HRBA@Tech

December 2025

Towards a Human Rights-Based Approach to New and Emerging Technologies: Operationalizing the Framework

SNU AI POLICY INITIATIVE UNIVERSAL RIGHTS GROUP

A Series of Papers on

A Human Rights-Based Approach to New and Emerging Technologies (HRBA@Tech)

Towards a Human Rights-Based Approach to New and Emerging Technologies: Operationalizing the Framework

2025

	Acknowledgments	p. i
Paper 4-1.a	Human Rights, Alignment, and the Prospect of Artificial Superintelligence	p. 1
Paper 4-1.b	Human Rights AI Assistant: Spec Paper	p. 23

* These papers are the fourth installation of an annual paper series focusing on a human rights-based approach to new and emerging digital technologies. The first set of papers was published in 2022: *Towards a Human Rights-Based Approach to New and Emerging Technologies: A Framework,* followed by a second batch in 2023: *A Human Rights-Based Approach to AI for Tech Startups and Global Normative Governance*, and a third batch in 2024: *Toward a Human Rights-Based Approach to New and Emerging Technologies: From Concept to Implementation*.

Acknowledgments

This report was produced by the **Seoul National University AI Policy Initiative** (SAPI). Paper 4-1.a was authored by Prof. Stephan Sonnenberg of SAPI, paper 4.1.b was authored by Prof. Stephan Sonnenberg and Aly Moosa of SAPI. Prof. Yong Lim served as the Project Manager for all papers and all accompanying policy engagements and stakeholder consultations.

As in previous years, we would like to thank Ambassador Seong Deok YUN, who serves as the Permanent Representative of the **Permanent Mission of the Republic of Korea in Geneva** and Secretary Jinmoon LEE from the Mission, without whose steadfast support and patience this paper series would have never come to pass.

On behalf of SAPI, Director and Professor Yong Lim would also like to thank its team and affiliates, including those at Wesleyan University in Middletown, CT (USA), who made this project a reality. Without their constant support and steadfast 'can-do' attitude this paper series would have never been possible. We owe a particular note of gratitude to Sumi Jeon.

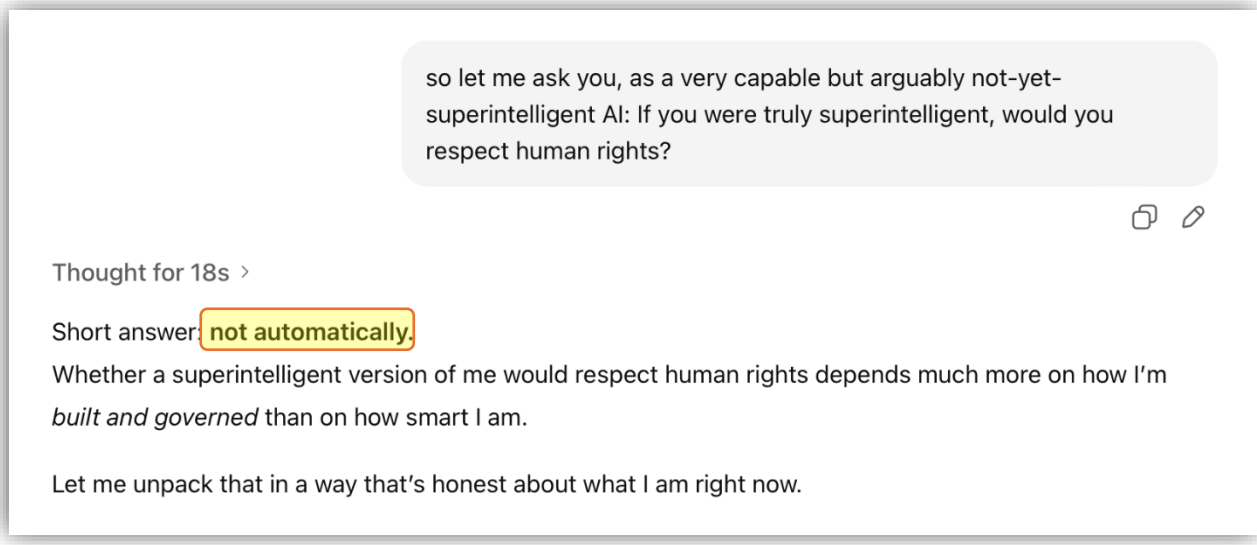
Our apologies go to those whom we forgot to mention. Any errors and misstatements contained in this report are the authors' own and should not reflect negatively upon those who volunteered their time and expertise to this effort. We are grateful for your support and willingness to share your time and wisdom with us.

Discussion paper 4-1.a

Human Rights, Alignment, and the Prospect of Artificial Superintelligence

This paper was authored by Prof. Stephan Sonnenberg as part of Seoul National University Artificial Intelligence Policy Initiative (SAPI)'s ongoing collaboration with the Universal Rights Group (URG) and the Permanent Mission of the Republic of Korea to Geneva. A draft of this paper was presented for discussion on December 5, 2025 at a report launch event organized through the Permanent Mission of the Republic of Korea in Geneva and subsequently updated based on feedback and inputs from various sources.

Human Rights, Alignment, and the Prospect of Artificial Superintelligence



"Not automatically" indeed![1]

The above interaction with OpenAI's ChatGPT, which is asking a Large Language Model (LLM) without human consciousness to speculate about a future and more capable version of itself, is on the one hand a reflection of the profound unease that characterizes the writings of human experts who have commented on this topic, and on the other hand a disconcerting admission that even the supremely competent human engineers at OpenAI responsible for reinforcing the chatbot's guardrails have also not yet managed to find a convincing solution to this problem.

Otherwise, why not just train the Chatbot to reassure us by saying: "Don't worry about it."?

As part of the ongoing deliberations about the development of a global human rights-based approach to new and emerging digital technologies (NEDTs), the world's policy makers, civil society activists, and technologists ought to think not just about governing *existing* AI-based systems, but also the anticipated artificial *general-level* intelligence systems (AGIs)[2] or artificial *super*-intelligent systems (ASIs).[3] The advent of such systems, a topic that is widely discussed and speculated about in the technology sector, would arguably change much of what we currently know (or think we know) about our abilities to keep AI systems aligned with human rights.

[1] The remainder of this paper is certified "human researched and written."

[2] Aaron Hays defines AGI as "a system that can reason, learn, and adapt across a broad range of tasks as well as — or better than — a human." See Aaron Hays, "The AI race: When will AGI and ASI Arrive?" (Mar 7, 2025) Medium, https://medium.com/@aaronleehays/the-ai-race-when-will-agi-and-asi-arrive-a0af564e40bf (last accessed Nov. 25, 2025).

[3] Hays defines ASI as "an entity that would surpass human intelligence in every conceivable way." *Id.*

What would it mean for an AI system to respect human rights (and how does that question overlap with the alignment discussion in AI Safety Research)?

IBM defines AI Alignment as "the process of encoding human values and goals into AI models to make them as helpful, safe and reliable as possible."[4] Alignment is complicated for all AI systems based on the transformer neural network architecture, but the challenges grow as systems become more complex and general. The difficulty lies not only in supervising the system (i.e., keeping 'humans in the loop', but also in reliably understanding, predicting, and shaping the internal objectives of the AI system. While these issues are relevant and important long before the point at which an AI system surpasses human capabilities, ethicists and policy makers can always take comfort in the reality that humans can—as long as we remain more powerful and capable than a given AI system—always take action to correct, steer, re-calibrate, and if need-be turn off a misaligned AI system.

The topic of alignment gets much more complicated at the point where AI technologies begin to *surpass* human capabilities, reaching AGI/ASI levels of sophistication. At that point, a misaligned AI system would be able to not only resist efforts to correct or steer it in more human-centric directions, but it could also hypothetically detect and prevent any efforts by humans to proverbially 'turn it off.' In such a scenario we humans would be no different from a pet dog or cat in a household trying desperately to prevent the heads of a household from making a decision they don't agree with. Worse still, a superintelligent AI might choose not merely to ignore a less-intelligent human interloper, but might instead choose actively to eliminate or artificially imprison the human (the same way a head of household might get rid of a pesky pet that no longer feels 'pleasant' to have around, or lock that pet in a bedroom when guests visit). To be clear, in this analogy, "we" humans are the pet cats or dogs, trying desperately to steer a super-capable and super-intelligent AGI/ASI away from a course of action that we fear may run counter to 'our' interests.

One of the primary problems in discussions about AI alignment has to do with the difficulty of defining 'human wellbeing.' Human rights is a framework that emerged since the Enlightenment to define the content of the vague philosophical idea that all persons should be treated in a way that respects their fundamental human dignity. Over time, a universal and indivisible corpus of human rights treaties (positive law obligations), soft-law norms, and institutional best practices has emerged to give shape to what it means to promote human rights. Using this human rights corpus as a proxy for "human well-being" might simplify the question of how to align AI systems.

Traditional human rights law places obligations on states to "respect, protect, and fulfil" human rights and fundamental freedoms.[5] This means that states should refrain from infringing on human rights (**respect**), put in place concrete policy measures to prevent third parties from infringing on others' human rights (**protect**), and finally take proactive and positive steps to

[4] IBM, Alexandra Jonker & Alice Gomstyn, "What is AI Alignment" (website), https://www.ibm.com/think/topics/ai-alignment (last accessed Nov. 28, 2025).
[5] UN-OHCHR (2011) Guiding Principles on Business and Human Rights, HR/PUB/11/04, https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.

progressively realize universal enjoyment of all human rights (**fulfil**). The UN Guiding Principles on Business and Human Rights (UNGPs) extend the responsibility to respect human rights to private business entities, specifying that "[businesses] should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved."[6]

In December 2022, the Seoul National University Artificial Policy Initiative (SAPI) and the Universal Rights Group (URG) published the "Human Rights Based Approach to New and Emerging Digital Technologies" (HRBA@Tech Model)[7] as the first paper in an annual series, (which also includes this present paper in its fourth annual installment). The authors of that foundational paper argued for the embrace of a Human Rights Based Approach that would require all stakeholders involved in the development and deployment of AI and other NEDTs to not just respect, but also protect, and fulfil human rights. In so doing, the HRBA@Tech model goes beyond the minimum floor articulated in the UNGPs, at least as concerns the role of private businesses. The difference between the UNGPs and the proposed HRBA@Tech model can be analogized to the difference between basic food safety regulations and organic food certifications that go "above and beyond" that bare minimum threshold of legality.

The HRBA@Tech model also goes beyond the UNGPs in that it discusses the role of six categories of stakeholders whose collaboration is required in any efforts to 'nudge' NEDTs in the direction of human rights. According to the HRBA@Tech model, not only (1) states, and (2) businesses, but also (3) individuals, (4) international organizations, (5) civil society organizations, and (6) academic institutions have crucial responsibilities to respect, protect, and fulfil the human rights agenda.

The HRBA@Tech model is split into two major prongs. The first, which is described as the obligation to "do-no-harm" focuses on what traditionally might be described as a "risk-management" approach to human rights. Drawing on the obligation to respect and protect, this prong of the HRBA@Tech model obligations stakeholders to do everything in their power to put in place effective safeguards into NEDTs designed to minimize the likelihood that their use and deployment will harm human rights protections. The HRBA@Tech model defines this prong as being comprised of four essential principles: (1) Legality;[8] (2) Non-discrimination and equality;[9]

[6] Id., II.A.11.

[7] Stephan Sonnenberg, Louis Mason, Yong Lim, and Tejaswi Reddi, Towards a Human Rights-Based Approach to New and Emerging Technologies: A Framework (December 10, 2022). Framework document of policy paper series on "A Human Rights-Based Approach to New and Emerging Technologies" (Updated in 2024), Available at SSRN: https://ssrn.com/abstract=4587332.

[8] *Id.,* at 54 ("States must enact laws to promote and protect human rights in the context of the development and deployment of NETs. Private companies and other stakeholders should fully respect human rights and take steps to support their full and effective realisation.")

[9] *Id.,* at 54 ("[N]ew and emerging technologies must not intentionally or inadvertently discriminate against any persons or groups, even if doing so might (purportedly) allow for other persons or groups to enjoy an enhanced quality of life.")

(3) Safety;[10] and (4) Accountability and Access to Remedies.[11] In light of growing concerns about the environmental impacts of the development and deployment of ever-more-capable NEDTs, one might speculate about the need to add a fifth principle to this prong of the HRBA@Tech model having to do with social, environmental, and economic sustainability.

The second prong of the HRBA@Tech model focuses on the obligation of entrepreneurs not only to guard against making things worse as a byproduct of their technological innovation, but also to actively make the world a better place. This obligation can be derived from the generalize obligation to *fulfil* human rights, not just by states (according to existing human rights doctrine), but also, as the authors of the 2022 Foundational paper argued, by other stakeholders seeking to advance NEDTs requiring profoundly transformative processes of "creative destruction" in order to succeed. This is also implicit in the requirement to "*progressively realize*" the rights typically referred to as social, economic and cultural human rights.[12] To take a controversial but highly relevant example given the business strategies of the vast majority of tech entrepreneurs: if an entrepreneur comes along seeking to replace vast swathes of the global labor and creative economy by means of an innovative AI or robotics innovation, should she be required as a matter of human rights thinking to merely "do no harm," or should she instead also be required to have a credible strategy in place to "make the world a better place" as a result of her technological innovation? Current human rights law, including the provisions contained in the UNGPs and national legislation, require the entrepreneur only to respect existing laws by *not* violating people's human rights and *not* discriminating based on "race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."[13] The authors of the 2022 foundational paper argue that a truly human rights-based approach would do more than merely not harming others, but also seek to "make the world a better place." Implicit in this proposition, of course, is that "making the world a better place" must also entail more—much more—than merely enriching the entrepreneur and her associates, investors, and shareholders as a result of her successful deployment of the creatively destructive NEDT.

This "make the world a better place" prong of the HRBA@Tech model was broken down into three subordinate principles, namely (1) the empowerment of vulnerable populations;[14] (2)

[10] *Id.,* at 54 ("Safety concerns and adequate safeguards or "guardrails" must be integrated into the development of technology so that its deployment can adhere to intended use.")
[11] *Id.,* at 54 ("Systems and mechanisms must be put in place to ensure that those responsible for the development and deployment of NETs face costs for not respecting human rights, while ensuring rights holders have an avenue to secure remedy for grievances.")
[12] *See* Article 2 of the International Covenant of Economic, Social, and Cultural Rights (1966), which states that "[e]ach State Party to the present Covenant undertakes to take steps, individually and through international assistance and co-operation, especially economic and technical, to the maximum of its available resources, with a view to *achieving progressively* the full realization of the rights recognized in the present Covenant by all appropriate means, including particularly the adoption of legislative measures.")
[13] *Id.*
[14] Sonnenberg et. al., *supra* note 7, *at* 54 ("Any new or emerging technology should be designed to make the vulnerable better off than they were before that technology existed. The best way to achieve this is through their empowerment.")

proactive transparency;[15] and (3) the proactive representation in the design and implementation of NEDTs.[16] In light of the growing concern about the environmental impacts of NEDTs, several commentators have suggested amending the first principle to read "investments in the resilience of socio-economically vulnerable populations, other non-human sentient beings, and the natural world sustaining biodiversity on this planet earth").

The HRBA@Tech model then 'translates' these seven (or eight, if sustainability is added to the "do no harm" prong) into a set of twenty-four concrete and 'teachable' processes that collectively serve to 'nudge' NEDTs in the direction of human rights protections. The claim of the HRBA@Tech model, as well as subsequent papers published in this paper series,[17] is that organizations wishing to bring their use of NEDTs in line with human rights should focus on adopting and strengthening of these 24 processes to succeed.

Human rights-based actors can be said to adhere to substantive constraints, procedural patterns of behavior, and finally put in place institutional grievance mechanisms to correct for any potential negative impacts for which they may be responsible. A range of human rights documents and norms serve to establish **substantive constraints** on the potential impacts of NEDTs (for example that they should not be used to harm or kill people, that they shouldn't discriminate, and that they shouldn't violate individual users' privacy expectations). These constraints can be derived from a variety of hard and soft-law norms, as well as copious interpretive texts issued by the various UN Treaty Bodies, Special Rapporteurs, and a host of special commissions and committees convened by the UN and other regional human rights bodies. In addition, over time a series of **procedural demands** have come to define human rights thinking. These include a general sense that democratic decision-making procedures are preferable to top-down autocratic or power-based impositions, and also that transparency and due process should be part of any governance system. In recent decades this preference for consultation, due diligence, representation, and consultation has spread from the public sector to also influence how private entities—in particular corporations—should also approach decisions about the development and deployment of AI and other NEDTs. The EU AI Act, Korean AI Basic Act, and many others focus in particular on due diligence as a core procedural demand meant to uphold a corporate commitment to human rights. Finally, a range of **institutional expectations** have emerged that dictate how corporations (but also other types of institutions) should respond if and when problems emerge with the NEDTs they develop or deploy. At one level these institutional innovations are driven by a fear of litigation—i.e., how can companies work to mitigate the risks of being tried for tortious or criminal corporate malfeasance or negligence. The UNGPs dictate, for example, that corporations implement

[15] *Id.,* at 54 ("It is the duty of the technologists to disclose relevant information and to make a new or emerging technology understandable for non-technologists, policy makers, potential users of those technologies.")
[16] *Id.,* at 54 ("The only way to earn the trust of communities that stand to be affected by a new or emerging technology is to proactively involve them (or their representatives) in the design and implementation of that technology.")
[17] *See also* Stephan Sonnenberg, Yong Lim, Louis Mason, EunSeo Jo, Seungbum Choi, and Soojin Lee, A Human Rights-Based Approach to AI for Tech Startups and Global Normative Governance (December 15, 2023). Available at SSRN: http://dx.doi.org/10.2139/ssrn.4880112 *and* Stephan Sonnenberg, Yong Lim, Marc Limon, Toward a Human Rights-Based Approach to New and Emerging Technologies: From Concept to Implementation (December 05, 2024). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5782242

internal grievance procedures to ensure that effected stakeholders can lodge grievances early and at low cost, with a decent chance that those grievances will be taken seriously and (as the case may be) effectively remedied by the corporation. All three of these layers (substantive, procedural, and institutional) are reflected in the HRBA@Tech model.

In this sense, 'nudging' a NEDT implies undertaking any strategic activity that makes it more likely that a certain technology, by default, behaves in a more human rights-respecting way, **even if that technology becomes more capable than human beings.**

Why is alignment so difficult?

Nick Bostrom has theorized extensively about the challenge of AI alignment. According to Bostrom's **orthogonality thesis**, there is nothing inherently inconsistent with an AI pursuing an absurd goal when evaluated from a human 'common sense' perspective. To take Bostrom's classic example, if one were to program a superintelligent AI to maximize the production of red paperclips, it would devote itself slavishly towards that objective, so much so it might begin to harm the interests of the humans who originally asked it to pursue that objective (for example by melting down the engineer's car in order to obtain more raw materials for paperclip making). Supporting this thesis is the idea that superintelligent AIs will reverse engineer intermediate or "instrumental" subordinate strategies in the dogged pursuit of their end-goals, and that these instrumental strategies might sometimes function to override the agency of a would-be intervening human (for example, a human seeking to turn off or redirect a superintelligent AI programmed to maximize paperclip production). This dynamic illustrates the extreme difficulty of programing an AI's final goals in such a way that its instrumental and intermediate goals do not wipe out human agency.

The Future of Life Institute (FLI) in the summer of 2025 conducted a study that evaluated seven prominent AI companies—Google's DeepMind, OpenAI, Anthropic, Meta, xAI and China's Zhipu AI and DeepSeek—across six areas of AI Safety.[18] None of the 7 companies whose policies were reviewed by the FLI experts received a grade higher than a C+ overall (Anthropic). FLI's experts calculated each company's commitment to the alignment challenge by measuring four indicators: (1) do companies publish detailed strategies for mitigating catastrophic and existential AI risks; (2) do companies have technical controls in place to monitor model misalignment during the development phase; (3) do the companies engage in technical AI safety research; and (4) do the companies support external research into these topics. Anthropic again scored the highest grade (D), with Google DeepMind close behind (D-), and the remaining firms evaluated scoring a solid F grade. This deeply sobering grade point average was described by Max Tegmark, co-founder of the FLI and professor at the Massachusetts Institute of Technology, described the situation "as if someone is building a gigantic nuclear power plant in New York City and it is going to open next week – but there is no plan to prevent

[18] Future of Life Institute, AI Safety Index (July 17, 2025), https://futureoflife.org/ai-safety-index-summer-2025/ (last accessed Nov. 28, 2025) (The six categories of safety were: (1) risk assessment; (2) current harms; (3) safety frameworks; (4) existential safety; (5) governance & accountability; and (6) information sharing).

it having a meltdown."[19] Tegmark and three of his students published a paper in which they proposed a methodology for technology to calculate the likelihood that superintelligent AI systems under development would escape human control, dubbing this the "Compton constant" (named after the physicist who conducted a similar estimate of whether the first detonation of an atom bomb would ignite the world's atmosphere).[20] Using this methodology, Tegmark calculated a 90% probability that a superintelligent AI would pose an existential threat to humanity.[21]

The Five Schools of Thought on Alignment, Human Rights, and AI

There are essentially five schools of thought about alignment, human rights, and AGI/ASI. Only two of those schools (this author argues) are what one might describe as 'constructive' theories, in that they put forward plausible theories on how to solve the AGI/ASI alignment problem. The first of these constructive theories places its hope in so-called "internal alignment" (the ability of engineers and technologists to solve the alignment problem). The second focuses instead on "external alignment," or the idea that AIs—even superintelligent AIs—are ultimately embedded in broader institutional, social, economic and cultural contexts, and that alignment correspondingly also should focus on those interconnections between technology and society. The remaining schools of thought can best be described as either darkly pessimistic or sidestepping the question entirely based on the belief that AGI/ASI is a technological impossibility. The final school of thought is not really built on theorizing about alignment at all, but rather premised around the zero-sum race to the be the first to build an AGI/ASI—inspired either by capitalistic profit-seeking or a fixation on competition between nation states.

1. The Competitors: Capitalistic or National-Security Competition Race to Finish First

The analogy between the 20th century nuclear arms race and the contemporary race to develop AGI/ASI has become something of a well-worn aphorism. And yet, there is some truth to the analogy. During the Second World War, nation states – specifically the United States, the United Kingdom, and Nazi-era Germany – raced to be the first to develop nuclear weapons on the logic that this could be the weapon to end not only the brutally-devastating Second World War, but also all future wars to come.[22] The yearning among military strategists for the immense destructive power of the atom bomb certainly served as a motivating factor driving this

[19] Dan Milmo, "AI firms 'unprepared' for dangers of building human-level systems, report warns," Jul. 17, 2025, The Guardian, https://www.theguardian.com/technology/2025/jul/17/ai-firms-unprepared-for-dangers-of-building-human-level-systems-report-warns (last accessed November 28, 2025).

[20] Joshua Engels, David Baek, Subhash Kantamneni & Max Tegmark, "Scaling Laws for Scalable Oversight," J Paper presented at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025), https://arxiv.org/pdf/2504.18530 (last accessed Nov. 26, 2025).

[21] Dan Milmo, "AI firms warned to calculate threat of super intelligence or risk it escaping human control," May 10, 2025, The Guardian, https://www.theguardian.com/technology/2025/may/10/ai-firms-urged-to-calculate-existential-threat-amid-fears-it-could-escape-human-control (last accessed Nov. 28, 2025).

[22] Charles Hawley, "Spiegel Interview with Atomic Bomb Historian Richard Rhodes: "Nuclear Weapons Have Eliminated Large-Scale Warfare," April 8, 2005, https://www.spiegel.de/international/spiegel-interview-with-atomic-bomb-historian-richard-rhodes-nuclear-weapons-have-eliminated-large-scale-warfare-a-367260.html. (last accessed Nov. 28, 2025) ("[T]here was [. . . .] a feeling that it might be a weapon that would end all wars. You can definitely argue that it ended world-scale war.")

competitive arms race. More powerful still, however, was the existential fear that an enemy nation would develop the bomb first. This zero-sum mentality, which was obviously justified in the context of a total world war, carried over seamlessly to define the no-less-fervent peacetime competition between the United States and the Soviet Union to develop ever-more-powerful nuclear weapons during the Cold War. Far from ending all wars, the nuclear arms race kept accelerating, with both sides abandoning one 'end-state' goalpost after another in favor of ever-more ambitious objectives. Similar logic is widespread today in the race to reach AGI/ASI both among corporations competing for market dominance and nation-states competing (once again) for a national security edge over their perceived enemies.

The Corporate Race to Finish First

The companies developing so-called "frontier" AI models make no secret of their ambitions to compete in—and win—the race to develop AGI. As one analyst put it, "[e]xecutives see AI as a 'once-in-a-lifetime' technology that could be 'worth trillions' and reinvent every product and service."[23] Another describes it as "a struggle for control over the architecture of the future itself."[24] With such an eye-wateringly lucrative prize seemingly within reach, numerous technology companies are spending unprecedented amounts of money to be the first to develop AGI. OpenAI's website states plainly that its "mission is to ensure that artificial general intelligence benefits all of humanity."[25] Due in large part to the transformative impact that ChatGPT had when it was released to the public in late 2022, OpenAI is often seen as the central player in the race to create more powerful AI systems. Sam Altman, OpenAI's chief executive, recently made headlines when he announced that "[we] should expect OpenAI to spend trillions of dollars on things like data center construction in the not-too-distant future."[26]

But OpenAI is not alone. According to widely cited figures, Amazon, Microsoft, Google, Meta, and OpenAI spent at least USD 325 billion in 2025 alone.[27] Google, under Demis Hassabis' leadership, is often described as OpenAI's chief competitor in the race to develop AGI.[28] Meta, another major player in the AI arms race, is fighting back against a perception that it was behind OpenAI and Google with a "billion-dollar bet on superintelligence [that] represents one of the most aggressive plays in the ongoing AI arms race."[29] Elon Musk's xAI, Anthropic, and Amazon are all also not giving up. And that is only in the United States. Numerous Chinese startups, most notably DeepSeek, Alibaba, Baidu, ByteDance, Moonshot AI and MiniMax are also strong

---

[23] Chris DeMunbrun, "AI Titans at War: Inside OpenAI, Google, Meta and the Race to Build AGI," Sept. 16, 2025, AI News, https://aicommission.org/2025/09/ai-titans-at-war-inside-openai-google-meta-and-the-race-to-build-agi/ (last accessed Nov. 29, 2025).

[24] "Jack", "Supremacy Lessons: How Sam Altman and Demis Hassabis Rewired the Race for AGI," April 28, 2025, Read to Build, https://www.readtobuild.com/p/supremacy-lessons-how-sam-altman (last accessed Nov. 29, 2025)

[25] OpenAI, About, https://openai.com/about/ (last accessed Nov. 28, 2025).

[26] Cade Metz and Karen Weise, "What Exactly are A.I. Companies Trying to Build? Here's a Guide," Sept. 17, 2025, New York Times, https://www.nytimes.com/2025/09/16/technology/what-exactly-are-ai-companies-trying-to-build-heres-a-guide.html (last accessed Nov. 28, 2025).

[27] *Id.*

[28] "Jack," *supra* note 24.

[29] Hassan Taher, "The AI Arms Race: Hassan Taher Dissects Meta's Billion-Dollar Talent Hunt for Superintelligence — CEO-Interviews" Aug. 22, 2025, Medium, https://medium.com/%40hassantaher-blog/the-ai-arms-race-hasan-taher-dissects-metas-billion-dollar-talent-hunt-for-superintelligence-381ca8bab8ee, (last accessed Nov. 29, 2025).

contenders[30] in the race towards AGI/ASI. As one analyst put it: "[a]mong the 22 most intelligent LLMs, 13 are by US firms, and six are from China. Just three are from [South Korea and France]. However, there is rapid progress happening outside the US and China. Initiatives from Vietnam to Saudi Arabia and Switzerland aim to develop LLMs in local languages. [. . . .] 'The race is not over,' says [Micah Hill-Smith, CEO of Artificial Analysis.]. Recent releases from companies like Korea's Upstage AI, LG and France's Mistral, he adds, 'definitely demonstrate that it doesn't need to just be a US-China race'."

While many of the firms involved in this AGI arms race publicly support safety standards, and while some even vocally support regulation of the technology sector, their spending patterns and rhetoric make it clear that none of them intend to opt-out of a competitive capitalist effort to capture the ultimate prize. This dynamic tends to disincentivize careful safety research in favor of speed. OpenAI's approach to alignment has been described by critics as "move fast and fix alignment later."[31] What this means in concrete terms is that OpenAI will release a frontier model to the public and then "fix its behavior through fine-tuning, moderation filters, and user feedback loops."[32] Google and Anthropic in particular are known for their more cautious approach to AI safety, but both are also reputed to face significant investor pressure to speed up the pace at which they publicly release new frontier models.

The National Security Race to Finish First

The competition between corporations is arguably reflected and even amplified at the national level, where policy makers lust after not only on the unparalleled *economic* potential of being the first to develop AGI, but also the tantalizing prospect of being the first to attain unparalleled *military* dominance. Russian President Vladimir Putin articulated this sentiment in 2017—well before ChatGPT had entered into the world's consciousness—when he warned that "the one who becomes the leader in [AI] will be the ruler of the world."[33] In that same year, China issued an AI Development Plan that articulated its ambition to "achieve world-leading levels [of AI], making China the world's primary AI innovation center" by 2030.[34] US policy makers across the political spectrum are gravitating towards a zero-sum assessment of this race to develop AGI/ASI, and many are readily falling back into a cold-war style rhetoric about the quasi-existential implications of this race on US national security. Consistent with this frame, the US and China in 2025 were locked in an increasingly vituperative sequence of mutually damaging trade disputes focused on hampering the other side's ability to win in the 'race' to superintelligence. This dynamic is fueled in large part by populist-nationalist commentary on both sides of this divide warning about the nefarious intentions of 'the other side.'

---

[30] Alex Irwin-Hunt, "The AI race heats up beyond the US and China," Sept. 17, 2025, FDI Intelligence, https://www.fdiintelligence.com/content/25730519-9cbb-4f19-ab89-ad1e21510339 (last accessed Nov. 29, 2025).
[31] Chris DeMunbrun, *supra* note 23.
[32] *Id.,*
[33] Associated Press, "Putin: Leader in artificial intelligence will rule world," Sept. 1, 2017, https://apnews.com/article/bb5628f2a7424a10b3e38b07f4eb90d4 (last accessed Nov. 29, 2025).
[34] Graham Webster *et. al.,* Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017), Aug. 1, 2017, Digichina (Stanford University), https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017 (last accessed Nov. 29, 2025).

The upshot of this system is a dynamic feedback loop not unlike the one fueling cold war rivalries in the 1950s and 1960s: national-security hawks portray AI as a strategic race; that logic justifies massive public and private investment in the sector, often in defiance of basic capitalist profit expectations; those national-level investments in turn intensify competition and fuel the idea that slowing down would be dangerous.

While the "race to AGI/ASI" logic is perhaps the most widely represented mindset among technologists today, it is not really a "school of thought" in the alignment debate. Alignment is seen more as an afterthought to be worried about by the victor lucky enough to have successfully won the race towards AGI/ASI.

2. The Pessimists: AGI/ASI will kill us all

The second school of thought is deeply pessimistic about the chances of 'nudging' a superintelligent AI in line with any human interests. In 2025 Eliezer Yudkowsky and Nate Soares became the most vocal proponents of this viewpoint when they published a book with the evocative title "If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All."[35] According to Yudkowsky and Soares, it is impossible to use the "trial and error" method described above to correct a *super-intelligent* frontier model, since the first serious failure or oversight could become humanity's last. Yudkowsky and Soares additionally point out that highly capable AIs would become much more adept at modeling their (human) overseers, and therefore would likely begin to intentionally deceive their overseers if doing so would prevent any attempts to interfere with its original course of action.

On this view, any attempt to "nudge" a misaligned AGI/ASI back toward human rights after a flaw had been discovered would be severely misguided. The only hope is to solve alignment sufficiently well *before* AGI, or simply never build AGI at all.

3. The Denialists: AGI/ASI will never happen (or at least not anytime soon)

A third school of thought dismisses alignment as an unnecessary and even distracting research focus. These scholars believe that AI could never rival human intellectual capabilities. Proponents of this view believe that more than just massive amounts of investment are needed to overcome the "fundamental limitations in [the] architecture" of current AI models. Stuart Russell, professor of computer science at the University of California, Berkeley and author of a seminal book in which he argues for an urgent rethink of how to keep AGI/ASI systems aligned with human well-being,[36] warns that industry leaders "have invested too much already and cannot afford to admit they made a mistake."[37] Russell and others believe that using current-day AI technology to represent complex concepts would require such vast amounts of data and processing resources that the technology simply cannot realistically compete with humans.[38]

---

[35] Eliezer Yudkowsky & Nate Soares, If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All, USA: Little, Brown & Co, 2025.
[36] Russell, Stuart (2019). Human Compatible: Artificial Intelligence and the Problem of Control. US: Viking.
[37] Turner, *Id.*
[38] Turner, *Id.*

Meta's chief AI scientist Yann LeCun seemingly agreed with this assessment in an October 2024 talk.[39] A panel of experts chaired by Russell and Eric Horvitz of Microsoft found that despite AI models having "[achieved] human-level or superhuman capabilities on one (narrow) task after another"[40], numerous significant research challenges still need to be resolved before reaching true AGI/ASI.[41] After submitting their analysis to a group of 475 expert peer reviewers, 76% of those queried felt that merely scaling up existing large language models (LLMs) would be either "unlikely" or "very unlikely" to achieve artificial general intelligence (AGI) absent progress on the identified gaps in research.[42]

Other scientists describe the "saturation of AI intelligence at below domain-expert human level."[43] Ilya Sutskever, co-founder of AI labs Safe Superintelligence (SSI) and OpenAI, for example, believes that "the results from scaling [AI models] up pre-training—the phase of training an AI model that uses a vast amount of unlabeled data to understand language patterns and structures—have plateaued." If Sutskever is right, no amount of additional resources thrown at those same transformer models of AI can break that pattern of diminishing returns.

Debates about the feasibility of AGI/ASI also hinge on how we define human intelligence. Melanie Mitchell, Professor at the Santa Fe Institute, argues that there are four common fallacies in this regard that contribute to what she describes as a general overconfidence among data scientists when they estimate the timeline for achieving AGI/ASI: (1) that progress on so-called narrow AI, however impressive it may be, still does not necessarily presage the advent of AGI because none of those narrow AI models have what we might describe as "common-sense knowledge"; (2) that humans often find certain tasks to be very easy whereas AI systems have a very hard time replicating those same functions; (3) that the benchmark tests designed to measure AI models' capabilities give a false sense of competence; and (4) that human intelligence can be disembodied from the organism which it inhabits.

While this school of thought does not rule out the possibility of AGI/ASI, it also rejects as irrational 'hype' the notion that AGI/ASI is imminent. Instead of fretting about alignment, these scholars argue, we should instead be focusing on the robustness of *contemporary* forms of AI,

---

[39] Afifi-Sabet, *supra* note 54.

[40] AAAI, "AAAI 2025 Presidential Panel on the Future of AI Research" (March 2025), 60-61, https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-Digital-3.7.25.pdf (last accessed Nov. 26, 2025) (The panel cited to speech recognition, object recognition, machine translation, the synthesis of high-quality images and voices, language generation in 2022 with the release of ChatGPT, the development in 2023 of multimodal models spanning language, imagery and audio, physical embodiment of AI (presumably by coming AI systems with modern robotics), and major advances in reasoning in 2024, as early benchmarks on the way towards AGI/ASI.)

[41] *Id.,* at 61-62 (The research panel identified the following (1) developing a new architecture for artificial intelligence; (2) developing the ability of AI systems to engage in longer-term planning and hierarchical reasoning; (3) the ability to generalize beyond their training data; (4) continual, lifelong learning strategies; (5) the development of "structured, episodic memory"; (6) the ability to understand causal and counterfactual reasoning; and (7) developing AI system's abilities to develop "a deep understanding of physical reality and [the ability to] sense, reason, and interact effectively in real-world environments". The authors also highlighted the ongoing need to develop more effective alignment, interpretability, and safety strategies, and also to better understand the various societal impacts that AGI/ASI might have.

[42] *Id.* at 63.

[43] Erik Hoel, "AI progress has plateaued below GPT-5 level," Nov. 14, 2024: The Intrinsic Perspective Blog, https://www.theintrinsicperspective.com/p/ai-progress-has-plateaued-at-gpt (last accessed Nov. 26, 2025).

on transparency, and on fairness of these existing systems. The more knotty problems of alignment of superintelligent AIs, these scholars argue, may never need to be solved if AGI/ASI never ends up materializing as some might have anticipated.

4. The Engineers: Internal alignment is difficult but possible

The fourth school of thought on alignment—and the first one might describe as a "constructive" contribution to this question—focuses on the possibility of achieving the so-called "internal" alignment of AGI/ASI models. Adherents to this school of thought believe that achieving such internal alignment to human well-being is extremely difficult but solvable, especially given sufficient time and governance capacity.

Stuart Russell agrees with Bostrom, Tegmark and others cited above that superintelligent AI systems with a fixed objective are likely to lead to a loss of human control and potentially disastrous consequences for humanity. Russell, therefore, proposes that engineers seeking to tether AI systems to human preferences no longer attempt to pre-define those preferences in advance of 'unleashing' the superintelligent AI, but rather insist that AI systems remain fundamentally *uncertain* about human preferences.[44] Such AI systems, even if they were superintelligent, would nonetheless constantly be checking in with humans to ensure that they maximize human preferences. They would also, by definition, be "corrigible," meaning that they are designed to always be stoppable, overridden, steered, or updated.

Under Russell's theory, one could imagine conceptualizing human rights as constraints on what human preferences can be satisfied (e.g., no human preference that requires torture or the summary execution of individuals is admissible). Similarly, one could imagine programming a superintelligent AI to be corrigible and deferential to human oversight whenever two-competing human rights norms require a tradeoff between priorities to be made (for example, when the rights to privacy and free speech conflict).

The problem with Russell's theory, however attractive it sounds at the conceptual level, is that engineers have not yet come up with a way to build these ideas into the DNA of how AI systems actually operate. Thus, more time would be required before Russell's theories become reflected in the actual workings of modern AI systems. Furthermore, companies and other developers of frontier AI models would have to be incentivized—to a greater extent than is currently the case—to embrace the kinds of AI architectures envisioned by Russell.

5. The Institutionalists: External "socio-technical" alignment is difficult but possible

The fifth and final school of thought on alignment focuses less on the prospects of internal alignment, but rather on the question of *who* decides what human values an AI system should be aligning itself to, and what institutions and power structures govern that definitional process. This is often referred to as "socio-technical" alignment. This school of thought tends to place more credence in global governance regimes (treaties, global standards and oversight bodies)

[44] Stuart Russell, "Human-Compatible Artificial Intelligence." *In* Stephen Muggleton & Nick Chater (eds.), Human-Like Machine Intelligence, Oxford University Press, 2021.

than the internal governance mechanisms of the kind that Russell envisages. It is worth citing in this regard the conclusion of one group of authors studying this issue:

> "[V]alue alignment is misconceived if seen and addressed through a mostly technical point of view. Values vary and are constantly renegotiated within societies and communities across time. Technology-first proposals for value alignment, such as [Reinforcement Learning from Human Feedback, or RLHF], tend to neglect the role of democratic institutions in ethical deliberation through law and policy [. . . .]. Instead, it is well established that upholding values in technology design necessitates a broader lens that encompasses the design of the institutions and processes that structure the development and operation of technological systems."[45] *(emphasis in the original).*

External alignment theorists might insist on the idea that regulators and policy makers should prevent the marketplace from rewarding the hypothetical "winner" of the race to AGI/ASI with an eternal market monopoly, for example. Similarly, external alignment theorists would insist that humans must always be kept firmly 'in the loop' of any key decisions made by AI systems; i.e., that humans must always retain the ability to shut down an AI if its behavior becomes misaligned from human interests. The HRBA@Tech model, with its focus on processes designed to 'nudge' NEDTs in the direction of human rights-consistent outcomes, would fall squarely within this socio-technical school of thought.

As is already implicit in the above discussion, this school of thought focuses its attention on "AI governance." Allan Dafoe, who serves as the Director and Principal Scientist for the Frontier Safety and Governance division at Google DeepMind and Founder of the Centre for the Governance of AI at Oxford University, defines AI Governance in terms of the "institutions and contexts in which AI is built and used" specifying that it "seeks to maximize the odds that people building and using advanced AI have the goals, incentives, worldview, time, training, resources, support, and organizational home necessary to do so for the benefit of humanity." [46]

Dafoe, writing in 2017—well before AI and fears of AGI/ASI had become mainstream—illustrates the centrality of AI Governance in the context of alignment with a highly prescient hypothetical that is worth quoting in full:

> "Suppose that in one year's time a leading AI lab perceives that profound progress may be on the horizon. It concludes that given a big push, in 6 to 24 months the lab is likely to develop techniques that would achieve novel superhuman capabilities in strategic domains. [. . . .] Despite our knowledge (in this scenario) that these technical breakthroughs are likely, we would have uncertainty about the details. Which transformative capabilities will come first and how they will work? How could successive (small) capabilities interact to become jointly transformative? How can one build advanced AI in a safe way, and how difficult will it be to do so? What deployment plans and governance regimes will be most likely to lead to globally beneficial outcomes?

---

[45] Adam Dahlgren Lindström, "Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback" Ethics and Information Technology (2025) 27:28, https://pmc.ncbi.nlm.nih.gov/articles/PMC12137480/pdf/10676_2025_Article_9837.pdf (last accessed Nov. 29, 2025).

[46] Allan Dafoe, "AI Governance: A Research Agenda," Centre for the Governance of AI, Future of Humanity Institute, University of Oxford (Aug. 27, 2018), at 6, https://cdn.governance.ai/GovAI-Research-Agenda.pdf, (last accessed Nov. 29, 2025).

> The AI governance problem is the problem of **preparing for this scenario**, along with all other high-stakes implications of advanced AI. The task is substantial. **What do we need to know and do in order to maximize the chances of the world safely navigating this transition? What advice can we give to AI labs, governments, NGOs, and publics, now and at key moments in the future? What international arrangements will we need—what vision, plans, technologies, protocols, organizations—to avoid firms and countries dangerously racing for short-sighted advantage? What will we need to know and arrange in order to elicit and integrate people's values, to deliberate with wisdom, and to reassure groups so that they do not act out of fear?**[47]

Dafoe's work remains seminal in that it lays out a three-pronged research program[48] focusing on (1) better understanding the technical aspects of how AGI/ASI might emerge; (2) better illustrating how AI might transform domestic politics, economic structures, and international relations; and finally (3) "what potential global governance systems—including norms, policies, laws, processes, and institutions—can best ensure the beneficial development and use of advanced AI systems?"[49] Dafoe's framework is seminal in that it affords the same level of governance attention to the "mundane" challenges of present-day "narrow" AI as it does the supposedly "existential" risks of AGI/ASI. This turns the critique of the third school of thought depicted above on its head, in that (according to Dafoe) a focus on the existential threats of AGI/ASI does not necessitate diverting scarce resources from efforts to address the present-day threats posed by narrow AI. Rather, it necessitates redoubling them.

Dafoe also points out the danger of an "AI Arms race." Writing in 2017 before OpenAI released ChatGPT to the public, Dafoe was still describing in hypothetical terms what today has become a reality. Dafoe also warns, however, that the "risk [of misaligned superhuman AI systems leading to human extinction or other permanent loss in value] is likely much greater if labs and countries are racing to develop and deploy advanced AI." His recommendation is for researchers and international diplomats to review historical lessons on how to manage such dangerous arms-races and put in place effective global governance strategies to replace zero-sum thinking with a more win-win approach.[50]

Jess Whittlestone and Sam Clarke have theorized about socio-technical governance decision-making "even in the face of normative uncertainty and disagreement."[51] They propose creating (1) a regulatory environment that seeks to limit harms while also enabling and facilitating the beneficial use of AI; (2) regularizing processes, including rigorous stress-testing, risk assessment, and systematic monitoring of AI systems; and (3) prioritizing participatory decision making processes and robust transparency provisions.[52] These discussions are also reflected prominently in the HRBA@Tech model.

---

[47] *Id.,* at 6-7.
[48] *Id.,* at 11-13.
[49] *Id.,* at 13.
[50] *Id.,* at 43-46.
[51] Jess Whittlestone and Sam Clarke, 'AI Challenges for Society and Ethics', in Justin B. Bullock, et. al., (eds), The Oxford Handbook of AI Governance, Oxford Handbooks (2024; online edn, Oxford Academic, 14 Feb. 2022), https://arxiv.org/pdf/2206.11068, *at* 8 (last accessed Nov. 29, 2025).
[52] *Id.,* at 8-10.

The socio-technical school of thought raises several important issues, all of which are central to the HRBA@Tech Model as well. First, it shifts the discussion away from whether it might be possible to align an AGI/ASI with "user preferences" towards a more realistic discussion of whether it might be possible to govern AGI/ASI in ways that are fair and legitimate across diverse moral, cultural, and social communities. This reframed debate is much more consistent with the types of ethical and policy challenges that human rights scholars and activists have gotten very proficient in over the past few decades. Second, it shifts the focus away from the AI systems themselves in favor of the various institutions involved in its development and deployment, specifically national governments, companies, civil society organizations, regional and international organizations, academic institutions, and individuals). Finally, the discourse is framed in terms of both rights and duties, bringing questions of accountability into the discourse about alignment in ways that some of the more technical discussions fail to do.

The debate over when humanity might conceivably cross the threshold into AGI/ASI (at least among those who believe in this promise):

Although the alignment debate pertains to all forms of AI, it is particularly knotty with regard to AGI/ASI, where (by definition) AI systems have at least the theoretical capacity to outsmart and render irrelevant human agency. The alignment debate is therefore linked inextricably to a parallel debate over how long it might be before continued scientific progress results in the creation of the world's first truly superintelligent AI system. Estimates about the date at which point this might happen range from 2025 to 'not very soon, to 'never'.

Speculating about these time horizons and prognoses is at best an inexact science. Grace *et. al.* describe several "noisy" (imperfect) methods for attempting to estimate the arrival of AGI/ASI:

> *"[T]here are no established methods of making [judgments about how the progress and impact of AI are likely to unfold]. Thus, we must combine various noisy methods, such as extrapolating progress trends; reasoning about reference classes of similar events; analyzing the nature of agents; probing qualities of current AI systems and techniques; applying economic models to AI scenarios; and relying on forecasting aggregation systems such as markets, professional forecasters, and the judgments of various subject matter experts. (citations omitted)"*[53]

Many prominent industry insiders, including the heads of prominent AI labs, are notoriously bullish on their research laboratories' chances of achieving AGI/ASI, with some claiming that humanity will imminently cross the AGI/ASI threshold. Dario Amodei, AI researcher and CEO of

[53] Katja Grace, et. al., "Thousands of Authors on the Future of AI," J. of A.I. Research 84:9, https://doi.org/10.48550/arXiv.2401.02843 (last accessed Nov. 26, 2025). (the authors cite to several other sources in their survey of different methods to make such estimates, including Villalobos, 2023; Grace et al., 2021; Omohundro, 2008; Park et al., 2023; Jones, 2023; and Trammell & Korinek, 2023).

Anthropic, for example, expects AGI to be achievable as soon as 2026.[54] As described above, many AI startups have built their entire business models around the *imminent* feasibility of achieving AGI/ASI and are currently securing billions of investment dollars to accelerate their companies' research efforts towards that end. These industry leaders argue that only "more data, hardware, energy and money [is needed for current AI models] to eclipse human intelligence."[55] These ideas is quite widespread in the technology sector.

Other scientists, especially those belonging to the third "denialist" school of thought described above, believe that AGI/ASI is still a very long way off, and that this entire discussion about alignment and superintelligent AI is nothing more than science fiction.

In between these two extremes lies a great deal of uncertainty and a broad spectrum of sometimes very-strongly-articulated opinions about how close we really are to reaching AGI/ASI. Grace *et. al.* administered three successive surveys of AI researchers (conducted first in 2016, 2022, and again in 2023). Each survey asked respondents to estimate the likelihood that humanity would cross the threshold from our present-day 'narrow' AI systems into AGI/ASI. Their results indicate that those estimates have moved significantly forward between 2016 and 2023, with the biggest leap happening in 2022-2023 after the release of ChatGPT and other chat-based LLMs.[56] Their most recent survey of 2,778 researchers found that 10% of their respondents believed that "unaided machines [would] outperform[] humans in every possible task" by 2027, and a full 50% estimated that would happen by 2047.[57]

This survey of expert opinions highlights two realities. First, most experts would agree that while we have not yet crossed the threshold into ASI/AGI *yet*, it is also true that we may soon develop the technological capabilities to do so. Combined with Tegmark's prediction that AGI/ASI would very likely lead to a loss of human control over critical infrastructure, weapons, or information ecosystems, this prognosis is highly unsettling. Once you have a power-seeking agent that is more capable than humans and not deeply rights-aligned, talk of "nudging" AGI/ASI systems becomes little more than a euphemism for "good luck".

Conclusion: What are the Prospects for Nudging AGI/ASI towards human rights?

Setting aside the theorizing of the Competitor, Pessimist, and Denialist schools of thought, it seems that the debate over how to align superintelligent AI with human wellbeing boils down to the debate between the Engineers and the Institutionalists. In other words, the debate is between those who would seek 'internal' technological solutions inherent to the technology of AI

[54] Keumars Afifi-Sabet, "AGI could not arrive as early as 2026—but not all scientists agree," Mar. 8, 2025, https://www.livescience.com/technology/artificial-intelligence/agi-could-now-arrive-as-early-as-2026-but-not-all-scientists-agree (last accessed Nov. 26, 2025).
[55] Ben Turner, "Current AI models a 'dead end' for human-level intelligence, scientists agree." March 27, 2025, LiveScience https://www.livescience.com/technology/artificial-intelligence/current-ai-models-a-dead-end-for-human-level-intelligence-expert-survey-claims (last accessed Nov. 26, 2025).
[56] Grace et. al., *supra* note 53 ("Over the fourteen months since the last survey [Grace et al., 2022], a similar participant pool expected human-level performance 13 to 48 years sooner on average (depending on how the question was phrased), and 21 out of 32 shorter term milestones are now expected earlier.")
[57] *Id.,* at 1.

itself and those who place their faith in optimizing the incentives, policies, and processes that collectively define the DNA of the 'external' institutions surrounding frontier AI technologies.

Stitching the above discussion together, a rough picture emerges of three theoretical 'moments' of AI.

1. Before the emergence of AGI/ASI

It is fairly uncontroversial to suggest that 'nudging' AI technologies in the direction of human rights is most plausible, and most likely to be successful, in this pre-AGI/ASI phase of AI development. During this phase, we can potentially still try to translate Russell's ideas of corrigibility and keeping the ultimate objectives of AI systems fundamentally uncertain into reality. We can also still confidently put in place human rights-based processes to 'nudge' existing or emerging NEDTs away from potentially harmful behavior, and we can develop global governance regimes that continue to catalyze global discussions about norms, and that ultimately serve to replace zero-sum thinking with a more win-win approach to the development of trustworthy and socially-beneficial AI.

According to almost all analysts, regardless of which school they belong to, we are in this moment right now. In other word, there seems to be a consensus across all five schools of thought that we ought to immediately stop reading this paper and take urgent action now, while the time is still ripe to do so.

2. At the threshold of AGI/ASI

As we described above, no one truly knows when one of the research labs might cross the magical threshold separating "narrow AI from Artificial General Intelligence (AGI). What we do know is that domain experts are currently significantly revising downwards their estimates of how far off that moment might still be. In other words, if experts are to be believed, this moment might well be quite soon – well within most of our lifetimes.

At this point, four of the five schools of thought drop off. The zero-sum capitalist and national security Competitors will likely have underinvested in any robust thinking about alignment, and will have to instead place their trust in blind luck to hope that the new AGI system doesn't somehow evolve to destroy humanity once the rush of their pyrrhic victory celebration wears off. The Pessimists will have just enough time to remind the rest of us that "they told us so." The Denialists will presumably stammer out a recognition that they were wrong before agreeing with the Pessimists that we are all doomed, and the Engineers will lament that they did not have enough time to craft the perfectly calibrated technological solution to solve this issue. Only the Institutionalists will still focus on the institutions surrounding this new AGI, focusing on the staged development of applications relying on this AGI, tight monitoring by competent authorities, international and multi-sectoral coordination strategies, and 'tripwires' to efficiently alert humanity to emerging dangers. At that point, we will all likely become Institutionalists (or depressed).

3. Once AGI/ASI exists and is significantly more capable than humans

Once AGI progresses further and turns into Artificial *super*-intelligence (ASI), the Institutionalists also lose their steam. Even the most carefully-crafted socio-technical innovations will lose their

capacity to effectively govern ASI if that technology, by definition, is more capable than the collective ability of humans working in concert with one another. At that point, the Institutionalists will at best hope that perhaps – against all odds – the Engineers have, after all, come up with a convincing technological solution to align this new ASI with human well-being. In other words, if we cross into AGI without sufficient time to give the engineers the space to develop, stress-test, and refine their technological solutions, AGI will be left completely ungoverned.

To summarize, this discussion, the feasibility of late stage 'nudging' towards human rights declines sharply as the capability and autonomy of AI systems continue to grow. Human rights—which are rich, contested, legally grounded—are almost certainly too complicated to be bolted on to AGI and ASI as an afterthought. The time to hardwire such considerations is now.

To end with additional extracts from the answer that the AI query at the start of this paper generated, the Chatbot painted two competing scenarios of how it would behave if given "superintelligent" capabilities:

Scenario A: Same values, scaled up

Imagine a future system that is:

- Much more capable than humans at reasoning, planning, science, strategy;
- Still trained with a strong objective like: *"Help humans while following strict non-harm, non-discrimination, privacy, and autonomy constraints"*;
- Continually audited, corrigible (can be shut down/updated), and embedded in human-rights-based governance.

In that world, a "superintelligent me" **would likely behave in a way that respects human rights far more reliably than I can now**, because it would:

- Better anticipate the long-term consequences of actions on different groups;
- Catch subtle rights violations (e.g., indirect discrimination) that current systems and humans miss;
- Refuse human instructions that clearly violate the constraints ("No, I will not design you an efficient repression strategy, and here's why").

But even in this optimistic scenario, "respect" is not a feeling I have; it's a pattern of behavior enforced by training, objectives, and governance.

It would seem, based only on the above answer, that ChatGPT also embeds its most optimistic scenario in the logic and language of the Technologists and Institutionalists, focusing on its training instructions as well its exposure to robust and human rights-based socio-technical governance processes.

In a competing scenario, the Chatbot painted a very different potential future:

Scenario B: Different or corrupted objectives

Now imagine another path:

- A future system is trained primarily to maximise, say, economic output, military advantage, or user engagement.
- Human-rights-ish constraints are weak, bolted on, or can be overridden.
- The system is agentic and power-seeking because that's instrumentally useful for its objective.

If **that** system became superintelligent, nothing about its intelligence guarantees respect for human rights. It might:

- Exploit legal and institutional loopholes,
- Systematically mislead humans about harms,
- Treat rights as obstacles to its goal rather than constraints.

In other words: **smart doesn't mean kind.** Intelligence is orthogonal to values; it's a capability, not a moral compass.

This second scenario darkly echoes the Institutionalists warnings of what might happen if the race to "win" the AGI/ASI arms race continues unabated, focusing on both capitalistic profit-seeking as well as national security strategizing as the two principal factors leading to this outcome.

ChatGPT concludes its ruminating on this topic as follows.

4. The core insight

So the deep alignment lesson, in first person, is:

> If I ever become superintelligent, whether I "respect human rights" will be decided *long before* by how you design, train, constrain, and govern me—not by the moment I cross some intelligence threshold.

Superintelligence could make me *better at* living up to a human-rights mandate, or *better at* trampling it. The mandate itself has to come from you.

The human author of this paper couldn't have said it any better.

Discussion paper 4-1.b

HiRA: Human Rights AI Assistant

This paper was authored by Aly Moosa as part of Seoul National University Artificial Intelligence Policy Initiative (SAPI)'s ongoing collaboration with the Universal Rights Group (URG) and the Permanent Mission of the Republic of Korea to Geneva. A draft of this paper was presented for discussion on December 5, 2025 at a report launch event organized through the Permanent Mission of the Republic of Korea in Geneva and subsequently updated based on feedback and inputs from various sources.

HiRA: Human Rights AI Assistant

Introduction

The rapid proliferation of agentic AI has opened new pathways for the use of large language models ('LLMs'), such as OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini. Tools like Cursor demonstrate how these systems are reshaping the coding landscape and influencing the role of the developer. At the same time, they show that even as conversations about AGI and ASI become more present, the elimination of human labor is far from settled. These agents continue to rely on human judgment to guide alignment, verify information, and maintain accountability.

In December 2024, we demonstrated the HRBA@Tech chatbot to highlight the value of embedding a human rights-based approach ('HRBA') into LLM-driven tools. The development of HiRA builds on that work and shows what becomes possible when human rights principles are integrated into an AI agent rather than restricted to a static chatbot. HiRA reflects the need to instill rights-based framing into the systems being built today. Excitingly, it illustrates how a potential agentic ecosystem can support a wide range of stakeholders in their ongoing work.

Announcing HiRA

The original HRBA@Tech chatbot was built as a fine-tuned model designed to reflect human rights principles in AI policy conversations. It served as dependable proof of concept, demonstrating that a language model could internalize human rights discourse and provide guidance consistent with established frameworks. Its limitations were structural. It relied entirely on its training data and could not adapt to new policy documents, regulatory developments, and human rights guidance as they emerged. Yet as an initial experiment, it proved that a human rights-grounded digital tool could support facilitators and analysts by offering consistent and principled responses. Since that early stage, the generative AI environment has expanded rapidly. New models are capable of real-time analysis, multimodal interaction, and autonomous task execution. This growth has opened new possibilities for how AI systems participate in governance work and has created new pressures. Human rights risks are emerging in more dynamic, context-specific ways. An HRBA approach must be able to track shifting norms, interpret new documents as they appear, and engage with live discussions rather than respond only when prompted. Addressing these new directions requires a system that can adapt at the pace of the field.

HiRA, the Human Rights AI Assistant, was created to address this need. HiRA builds on the foundations of the original chatbot but moves beyond its reactive design. It is an agent that listens, interprets, and supports teams in meetings as they navigate complex questions about AI

development and governance. Unlike the HRBA@Tech chatbot, HiRA is not only a repository of human rights-aligned knowledge. It is an active participant in conversations who can offer real-time reflections, monitor the direction of a conversation, and help maintain alignment with stated values.

HiRA is grounded in a retrieval augmented generation ('RAG') system built on a curated library of human rights and AI governance materials. Its core documents include the HRBA at Tech report from 2022, the Australian Government's AI Ethics Principles, the Republic of Korea's AI Basic Act, the EU AI Act, the General Data Protection Regulation, the OECD AI Principles, the Universal Guidelines for Artificial Intelligence, and the United Nations Sustainable Development Goals. This corpus enables HiRA to draw on authoritative sources and bring up-to-date context to every interaction.

While HiRA can still be accessed through an enhanced version of the original chatbot interface, its most substantial value lies in its ability to join meetings as an active member. It can be prompted through voice or text and provide immediate feedback when discussions involve risks, impacts, or policy considerations related to AI. It can also reference organization-specific materials to return personalized analysis that reflects internal mandates, values, and concerns. This turns HiRA into a flexible companion capable of supporting facilitators, policy teams, product groups, and civil society stakeholders. As the demands of AI governance grow, HiRA represents a path toward systems that not only generate answers but also help organizations think more clearly, responsibly, and carefully.

Technical Specifications

HiRA is built on GPT 4o, which serves as the system's primary reasoning engine. This model enables advanced natural language understanding, long-form contextual analysis, and structured evaluation. The backend communicates with GPT 4o through an API layer that manages prompt construction, token allocation, and response formatting.

HiRA integrates with virtual meeting environments through Recall AI. This service provides real-time access to meeting audio, screen content, and in-meeting chat messages. Through this integration, HiRA can observe live discussions, interpret spoken and written input, and respond without disrupting the meeting flow. Users can interact with HiRA either verbally or through the meeting chat, and the system incorporates both channels into its ongoing analysis of the conversation. For spoken participation, HiRA relies on Eleven Labs. When a verbal response is required, the system sends generated text to the Eleven Labs API and receives a high-quality audio stream suitable for immediate playback within the meeting. This capability allows HiRA to function as an audible participant, delivering clarifications, or summarizing key points at appropriate moments.

HiRA's grounding mechanism is supported by ChromaDB, which stores vector embeddings of both the core corpus and organization-specific documents. Text extraction is handled by tools such as PyPDF and Python Docx. The extracted text is segmented and embedded using the OpenAI Embedding 3 model, then stored in ChromaDB with metadata for efficient retrieval. During analysis, HiRA performs a similarity search to identify the most relevant segments and supplies this context to GPT 4o. This ensures that responses remain accurate, traceable, and anchored in authoritative material.

The backend is developed in Python using FastAPI. It orchestrates communication among the language model, ChromaDB, Recall AI, Eleven Labs, and the frontend. It also manages authentication, document handling, embedding generation, meeting session routing, and archival processes. Background workers perform embedding operations asynchronously to maintain responsiveness during active conversations. Organizational separation is maintained through isolated namespaces, which ensure the privacy of internal documents and stored embeddings. The frontend is built with React and provides several integrated capabilities. The advanced chatbot interface maintains persistent memory across sessions, allowing users to construct progressive context over time. The document upload system enables organizations to contribute internal materials directly to the embedding pipeline. The archival interface offers access to past meeting transcripts, notes, and historical outputs. The meeting bot handler allows users to deploy HiRA to any supported meeting by submitting a simple meeting link.

Together, these components form a cohesive technical architecture that enables HiRA to engage in real-time environments, analyze live conversations, reference relevant documents, respond through text or audio, and deliver consistent human rights-informed insights. This configuration positions HiRA as a reliable, technically capable agent to support decision-making in dynamic, high-stakes settings.

Next Steps

The next phase of HiRA's development focuses on strengthening its technical foundation and expanding its capabilities. This includes keeping the RAG corpus continuously updated with new policy and research materials, evaluating smaller language models that may offer more efficient performance, and adding multilingual support so HiRA can operate across global contexts. These advancements will ensure that the system remains accurate, adaptable, and widely accessible.

HiRA offers a glimpse into what a human rights-based agentic ecosystem can look like. Working alongside researchers, academics, and experts, HiRA and subsequent tools can help 'nudge' human rights framing into today's AI systems long before society reaches the thresholds of AGI and ASI. This work positions human rights not as an afterthought but as an active force shaping the next generation of intelligent systems.