESG) reports and align internal audits and governance processes with frameworks like the NIST AI RMF[143] and ISO 42001.[144]

- **Legal and Academic Community:** Conduct further research to refine legal definitions, operationalization strategies, and enforcement mechanisms for the Firewall principles. Develop robust arguments for their non-derogable status under international law. The development and adoption of such frameworks, including significant investment in AI safety and interpretability research as advocated by figures like Hinton and Amodei,[145] by proposing concrete ethical and governance structures essential for that safety. Integrate Firewall concepts into legal and technical curricula and Institutional Review Board (IRB) processes.

- **Civil Society:** Develop grassroots educational materials and "red-flag" checklists based on the Firewall principles for public awareness. Utilize the Firewall framework in advocacy, public interest litigation, and corporate accountability campaigns.

***Navigating Geopolitical Realities*:**

Universal adoption will be gradual. Momentum can build through "coalitions of the willing"—alliances of states, organizations, and companies committed to these principles.[146] Multi-stakeholder diplomacy involving civil society, academia, and ethical businesses is vital to foster broader acceptance and implementation.[147] The AI-2027 Initiative's call to "spark a broad conversation about where we're headed and how to steer toward positive futures" underscores the need for such inclusive dialogues.[148]

Incentives, such as preferential trade conditions for AI systems certified as Firewall-compliant, or access to collaborative research initiatives, can accelerate convergence.[149] The Firewall's insistence on non-derogable thresholds and verifiable transparency inherently counters "ethics-washing," demanding demonstrable commitment rather than superficial declarations.[150]

## VII. Future Directions: Evolving the Moral Firewall and Advancing Foundational Research

The accelerated timelines and transformative potential highlighted by foresight initiatives like AI-2027, expert commentary on the imminent arrival of AGI,[151] including Demis Hassabis's estimate that AGI is potentially "5 to 10 years away,"[152] warnings of AI as an "imminent security threat,"[153] and the critical lag in our ability to understand these complex systems,[154] underscore the critical need for this ongoing development and research to ensure the Firewall remains a robust safeguard. The need for more research to quantify the risks associated with AI development is clear.[155]

The Moral Firewall, as presented, offers a robust framework for human-centered AI governance. However, its principles and implementation must be part of a living, iterative process, adapting to the evolving technological landscape and our deepening understanding of AI's societal impact. This section outlines key areas for future development of this white paper and a concurrent research agenda crucial for strengthening and operationalizing the Firewall, including calls for greater investment in AI safety, interpretability research,[156] enhanced international cooperation, and robust workforce development strategies.[157] The current "absence of concrete actions to address the risks"[158] makes this forward-looking agenda even more imperative.

### A. Operationalization, Verification, and Global Enforcement

- ***Future White Paper Development (v1.5 / 2.0)*:**
    - Detail phased global implementation strategies, including pilot programs and the role of "coalitions of the willing."
    - Elaborate on robust, cross-jurisdictional verification and auditing frameworks for Firewall compliance, incorporating advanced interpretability metrics and measures to detect and counter "ethics-washing."
    - Further specify mechanisms for ensuring compliance and addressing violations by powerful state and corporate actors, considering international legal and diplomatic avenues.
- ***Recommended Research Agenda*:**
    - Developing scalable, technically sound, and culturally adaptable auditing methodologies for the Moral Firewall's principles, with a strong focus on assessing levels of AI interpretability.
    - Modeling international cooperation dynamics, competitive pressures, and enforcement strategies in global AI governance.
    - Designing and evaluating techno-legal mechanisms for verifiable accountability and transparency across complex, international AI development and deployment supply chains.
    - Investigating effective institutional designs for international bodies tasked with overseeing AI safety, interpretability standards, and ethical compliance.

### B. The Moral Firewall's Principles and the Evolution of International Law

- ***Future White Paper Development (v1.5 / 2.0)*:**
    - Provide an expanded jurisprudential analysis of the *sui generis* nature of the Firewall principles, drawing clearer distinctions and connections to the underlying values of *jus cogens* in light of AI's unique characteristics.
    - Incorporate further comparative analysis mapping severe, novel AI-driven harms to the core protections offered by existing peremptory norms, strengthening the argument for their analogous moral and legal weight.
    - Discuss pathways and precedents for the progressive development of international law, including the potential role of state practice and *opinio juris* in solidifying AI-specific governance norms.
- ***Recommended Research Agenda*:**
    - Analyzing the doctrinal evolution of *jus cogens* and general international law in response to previous transformative technological and societal shifts.
    - Conducting comparative legal studies on the thresholds for peremptory norms and their application to systemic algorithmic harms, particularly those with cross-border implications.
    - Identifying and analyzing gaps in existing international human rights and humanitarian law concerning novel AI-driven harms, informing the necessity and scope of *sui generis* principles.

### C. Ensuring Robust, Independent, and Competent Adjudication and Oversight

- ***Future White Paper Development (v1.5 / 2.0)*:**
    - Detail specific governance models for national and international multi-stakeholder oversight bodies, emphasizing structures that ensure independence, technical competence, and resilience against undue influence or regulatory capture.
    - Outline best practices for transparency, due process, right to appeal, and access to effective remedies in adjudicating alleged violations of the Moral Firewall.
    - Explore the integration of participatory and deliberative democratic mechanisms (e.g., citizen assemblies) to enhance the legitimacy and societal alignment of Firewall governance.

- ***Recommended Research Agenda*:**
    - Developing best-practice models for establishing and maintaining independent, technically proficient AI governance and adjudicatory bodies, including safeguards against capture.
    - Assessing the efficacy and scalability of various participatory mechanisms (e.g., citizen juries, deliberative polling) in the context of complex AI policy and oversight.
    - Establishing international standards for capacity building within regulatory, judicial, and legislative bodies to effectively address AI-related challenges, including the interpretation of complex AI systems.
    - Investigating the interface and optimal information flow between internal AI ethics/safety boards within organizations and external public oversight authorities.
    - Pioneering novel methods for public education and engagement to build societal understanding of AI capabilities, risks, and the importance of interpretability (aligning with Amodei n.d.; Klein and Buchanan 2025). Given that deep technical understanding is crucial for effectively utilizing AI tools,[159] this is vital.

Addressing these areas through continued development of the Moral Firewall framework and dedicated research will be essential to translating its principles into effective, globally recognized safeguards for a human-centric AI future.

## VIII. Conclusion: A Non-Negotiable Ethic for an Intelligent Future – An Invitation to Action

Artificial intelligence now shapes decisions about who is seen, served, protected, and who is not. Current governance frameworks are fragmented and reactive, struggling to keep pace with technology's relentless advance—one that forecasts suggest could lead to Artificial Superintelligence within years,[160] presenting what some, like Paul Tudor Jones, term an "imminent security threat to humanity,"[161] while our fundamental understanding of these systems remains dangerously incomplete.[162] The world urgently needs a moral perimeter: a clear, unambiguous boundary that cannot be overridden by code, convenience, or national interest, defining not how advanced AI can become, but how harmful it must never be.

The Moral Firewall offers this shared ethical floor. Rooted in the enduring legal and moral authority of *jus cogens*, it provides a framework adaptable across jurisdictions and deployable at every stage of the AI lifecycle. It is legally grounded, designed to be compatible with and strengthen existing regulatory and standardization efforts like ISO 42001,[163] the NIST AI RMF,[164] and the EU AI Act,[165] and is immediately adaptable for public procurement, comprehensive risk assessments, and independent oversight.

It does not seek to control technology's progress; it seeks to ensure that wherever technology governs, human dignity remains sovereign. This is not a rejection of technology, but a profound recommitment to the fundamental human values that technology must serve. It does not slow innovation;

it gives innovation something profoundly worth accelerating toward—a future where intelligence serves humanity and integrity. Automation must not erase accountability, and opacity must not obscure responsibility.[166] The Moral Firewall is how we protect, and indeed reaffirm, what makes us human in an age of intelligent machines. It is a vital step in preparing society for a future that may arrive sooner, be more transformative, and carry greater inherent risks than widely anticipated.[167]

**Next Steps: From Principle to Practice – *Join Us in Forging This Future***

- **Launch Public Pilots**: Test and refine the Moral Firewall framework through practical application with regulators, industry leaders, and civil society organizations on high-impact AI systems in diverse sectors.

- **Develop Model Clauses and Legislation:** Translate the Firewall principles into concrete model legislative language, contractual clauses for procurement, and standardized modules for international agreements and auditing protocols.

- **Champion Global Treaty Integration**: Advocate vigorously for the inclusion of the Moral Firewall's core tenets within emerging global governance mechanisms like the Global Digital Compact, and ongoing AI treaty negotiations at the OECD, Council of Europe, and other key international fora.

- **Invest in AI Safety, Interpretability, Workforce Development, and Ethical Guidelines:** Drive significant investment in AI safety research and development, with a dedicated acceleration of interpretability research as urged by leading figures;[168] develop strategies for workforce retraining to address labor market disruptions; and continuously refine ethical guidelines and regulations for AI development and deployment.[169]

- **Equip Civil Society and Academia:** Develop and disseminate open-access training guides, practical checklists, and educational curricula to empower civil society organizations, researchers, educators, and the broader public to understand, apply, and advocate for the Moral Firewall.

- **Urge Proactive and Accelerated Government Action:** Discuss the Moral Firewall with government officials and agencies, legislative bodies, sectoral agencies to promote awareness and then call them to act on obtaining technical proficiency to effectively govern AI and potentially form and launch AI national and cross-border ombudspersons offices. Encourage a policy for a "race to the top" in AI Safety, Interpretability, and AI Ethics research and implementation.

We invite policymakers, developers, academics, foresight researchers, investors, and all members of civil society to engage critically with this framework, to test its precepts, refine its applications, and collaborate in its implementation.

The future of artificial intelligence, particularly one that may include ASI, cannot be entrusted to technical audits or market forces alone, especially given the intense competition and current lack of overarching control.[170] Dignity, transparency, and accountability are not design preferences; they are non-negotiable obligations we owe to current and future generations. Let us forge this future together, ensuring that intelligence serves humanity, always.

**Endnotes**

[1] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. Available at: https://ai-2027.com/ (Accessed: 10 May 2025).
[2] Jones, P. T. (2025) *Paul Tudor Jones: AI poses an imminent threat to humanity in our lifetime* [Video]. YouTube (CNBC Television channel). Available at: http://www.youtube.com/watch?v=wrESBnPYoZU (Accessed: 10 May 2025).
[3] Amodei, D. (n.d.) *The Urgency of Interpretability*. Available at: https://www.darioamodei.com/post/the-urgency-of-interpretability (Accessed: 10 May 2025).
[4] Klein, E. & Buchanan, B. (2025) *The Government Knows AGI is Coming | The Ezra Klein Show* [Video]. YouTube (The Ezra Klein Show channel). Available at: https://www.youtube.com/watch?v=Btos-LEYQ30 (Accessed: 10 May 2025).
[5] Time. (2025) *Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic*. [Online video]. 00:01:00. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[6] 60 Minutes. (2025) *What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute*. [Online video]. 00:00:30. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[7] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:05:31. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[8] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:13:23. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[9] Jones, P. T. (2025) op. cit.
[10] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.
[11] Klein, E. & Buchanan, B. (2025) op. cit.
[12] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:05:40. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[13] Jones, P. T. (2025) op. cit.
[14] Amodei, D. (n.d.) op. cit.
[15] Ibid.
[16] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:11:41. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[17] LBC (2025) *'Godfather of AI' predicts it will take over the world | LBC* [Video]. YouTube. Available at: https://www.youtube.com/watch?v=vxkBE23zDmQ (Accessed: 10 May 2025).
[18] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.
[19] Jones, P. T. (2025) op. cit., 00:02:46; Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.; LBC (2025) op. cit.
[20] Amodei, D. (n.d.) op. cit., cited in Digitrendz (2025) 'Anthropic's Dario Amodei Calls for Urgent "Race" to Understand AI's Inner Workings'. *Digitrendz Blog*, 25 April. Available at: https://digitrendz.blog/tech-news/10172/anthropics-dario-amodei-calls-for-urgent-race-to-understand-ais-inner-workings/ (Accessed: 10 May 2025).
[21] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:11:53. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[22] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying

Optimistic. [Online video]. 00:03:55. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[23] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:04:14. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[24] Klein, E. & Buchanan, B. (2025) op. cit.
[25] Jones, P. T. (2025) op. cit. [26] Klein, E. & Buchanan, B. (2025) op. cit.
[27] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:11:46. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[28] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:07:59. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[29] Vienna Convention on the Law of Treaties (1969) Art. 53. United Nations Treaty Series, vol. 1155, p. 331. Available at: https://legal.un.org/ilc/texts/instruments/english/conventions/1_1_1969.pdf (Accessed: 11 May 2025).
[30] United Nations (1969) *Vienna Convention on the Law of Treaties*. United Nations Treaty Series, vol. 1155, p. 331. Available at: https://treaties.un.org/doc/Publication/UNTS/Volume%201155/volume-1155-I-18232-English.pdf (Accessed: 10 May 2025).
[31] ILC (International Law Commission) (2022) *Draft conclusions on identification and legal consequences of peremptory norms of general international law (jus cogens), with commentaries*. A/77/10. Available at: https://legal.un.org/ilc/reports/2022/english/chp4.pdf (Accessed: 10 May 2025).
[32] Ibid.
[33] Ibid.
[34] United Nations Treaty Collection (2024) *Vienna Convention on the Law of Treaties*. Status as at: 10-05-2025. Available at: https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXIII-1&chapter=23&clang=_en (Accessed: 10 May 2025).
[35] American Law Institute (1987) *Restatement (Third) of the Foreign Relations Law of the United States*. (Excerpt). Available at: https://opencasebook.org/casebooks/393-international-law-and-human-rights-fall-2016/resources/2.6.1-restatement-third-of-foreign-relations-law-of-the-united-states-1987-excerpt/ (Accessed: 10 May 2025).
[36] *Filártiga v. Peña-Irala*, 630 F.2d 876 (2d Cir. 1980). Available at: https://hrp.law.harvard.edu/wp-content/uploads/2011/04/filartiga-v-pena-irala.pdf (Accessed: 11 May 2025).
[37] *Siderman de Blake v. Republic of Argentina*, 965 F.2d 699 (9th Cir. 1992). Available at: https://ihl-databases.icrc.org/en/national-practice/siderman-de-blake-v-republic-argentina-court-appeals-ninth-circuit-22-may-1992 (Accessed: 11 May 2025).
[38] ICJ (International Court of Justice) (1986) *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*. Merits, Judgment. I.C.J. Reports 1986, p. 14. Available at: https://www.icj-cij.org/case/70 (Accessed 10 May 2025).
[39] LBC (2025) op. cit.
[40] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., AI-2027 Initiative. (2025) op. cit.
[41] Jones, P. T. (2025) op. cit.
[42] Teo, S.A. (2024) 'Artificial intelligence and its 'slow violence' to human rights'. *AI and Ethics*. Published online 02 February 2024. Available at: https://link.springer.com/article/10.1007/s43681-024-00547-x (Accessed: 11 May 2025).
[43] Ibid.
[44] Teo, S.A. (2024) op. cit., p. 1.
[45] Teo, S.A. (2024) op. cit.
[46] Ibid.

[47] Ibid.
[48] Bygrave, L.A., (2020) 'Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions'. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3721118 (Accessed: 11 May 2025).
[49] Robertson, C.T. (2023) 'Online Echo Chambers, Online Epistemic Bubbles, and Open-Mindedness'. *Episteme*, 20(4), pp. 700-718. Available at: https://www.researchgate.net/publication/375712534_Online_Echo_Chambers_Online_Epistemic_Bubbles_and_Open-Mindedness (Accessed: 11 May 2025).
[50] Bygrave, L.A., (2020) op. cit.
[51] Teo, S.A. (2024) op. cit.
[52] Amodei, D. (n.d.) op. cit.
[53] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.
[54] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:12:41. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[55] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:12:41. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[56] Kramer, A. (2024) *Jus Cogens and Privacy: A Critical Review of Internet Regulation and Personal Freedoms*. ResearchGate. Available at: https://www.researchgate.net/publication/385039754_Jus_Cogens_and_Privacy_A_Critical_Review_of_Internet_Regulation_and_Personal_Freedoms (Accessed: 11 May 2025).
[57] Klein, E. & Buchanan, B. (2025) op. cit.
[58] Jones, P. T. (2025) op. cit.
[59] LBC (2025) op. cit.
[60] Amodei, D. (n.d.) op. cit.
[61] Ibid.
[62] Amodei, D. (n.d.) op. cit., cited in Scale AI (n.d.) *Diagnosing AI: Advancing Interpretability and Evaluations*. Available at: https://scale.com/blog/advancing-interpretability (Accessed: 10 May 2025).
[63] AuditBoard (2024) *Navigating New Regulations for AI in the EU*. Available at: https://auditboard.com/blog/eu-ai-act (Accessed: 9 May 2025); TechPolicy.Press (2023) *Navigating AI Safety: A Socio-Technical and Risk-based Approach to Policy Design*. Available at: https://www.techpolicy.press/navigating-ai-safety-a-sociotechnical-and-riskbased-approach-to-policy-design/ (Accessed: 9 May 2025).
[64] European Union (2024) *Artificial Intelligence Act*. Regulation (EU) 2024/1084. Official Journal of the European Union. Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (Accessed: 11 May 2025).
[65] Amodei, D. (n.d.) op. cit., cited in AiNews.com (2025). Available at: https://www.ainews.com/p/anthropic-ceo-we-must-understand-ai-models-before-2027?utm_source=newsletter&utm_medium=email&utm_campaign=AiNews (Accessed: 11 May 2025).
[66] European Union (2024) op. cit.; Toader, A. (2019) 'Auditability of AI systems – brake or acceleration to innovation?', *SSRN Electronic Journal*. Available at: https://doi.org/10.2139/ssrn.3526222 (Accessed: 9 May 2025).
[67] Felzmann, H., Fosch-Villaronga, E., Lutz, C., and Tamo-Larrieux, ` A. *Towards transparency by design for artificial intelligence, Science and Engineering Ethics*, 26(6), 2020, 3333-3361 cited in Haresamudram, K., Larsson, S., & Heintz, F. (2023). *Three Levels of AI Transparency.* Computer, 56(2), 93-100. Available at: https://doi.org/10.1109/MC.2022.3213181 (Accessed: 11 May 2025).
[68] HÄRTING Rechtsanwälte (2024) *Transparency Obligations in the AI Act*. Available at: https://haerting.de/en/insights/transparenzpflichten-in-der-ki-verordnung/ (Accessed: 9 May 2025).
[69] OECD (2019) *Recommendation of the Council on Artificial Intelligence (OECD AI Principles)*. OECD/LEGAL/

0449. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (Accessed: 9 May 2025).
[70] Amodei, D. (n.d.) op. cit.
[71] APQ, Journal (2025) 'The Four Fundamental Components for Intelligibility and Interpretability in AI Ethics'. *American Philosophical Quarterly*, 62(2), p.103.
[72] OECD (2019) op. cit.; EU AI Act, Recital 27.
[73] Oye, E., et al (2022). *A Comprehensive Comparative Analysis of Explainable AI Techniques.* Accessible at: https://www.researchgate.net/publication/388353445_A_Comprehensive_Comparative_Analysis_of_Explainable_AI_Techniques (Accessed: 11 May 2025).
[74] ASEAN (2023) *ASEAN Guide on AI Governance and Ethics*. Available at: https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf (Accessed: 9 May 2025); Palo Alto Networks (n.d.) *NIST AI Risk Management Framework (AI RMF)*. Available at: https://www.paloaltonetworks.com/cyberpedia/nist-ai-risk-management-framework (Accessed: 9 May 2025).
[75] Burrell, J. (2016) 'How the machine 'thinks': Understanding opacity in machine learning algorithms'. *Big Data & Society*, 3(1). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2660674 (Accessed: 11 May 2025).
[76] UNESCO (2021) *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 9 May 2025).
[77] NTIA (National Telecommunications and Information Administration) (2024) *AI Accountability Policy Report*. Available at: https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report (Accessed: 10 May 2025).
[78] CNBC Television (2025) *OpenAI CEO Sam Altman testifies on AI competition before Senate committee — 5/8/2025* [Video]. YouTube. 00:56:01. Available at: http://www.youtube.com/watch?v=jOqTg1W_F5Q (Accessed: 9 May 2025).
[79] European Union (2024) op. cit., Art. 16.
[80] Formosa, P. (2017). Who Has Dignity? Rational Agency and the Limits of the Formula of Humanity. In: *Kantian Ethics, Dignity and Perfection*. Cambridge University Press; 2017:120-162. Available at: https://www.cambridge.org/core/books/abs/kantian-ethics-dignity-and-perfection/who-has-dignity-rational-agency-and-the-limits-of-the-formula-of-humanity/A3BAFF11C0A5DC6039BB1743FD7885A2 (Accessed: 11 May 2025).
[81] UN General Assembly (1948) *Universal Declaration of Human Rights (UDHR)*. (A/RES/217(III)). Available at: https://www.un.org/en/about-us/universal-declaration-of-human-rights (Accessed: 9 May 2025).
[82] BBC. (2025) *Why Pope Leo chose his name: AI, workers' rights, new Industrial Revolution* [Online]. [10 May]. Available at: https://www.cnbc.com/2025/05/10/pope-leo-name-ai-workers-catholic.html (Accessed: 11 May 2025).
[83] CNBC Television (2025) op. cit.
[84] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.
[85] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:11:35. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[86] Council of Europe (n.d.b) *HUDERIA: New tool to assess the impact of AI systems on human rights*. Council of Europe Portal. Available at: https://www.coe.int/en/web/portal/-/huderia-new-tool-to-assess-the-impact-of-ai-systems-on-human-rights (Accessed: 9 May 2025).
[87] Consensus Academic Search Engine (n.d.) *What Are The Ethical Considerations In The Design And Deployment Of Social Robots?* Available at: https://consensus.app/questions/what-ethical-considerations-design-deployment-social/ (Accessed: 9 May 2025); Kovach, K.A., Funke, A. and Altman, M. (2024) 'Cultivating Dignity in Intelligent Systems'. *Philosophies*, 9(2), p.46.
[88] EDRi (European Digital Rights) (n.d.) *EU privacy regulators and Parliament demand AI and biometrics red lines*.

Available at: https://edri.org/our-work/eu-privacy-regulators-and-parliament-demand-ai-and-biometrics-red-lines/ (Accessed: 9 May 2025).
[89] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:04:39. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[90] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit. [91] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:11:27. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[92] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:11:33. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[93] Ogunleye, I. (n.d.) *AI's Redress Problem: Recommendations to Improve Consumer Protection from Artificial Intelligence*. OCLTC Berkeley (Center for Long-Term Cybersecurity, UC Berkeley). Available at: https://cltc.berkeley.edu/publication/cltc-white-paper-ais-redress-problem/ (Accessed: 9 May 2025).
[94] Ibid.
[95] OECD (2025b) *Towards a common reporting framework for AI incidents*. OECD Publishing. (Information available via OECD.AI). Available at: https://oecd.ai/en/wonk/deepfake-scams-biased-ai-incidents-framework-reporting-can-keep-ahead-ai-harms (Accessed: 10 May 2025).
[96] Cooper et al. (2022) 'Liability in AI Systems'. (Cited in: IRJET (International Research Journal of Engineering and Technology) (2023). 'AI Accountability: Navigating Legal, Ethical, and Transparency Challenges'. *IRJET*, 11(12)).
[97] ScienceOpen (2022) *Algorithmic Impact Assessment for an Ethical Use of AI in SMEs*. (Mbuy, N.K., et al., HCI International 2022). Available at: https://www.scienceopen.com/document_file/c313b027-1620-49b8-99d9-e455d3ea70bf/ScienceOpen/001_Mbuy_HCI2022.pdf (Accessed: 9 May 2025).
[98] U.S. Government (2023) 'AI Accountability Policy Request for Comment'. *Federal Register*, 88(71), pp. 22433-22439.
[99] ISO/IEC (2023) *ISO/IEC 42001:2023 Artificial Intelligence – Management System*. Available at: https://www.iso.org/standard/81228.html (Accessed: 9 May 2025).
[100] NIST (National Institute of Standards and Technology) (2023a) *AI Risk Management Framework (AI RMF 1.0)*. Available at: https://www.nist.gov/itl/ai-risk-management-framework (Accessed: 9 May 2025).
[101] Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) 'Machine bias'. *ProPublica*. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed: 9 May 2025).
[102] Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y.C. (2025) 'The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships'. (Forthcoming, CHI 2025; Preprint available). *arXiv*. Available at: https://arxiv.org/abs/2410.20130 (Accessed: 10 May 2025).
[103] Eysenbach G. Crisis Text Line and Loris.ai Controversy Highlights the Complexity of Informed Consent on the Internet and Data-Sharing Ethics for Machine Learning and Research. J Med Internet Res. 2025 Jan 22;27:e67878. doi: 10.2196/67878. PMID: 39841991; PMCID: PMC11799832. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11799832/ (Accessed: 10 May 2025).
[104] Khera, R. (2017) 'Impact of Aadhaar on Welfare Programmes'. *Economic and Political Weekly*, 52(50). Available at: https://www.jstor.org/stable/45132600 (Accessed: 11 May 2025).
[105] Angwin, J., et al. (2016) op. cit.
[106] European Union (2024) op. cit.
[107] Ahmed, N. and Echi, M. (2021) 'AI surveillance risks discussion'. (Cited in: IJFMR (2025). 'AI-Powered Surveillance vs. Privacy Rights: Striking the Right Balance'. *International Journal for Multidisciplinary Research*,

2(2)). Available at: https://www.ijfmr.com/papers/2025/2/42672.pdf (Accessed: 11 May 2025).
[108] European Union (2024) op. cit.
[109] Klein & Buchanan, (2025) op. cit.; AI-2027 Initiative, (2025) op. cit.
[110] Jones, P. T. (2025) op. cit.
[111] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:11:53. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[112] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.
[113] Klein & Buchanan, (2025) op. cit.
[114] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:00:00. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[115] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:05:44. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[116] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:06:01. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[117] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:12:21. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[118] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:12:28. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[119] 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:12:33. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
[120] Klein & Buchanan, (2025) op. cit.
[121] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.
[122] CNBC Television (2025) op. cit.
[123] Henriques, A. (2020) *Robodebt Investigation*. ABC News; Federal Court of Australia (2021) *Prygodicz v Commonwealth of Australia (No 2)* [2021] FCA 634.
[124] Foxglove Legal (2020) *UK Visa Algorithm Case*. (Information available via Incident Database). Available at: https://incidentdatabase.ai/cite/335/ (Accessed: 10 May 2025).
[125] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.; Klein and Buchanan (2025) op. cit.; Jones (2025) op. cit.
[126] Amodei, D. (n.d.) op. cit.
[127] Klein and Buchanan (2025) op. cit.
[128] Jones (2025) op. cit.
[129] Federal Court of Australia (2021) op. cit.
[130] CNBC Television (2025) op. cit.
[131] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit. [132] UNESCO (2021) op. cit.
[133] OECD (2019) op. cit.
[134] NIST (2023a) op. cit.
[135] ISO/IEC (2023) op. cit.
[136] LBC (2025) op. cit.
[137] AI Ethics Lab (n.d.) *REG: Regulatory and Oversight Bodies*. Rutgers University. Available at: https://aiethicslab.rutgers.edu/glossary/reg/ (Accessed: 9 May 2025).

[138] Vidhi Centre for Legal Policy (2024) *Participatory AI Approaches in AI Development and Governance*. Available at: https://vidhilegalpolicy.in/research/participatory-ai-approaches-in-ai-development-and-governance/ (Accessed: 9 May 2025).
[139] Tandfonline (n.d.) 'On the need for a global AI ethics'. (Possibly Floridi, L., et al.). Available at: https://www.tandfonline.com/doi/full/10.1080/17449626.2024.2425366?af=R (Accessed: 9 May 2025); WHO (World Health Organization) (2022) 'Call for Governance papers: Ethics of artificial intelligence in global health research meeting'. Available at: https://www.who.int/news-room/articles-detail/ethics-of-artificial-intelligence-in-global-health-research-meetings-governance-papers (Accessed: 9 May 2025); Chatham House (2023) *AI governance and human rights*. Available at: https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights (Accessed: 9 May 2025).
[140] Cf. Choung, H., David, P., & Seberger, J.S. (2023) *A multilevel framework for AI governance*. arXiv:2307.03198 [cs.CY]. Available at: https://www.researchgate.net/publication/372234650_A_multilevel_framework_for_AI_governance (Accessed: 11 May 2025).
[141] Cao, et al (2020) 'Adaptive Governance, Loose Coupling, Forward-Looking Strategies and Responsible Innovation'. *ResearchGate*.
[142] Amodei D. (n.d.) op. cit.; Digitrendz (2025) op. cit.
[143] NIST (2023a) op. cit.
[144] ISO/IEC (2023) op. cit.
[145] LBC (2025) op. cit.; Amodei (n.d.) op. cit.
[146] Bepress Legal Repository (n.d.) *'"Coalitions of the Willing" and the Evolution of Informal International Law"*. Available at: https://law.bepress.com/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1031&context=taulwps (Accessed: 9 May 2025).
[147] CIGI Online (Centre for International Governance Innovation) (n.d.) *Advancing Multi-stakeholderism for Global Governance of the Internet and AI*. (Referencing Diya Uberoi). Available at: https://www.cigionline.org/documents/3278/_Diya.pdf (Accessed: 9 May 2025).
[148] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.
[149] Brookings Institution (n.d.) *Strengthening international cooperation on artificial intelligence*. Available at: https://www.brookings.edu/articles/strengthening-international-cooperation-on-artificial-intelligence/ (Accessed: 9 May 2025).
[150] Bernini, et al (2024) 'Artificial Intelligence's (AI's) Responsible Use: How to Manage Digital Ethicswashing'. *ResearchGate*.
[151] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.; Klein and Buchanan (2025) op. cit.
[152] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:11:03. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[153] Jones (2025) op. cit.
[154] Amodei (n.d.) op. cit.
[155] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:07:52. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[156] Amodei (n.d.) op. cit.
[157] Klein and Buchanan (2025) op. cit.
[158] Jones (2025) op. cit.
[159] Time. (2025) Google DeepMind CEO Worries About a "Worst-Case" A.I Future, But Is Staying Optimistic. [Online video]. 00:10:37. Available at: https://youtu.be/i2W-fHE96tc?si=1-Wr1zCNoM0FDh5s (Accessed: 11 May 2025).
[160] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., Kokotajlo, D., Alexander, S., Larsen, T.,

Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.; Klein and Buchanan (2025) op. cit.
[161] Jones (2025) op. cit.
[162] Amodei (n.d.) op. cit.
[163] ISO/IEC (2023) op. cit.
[164] NIST (2023a) op. cit.
[165] European Union (2024) op. cit.
[166] Amodei (n.d.) op. cit.
[167] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.; Klein and Buchanan (2025) op. cit.; Jones (2025) op. cit.
[168] Amodei (n.d.) op. cit.
[169] Klein and Buchanan (2025) op. cit.
[170] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. op. cit.; Jones (2025) op. Cit.

**References**


1. Access Now (2022). *Recommendations on Ethical AI Governance*. Available at: https://www.accessnow.org/recommendations-on-ethical-ai-governance (Accessed: 9 May 2025).
2. Ahmed, N. and Echi, M. (2021). 'AI surveillance risks discussion'. (A direct link to the original paper was not found. This work is cited in the context of AI surveillance risks in: IJFMR (2025). 'AI-Powered Surveillance vs. Privacy Rights: Striking the Right Balance'. *International Journal for Multidisciplinary Research*, 2(2)). Available at: https://www.ijfmr.com/papers/2025/2/42672.pdf (Accessed: 10 May 2025).
3. AI Ethics Lab (n.d.). *REG: Regulatory and Oversight Bodies*. Rutgers University. Available at: https://aiethicslab.rutgers.edu/glossary/reg/ (Accessed: 9 May 2025).
4. American Law Institute (1987). *Restatement (Third) of the Foreign Relations Law of the United States*. (Excerpt). Available at: https://opencasebook.org/casebooks/393-international-law-and-human-rights-fall-2016/resources/2.6.1-restatement-third-of-foreign-relations-law-of-the-united-states-1987-excerpt/ (Accessed: 10 May 2025).
5. Amodei, D. (n.d.). *The Urgency of Interpretability*. Available at: https://www.darioamodei.com/post/the-urgency-of-interpretability (Accessed: 10 May 2025).
6. Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). 'Machine bias'. *ProPublica*. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed: 9 May 2025).
7. APQ Journal (2025). 'The Four Fundamental Components for Intelligibility and Interpretability in AI Ethics'. *American Philosophical Quarterly*, 62(2), p.103. Available at: https://scholarlypublishingcollective.org/uip/apq/article/62/2/103/397186/The-Four-Fundamental-Components-for (Accessed: 9 May 2025).
8. ASEAN (2023). *ASEAN Guide on AI Governance and Ethics*. Available at: https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf (Accessed: 9 May 2025).
9. AuditBoard (2024). *Navigating New Regulations for AI in the EU*. Available at: https://auditboard.com/blog/eu-ai-act (Accessed: 9 May 2025).
10. BBC. (2025) *Why Pope Leo chose his name: AI, workers' rights, new Industrial Revolution.* [Online]. [10 May]. Available at: https://www.cnbc.com/2025/05/10/pope-leo-name-ai-workers-catholic.html (Accessed: 11 May 2025).
11. Bepress Legal Repository (n.d.). *'"Coalitions of the Willing" and the Evolution of Informal International*

*Law"*. (Referencing work by Abraham L. Sofaer). Available at: https://law.bepress.com/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1031&context=taulwps (Accessed: 9 May 2025).
12. Bernini, et al (2024). 'Artificial Intelligence's (AI's) Responsible Use: How to Manage Digital Ethicswashing'. *ResearchGate*. Available at: https://www.researchgate.net/publication/387510746_Artificial_Intelligence's_AI's_Responsible_Use_How_to_Manage_Digital_Ethicswashing (Accessed: 9 May 2025).
13. Brookings Institution (n.d.). *Strengthening international cooperation on artificial intelligence*. Available at: https://www.brookings.edu/articles/strengthening-international-cooperation-on-artificial-intelligence/ (Accessed: 9 May 2025).
14. Burrell, J. (2016). 'How the machine 'thinks': Understanding opacity in machine learning algorithms'. *Big Data & Society*, 3(1). Available at: https://doi.org/10.1177/2053951715622512 (Accessed: 10 May 2025).
15. Calo, R. (2018). 'Artificial Intelligence Policy: A Primer and Roadmap'. *UC Davis Law Review*, 51(2), pp. 399–435. Available at: https://digitalcommons.law.uw.edu/faculty-articles/640/ (Accessed: 10 May 2025).
16. Cambridge University Press (n.d.). 'Artificial Intelligence and the Right to Algorithmic Transparency (Chapter 12)'. In: *The Cambridge Handbook of Information Technology, Life Sciences and Human Rights*. Available at: https://www.cambridge.org/core/books/cambridge-handbook-of-information-technology-life-sciences-and-human-rights/artificial-intelligence-and-the-right-to-algorithmic-transparency/A92EE127AF24D868066EC0AEAE3A370C (Accessed: 9 May 2025).
17. Cao, et al (2020). 'Adaptive Governance, Loose Coupling, Forward-Looking Strategies and Responsible Innovation'. *ResearchGate*. Available at: https://www.researchgate.net/publication/347930843_Adaptive_Governance_Loose_Coupling_Forward-Looking_Strategies_and_Responsible_Innovation_September_2020 (Accessed: 9 May 2025).
18. Chatham House (2023). *AI governance and human rights*. Available at: https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights (Accessed: 9 May 2025).
19. Choung, H., David, P., & Seberger, J.S. (2023). *A multilevel framework for AI governance*. arXiv:2307.03198 [cs.CY]. Available at: https://arxiv.org/abs/2307.03198 (Accessed: 9 May 2025).
20. CIGI Online (Centre for International Governance Innovation) (n.d.). *Advancing Multi-stakeholderism for Global Governance of the Internet and AI*. (Referencing Diya Uberoi). Available at: https://www.cigionline.org/documents/3278/_Diya.pdf (Accessed: 9 May 2025).
21. CLTC Berkeley (Center for Long-Term Cybersecurity, UC Berkeley) (n.d.). *AI's Redress Problem: Recommendations to Improve Consumer Protection from Artificial Intelligence*. Available at: https://cltc.berkeley.edu/publication/cltc-white-paper-ais-redress-problem/ (Accessed: 9 May 2025).
22. CNBC Television (2025). *OpenAI CEO Sam Altman testifies on AI competition before Senate committee — 5/8/2025* [Video]. YouTube. Available at: http://www.youtube.com/watch?v=jOqTg1W_F5Q (Accessed: 9 May 2025).
23. Consensus Academic Search Engine (n.d.). *What Are The Ethical Considerations In The Design And Deployment Of Social Robots?* (Discusses Ethical Participatory Design). Available at: https://consensus.app/questions/what-ethical-considerations-design-deployment-social/ (Accessed: 9 May 2025).
24. Cooper et al. (2022). 'Liability in AI Systems'. (Cited in: IRJET (International Research Journal of Engineering and Technology) (2023). 'AI Accountability: Navigating Legal, Ethical, and Transparency Challenges'. *IRJET*, 11(12)). Available at: https://www.irjet.net/archives/V11/i12/IRJET-V11I1228.pdf (Accessed: 9 May 2025).
25. Council of Europe (n.d.b). *HUDERIA: New tool to assess the impact of AI systems on human rights*. Council of Europe Portal. Available at: https://www.coe.int/en/web/portal/-/huderia-new-tool-to-assess-the-impact-of-ai-systems-on-human-rights (Accessed: 9 May 2025).

26. Council of Europe (2024). *Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law (CAI)*. Available at: https://www.coe.int/en/web/artificial-intelligence/cai (Accessed: 9 May 2025).
27. Creemers, R. (2018). 'China's Social Credit System: An Evolving Practice of Control'. *MERICS*. Available at: https://www.researchgate.net/publication/325749336_China's_Social_Credit_System_An_Evolving_Practice_of_Control (Accessed: 10 May 2025).
28. Digitrendz (2025). 'Anthropic's Dario Amodei Calls for Urgent "Race" to Understand AI's Inner Workings'. *Digitrendz Blog*, 25 April. Available at: https://digitrendz.blog/tech-news/10172/anthropics-dario-amodei-calls-for-urgent-race-to-understand-ais-inner-workings/ (Accessed: 10 May 2025).
29. EDRi (European Digital Rights) (n.d.). *EU privacy regulators and Parliament demand AI and biometrics red lines*. Available at: https://edri.org/our-work/eu-privacy-regulators-and-parliament-demand-ai-and-biometrics-red-lines/ (Accessed: 9 May 2025).
30. Eysenbach, G. (2025). *Crisis Text Line and Loris.ai Controversy Highlights the Complexity of Informed Consent on the Internet and Data-Sharing Ethics for Machine Learning and Research.* J Med Internet Res 2025;27:e67878 Available at: https://www.jmir.org/2025/1/e67878 (Accessed: 10 May 2025).
31. European Union (2024). *Artificial Intelligence Act*. Regulation (EU) 2024/1084. Official Journal of the European Union. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401330 (Accessed: 10 May 2025).
32. Federal Court of Australia (2022). *Prygodicz v Commonwealth of Australia (No 2)* [2021] FCA 634. Available at: https://gordonlegal.com.au/app/uploads/2023/08/prygodicz-v-commonwealth-of-australia-no-2-2021-fca-634-summary.pdf (Accessed 10 May 2025).
33. *Filártiga v. Peña-Irala*, 630 F.2d 876 (2d Cir. 1980). Available at: https://hrp.law.harvard.edu/wp-content/uploads/2011/04/filartiga-v-pena-irala.pdf (Accessed: 10 May 2025).
34. Foxglove Legal (2020). *UK Visa Algorithm Case*. (Information available via Incident Database). Available at: https://incidentdatabase.ai/cite/335/ (Accessed: 10 May 2025).
35. Garvie, C., Frankle, J. and Wide, A. (2016). *The Perpetual Line-Up: Unregulated Police Face Recognition in America*. Georgetown Law Center on Privacy & Technology. Available at: https://www.perpetuallineup.org/ (Accessed: 10 May 2025).
36. HÄRTING Rechtsanwälte (2024). *Transparency Obligations in the AI Act*. Available at: https://haerting.de/en/insights/transparenzpflichten-in-der-ki-verordnung/ (Accessed: 9 May 2025).
37. Haresamudram, H., HnLiem, M.P., & Rijen, F.v. (2023). 'Three Levels of AI Transparency'. *AI and Ethics*, 3, pp. 919-932. Available at: https://doi.org/10.1007/s43681-022-00232-8 (Accessed: 10 May 2025).
38. Clarke, R., et al (2024). *Robodebt: A Socio-Technical Case Study of Public Sector Information Systems Failure*. Australian Journal of Information Systems. Available at: https://www.researchgate.net/publication/383797766_Robodebt_A_Socio-Technical_Case_Study_of_Public_Sector_Information_Systems_Failure (11 May 2025).
39. ICJ (International Court of Justice) (1986). *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*. Merits, Judgment. I.C.J. Reports 1986, p. 14. Available at: https://www.icj-cij.org/case/70 (Accessed 10 May 2025).
40. ILC (International Law Commission) (2022). *Draft conclusions on identification and legal consequences of peremptory norms of general international law (jus cogens), with commentaries*. A/77/10. Available at: https://legal.un.org/ilc/reports/2022/english/chp4.pdf (Accessed: 10 May 2025).
41. IRJET (International Research Journal of Engineering and Technology) (2023). 'AI Accountability: Navigating Legal, Ethical, and Transparency Challenges'. *IRJET*, 11(12). (Citing Cooper et al., 2022). Available at: https://www.irjet.net/archives/V11/i12/IRJET-V11I1228.pdf (Accessed: 9 May 2025).

42. ISO/IEC (2023). *ISO/IEC 42001:2023 Artificial Intelligence – Management System*. Available at: https://www.iso.org/standard/81230.html (Accessed: 9 May 2025).
43. Jones, P. T. (2025). *Paul Tudor Jones: AI poses an imminent threat to humanity in our lifetime* [Video]. YouTube (CNBC Television channel). Available at: http://www.youtube.com/watch?v=wrESBnPYoZU (Accessed: 10 May 2025).
44. Khera, R. (2017). 'Impact of Aadhaar on Welfare Programmes'. *Economic and Political Weekly*, 52(50). Available at: https://www.epw.in/journal/2017/50/special-articles/impact-aadhaar-welfare-programmes.html (Accessed: 10 May 2025).
45. Klein, E. & Buchanan, B. (2025). *The Government Knows AGI is Coming | The Ezra Klein Show* [Video]. YouTube (The Ezra Klein Show channel). Available at: http://www.youtube.com/watch?v=Btos-LEYQ30 (Accessed: 10 May 2025).
46. Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R., (2025). AI-2027 Initiative. Available at: https://ai-2027.com/ (Accessed: 10 May 2025).
47. Kovach, K.A., Funke, A. and Altman, M. (2024). 'Cultivating Dignity in Intelligent Systems'. *Philosophies*, 9(2), p.46. Available at: https://www.mdpi.com/2409-9287/9/2/46 (Accessed: 9 May 2025).
48. Kramer, A. (2024) *Jus Cogens and Privacy: A Critical Review of Internet Regulation and Personal Freedoms*. ResearchGate. Available at: https://doi.org/10.13140/RG.2.2.11876.59522 (Accessed: 9 May 2025).
49. LBC (2025). *'Godfather of AI' predicts it will take over the world | LBC* [Video]. YouTube. Available at: http://www.youtube.com/watch?v=vxkBE23zDmQ (Accessed: 9 May 2025).
50. MDPI (2024). 'Prediction of Students' Adaptability Using Explainable AI in Educational Machine Learning Models'. *Applied Sciences*, 14(12), p.5141. Available at: https://www.mdpi.com/2076-3417/14/12/5141 (Accessed: 9 May 2025).
51. NIST (National Institute of Standards and Technology) (2023a). *AI Risk Management Framework (AI RMF 1.0)*. Available at: https://www.nist.gov/itl/ai-risk-management-framework (Accessed: 9 May 2025).
52. NTIA (National Telecommunications and Information Administration) (2024). *AI Accountability Policy Report*. Available at: https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report (Accessed: 10 May 2025).
53. OECD (2019). *Recommendation of the Council on Artificial Intelligence (OECD AI Principles)*. OECD/LEGAL/0449. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (Accessed: 9 May 2025).
54. OECD (2025b). *Towards a common reporting framework for AI incidents*. OECD Publishing. (Information available via OECD.AI). Available at: https://oecd.ai/en/wonk/deepfake-scams-biased-ai-incidents-framework-reporting-can-keep-ahead-ai-harms (Accessed: 10 May 2025).
55. Oye, E., et al (2022). *A Comprehensive Comparative Analysis of Explainable AI Techniques.* Accessible at: https://www.researchgate.net/publication/388353445_A_Comprehensive_Comparative_Analysis_of_Explainable_AI_Techniques (Accessed: 11 May 2025).
56. Palo Alto Networks (n.d.). *NIST AI Risk Management Framework (AI RMF)*. Available at: https://www.paloaltonetworks.com/cyberpedia/nist-ai-risk-management-framework (Accessed: 9 May 2025).
57. Richmond, K.M., Muddamsetty, S.M., Gammeltoft-Hansen, T. *et al.* Explainable AI and Law: An Evidential Survey. *DISO* 3, 1 (2024). https://doi.org/10.1007/s44206-023-00081-z
58. Turner, C. (2023). 'Online Echo Chambers, Online Epistemic Bubbles, and Open-Mindedness'. *Episteme*, 20(4), pp. 700-718. Available at: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/16DAD288417F00A11635C7B129B258BB/

S1742360023000527a.pdf/
online_echo_chambers_online_epistemic_bubbles_and_openmindedness.pdf (Accessed: 9 May 2025).
59. Scale AI (n.d.). *Diagnosing AI: Advancing Interpretability and Evaluations*. Available at: https://scale.com/blog/advancing-interpretability (Accessed: 10 May 2025).
60. ScienceOpen (2022). *Algorithmic Impact Assessment for an Ethical Use of AI in SMEs*. (Mbuy, N.K., et al., HCI International 2022). Available at: https://www.scienceopen.com/document_file/c313b027-1620-49b8-99d9-e455d3ea70bf/ScienceOpen/001_Mbuy_HCI2022.pdf (Accessed: 9 May 2025).
61. 60 Minutes. ([Actual Year of Video Publication]) *[Exact Title of 60 Minutes YouTube Video Featuring Demis Hassabis Here]*. [Online video]. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
62. *Siderman de Blake v. Republic of Argentina*, 965 F.2d 699 (9th Cir. 1992). Available at: https://ihl-databases.icrc.org/en/national-practice/siderman-de-blake-v-republic-argentina-court-appeals-ninth-circuit-22-may-1992 (Accessed: 10 May 2025).
63. *Sosa v. Alvarez-Machain*, 542 U.S. 692 (2004). Available at: https://supreme.justia.com/cases/federal/us/542/692/ (Accessed: 10 May 2025).
64. Tandfonline (n.d.). 'On the need for a global AI ethics'. (Possibly Floridi, L., et al.). Available at: https://www.tandfonline.com/doi/full/10.1080/17449626.2024.2425366?af=R (Accessed: 9 May 2025).
65. Misra, G., et al (2024) *Navigating AI Safety: A Socio-Technical and Risk-based Approach to Policy Design*. TechPolicy.Press Available at: https://www.techpolicy.press/navigating-ai-safety-a-sociotechnical-and-riskbased-approach-to-policy-design/ (Accessed: 9 May 2025).
66. Teo, S.A. (2024). 'Artificial intelligence and its 'slow violence' to human rights'. *AI and Ethics*. Published online 02 February 2024. Available at: https://link.springer.com/article/10.1007/s43681-024-00547-x (Accessed: 9 May 2025).
67. 60 Minutes. (2025) What's next for AI at DeepMind, Google's artificial intelligence lab | 60 Minute. [Online video]. 00:12:41. Available at: https://youtu.be/1XF-NG_35NE?si=QZ66zUZJtNeAxDmB (Accessed: 11 May 2025).
68. Toader, A. (2019) 'Auditability of AI systems – brake or acceleration to innovation?', *SSRN Electronic Journal*. Available at: https://doi.org/10.2139/ssrn.3526222 (Accessed: 9 May 2025).
69. UN General Assembly (1948). *Universal Declaration of Human Rights (UDHR)*. (A/RES/217(III)). Available at: https://www.un.org/en/about-us/universal-declaration-of-human-rights (Accessed: 9 May 2025).
70. UN HRC (United Nations Human Rights Council) (2018). *Report of the Special Rapporteur on the right to privacy, Joseph Cannataci*. (A/HRC/39/29). United Nations Digital Library. Available at: https://digitallibrary.un.org/record/1640588/files/A_HRC_39_29-EN.pdf (Accessed: 10 May 2025).
71. UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 9 May 2025).
72. United Nations (1969). *Vienna Convention on the Law of Treaties*. United Nations Treaty Series, vol. 1155, p. 331. Available at: https://treaties.un.org/doc/Publication/UNTS/Volume%201155/volume-1155-I-18232-English.pdf (Accessed: 10 May 2025).
73. United Nations Treaty Collection (2024). *Vienna Convention on the Law of Treaties*. Status as at: 10-05-2025. Available at: https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXIII-1&chapter=23&clang=_en (Accessed: 10 May 2025).
74. U.S. Government (2023). 'AI Accountability Policy Request for Comment'. *Federal Register*, 88(71), pp. 22433-22439. Available at: https://www.federalregister.gov/documents/2023/04/13/2023-07776/ai-accountability-policy-request-for-comment (Accessed: 9 May 2025).
75. Bygrave, L.A., (2020). 'Machine Learning, Cognitive Sovereignty and Data Protection Rights with

Respect to Automated Decisions'. In: Casey, B., Mittelstadt, B. and Vayena, E. (eds.) *The Cambridge Handbook of Information Technology, Life Sciences and Human Rights*. Cambridge: Cambridge University Press, pp. 205-220. Available at: https://www.cambridge.org/core/books/cambridge-handbook-of-information-technology-life-sciences-and-human-rights/machine-learning-cognitive-sovereignty-and-data-protection-rights-with-respect-to-automated-decisions/A1D153F5D7D4461EAF5B3B965E4B9612 (Accessed: 9 May 2025).

76. Vidhi Centre for Legal Policy (2024). *Participatory AI Approaches in AI Development and Governance*. Available at: https://vidhilegalpolicy.in/research/participatory-ai-approaches-in-ai-development-and-governance/ (Accessed: 9 May 2025).
77. WHO (World Health Organization) (2022). 'Call for Governance papers: Ethics of artificial intelligence in global health research meeting'. Available at: https://www.who.int/news-room/articles-detail/ethics-of-artificial-intelligence-in-global-health-research-meetings-governance-papers (Accessed: 9 May 2025).
78. Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y.C. (2025). 'The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships'. (Forthcoming, CHI 2025; Preprint available). *arXiv*. Available at: https://arxiv.org/abs/2410.20130 (Accessed: 10 May 2025).