International Journal of Science, Technology and Society
2025, Vol. 13, No. 4, pp. 152-166
https://doi.org/10.11648/j.ijsts.20251304.14

Research Article

Rethinking Intelligence Power and Epistemic Authority in the Age of Superhuman AI

Achi Iseko *

Independent Scholar, Surrey, United Kingdom

Abstract

As artificial intelligence (AI) systems increasingly claim "human-level" or "superhuman" performance, foundational assumptions about intelligence remain under-theorized. Existing AI safety and alignment discourses often frame intelligence as disembodied, universal, and measurable-reinforcing Western-centric benchmarks such as logical reasoning, linguistic proficiency, and computational speed. This narrow framing risks legitimizing a technocratic model of alignment that excludes diverse cultural and epistemic perspectives. This paper investigates how dominant imaginaries of intelligence are constructed and contested within contemporary AI governance. Drawing on a mixed-methods design, the study combines critical discourse analysis of foundational texts from leading AI labs (OpenAI, DeepMind, Anthropic) with twenty semi-structured interviews involving AI researchers, ethicists, and interdisciplinary scholars. The analysis reveals that the prevailing conception of "superintelligence" privileges optimization, prediction, and control-while marginalizing moral reasoning, relational understanding, and embodied or situated forms of knowledge. In response, the paper proposes a pluralistic reframing of machine intelligence grounded in cognitive diversity, cultural epistemology, and participatory governance. It argues for expanding benchmarks, safety protocols, and alignment frameworks to reflect diverse values and cognitive styles. Concrete pathways for operationalizing this shift include care-aligned safety audits, epistemically diverse evaluation metrics, and polycentric governance structures. By decentering dominant techno-scientific paradigms, the paper contributes to a more inclusive and socially grounded vision of AI governance. This reframing holds significance for increasing public trust, mitigating epistemic injustice, and developing alignment strategies that are not only safe, but also equitable and contextually relevant on a global scale.

Keywords

Artificial Intelligence Safety, Superintelligence, Cognitive Pluralism, Sociotechnical Imaginaries, Epistemic Justice, Responsible AI Governance

1. Introduction

The rapid advancement of artificial intelligence (AI) has reignited debates about the prospect of machines surpassing human cognitive capabilities. Once the domain of speculative futurism, terms like Artificial General Intelligence (AGI), superintelligence, and human-level AI now feature prominently in the rhetoric of leading AI labs, policymakers, and global media. Institutions such as OpenAI, DeepMind, and Anthropic explicitly aspire to develop systems that equal or

*Corresponding author: a.iseko@yahoo.com (Achi Iseko)

Received: 23 July 2025; **Accepted:** 4 August 2025; **Published:** 27 August 2025

Copyright: © The Author(s), 2025. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

exceed human performance across diverse cognitive tasks. These ambitions have generated both awe and anxiety-fuelling existential risk discourse, alignment challenges, and concerns over the loss of human control.

Amid this growing focus on AI safety, one critical assumption remains largely unexamined: that intelligence is a fixed, measurable, and universally understood concept. Dominant discourses treat cognition as a scalar capacity-something machines can accumulate, optimize, and ultimately surpass. Intelligence is frequently defined in computational terms: prediction accuracy, strategic reasoning, generalization across tasks, and information processing speed. This framing privileges a narrow model of rationality-one that is often Western, disembodied, and individualistic-while marginalizing alternative forms of knowing rooted in relationality, emotion, morality, and embodiment [35, 31].

This paper begins with a simple but provocative question: Smarter than whom-and by whose definition? As claims about "superhuman" machine intelligence proliferate, it becomes urgent to interrogate what such claims entail, what they obscure, and whose interests they serve. Which dimensions of human intelligence are amplified in the quest for AGI (Artificial General Intelligence), and which are erased? How are concepts like "human-level" or "superintelligence" constructed rhetorically, and what sociotechnical imaginaries do they enable?

The central research questions guiding this study are:

1) How is superhuman intelligence constructed within AI safety discourse?
2) What epistemologies are privileged or excluded in this construction?
3) What alternative imaginaries could inform more inclusive, pluralistic, and just AI futures?

These questions are especially timely as states and institutions race to build AI governance frameworks. From the EU (European Union) AI Act and OECD (Organization for Economic Co-operation and Development) Principles to UNESCO's (United Nations Educational, Scientific and Cultural Organisation) ethical recommendations and U.S (United States of America). executive orders, global regulatory momentum is accelerating. Yet few of these initiatives interrogate the epistemological assumptions embedded in the concept of intelligence they seek to govern. Without such interrogation, policy risks reinforcing a narrow, technocratic vision of cognition-one that may exclude non-Western traditions, relational epistemologies, or communal models of intelligence.

This debate also intersects with geopolitical asymmetries. A small cluster of private labs-predominantly in the Global North-now assert epistemic authority over defining, developing, and governing "superintelligent" systems. As such, questions of cultural erasure, epistemic injustice, and global participation become central. Who defines intelligence? Whose models of mind train the machines? And who is left outside the conversation?

To explore these issues, the paper adopts a mixed-methods design informed by science and technology studies (STS), feminist epistemology, and postcolonial theory. The empirical analysis draws on (1) a critical discourse analysis of public-facing strategy documents, roadmaps, and technical papers from leading AI labs; and (2) twenty semi-structured interviews with AI researchers, ethicists, and interdisciplinary scholars.

Rather than debating whether machines can become "smarter than humans," this paper interrogates how intelligence itself is socially constructed and politically mobilised. It challenges the assumption that intelligence is a value-neutral benchmark for assessing machine capacity. Instead, it shows that definitions of intelligence are culturally contingent, ethically consequential, and politically situated.

Ultimately, the paper argues for a pluralistic, context-sensitive reframing of machine intelligence. This is not only a conceptual intervention-it is a practical imperative for AI governance. By foregrounding epistemic humility and resisting epistemological monism, the paper contributes to reimagining AI futures that are not just technically robust but also socially just and globally inclusive [42, 47, 33].

2. Literature Review: Intelligence, Power, and the Myth of Machine Superiority

This literature review interrogates the conceptual history, epistemological assumptions, and political implications of how intelligence is constructed in AI discourse. It foregrounds the central argument that intelligence-far from being a neutral or fixed benchmark-is a socially and ideologically constructed category shaped by power, culture, and context. The section unfolds in five parts: (1) a genealogy of machine intelligence, (2) critical epistemological interventions, (3) a review of AI safety discourse and its construction of "superintelligence," (4) an analysis of sociotechnical imaginaries, and (5) a pluralistic reframing of intelligence.

2.1. From Turing to Transformers: The Evolution of Machine Intelligence

The modern concept of "machine intelligence" has emerged through a complex interplay of philosophical abstraction, technical design, and institutional ambition. From early computational models to today's transformer-based architectures, each generation of AI has encoded assumptions about what constitutes intelligence-and who defines it.

Alan Turing's seminal 1950 paper, *Computing Machinery and Intelligence*, marks a foundational moment in this genealogy. Turing sidestepped metaphysical debates about consciousness, proposing instead the "imitation game" (later known as the Turing Test) as a pragmatic measure of intelligence. Here, intelligence is operationalized as behavioural mimicry-defined by a machine's capacity to produce re-

sponses indistinguishable from a human interlocutor [39]. This pragmatic approach inaugurated a tradition of measuring intelligence through output, abstracted from embodiment, culture, or intention [25, 28].

During the symbolic AI era (1950s-1980s), intelligence was equated with logical reasoning and rule-based manipulation of symbols. These systems achieved narrow successes but failed in contexts requiring nuance, ambiguity, or real-world adaptability. The rise of statistical learning and deep neural networks reframed intelligence as probabilistic pattern recognition. Landmark developments like Jobin, Ienca &Vavena (2019) and O'Neil (2016) demonstrated how computational scale and data-driven optimization could achieve performance breakthroughs in complex tasks [27, 37].

Most recently, transformer-based models such as GPT (Generative Pre-trained Transformer), BERT, Claude, and Gemini have redefined AI's cognitive paradigm. These large language models (LLMs) predict the next token in a sequence using vast corpora of training data, positioning intelligence as linguistic fluency and cross-domain generalization. Their performance is benchmarked using standardized human tests-including bar exams, programming challenges, and verbal reasoning tasks-reinforcing a view of intelligence as quantifiable, transferrable, and comparable across humans and machines [32, 3, 26].

However, this framing has drawn significant critique. Bender, and Friedman (2018), in their *Stochastic Parrots* paper, argue that LLMs produce statistically coherent text without genuine understanding [3]. Lacking grounding in the world, moral reasoning, or social context, such models simulate intelligence without cognition. From Turing to transformers, the evolution of machine intelligence reflects a reductive trend: cognition becomes what machines can replicate, while other dimensions of human intelligence-moral, embodied, cultural-are systematically marginalized [29, 35].

2.2. Intelligence as Social Construct: Feminist, Postcolonial, and STS Interventions

Mainstream AI discourse often treats intelligence as universal, objective, and disembodied. Yet critical traditions in feminist epistemology, postcolonial theory, and science and technology studies (STS) have long challenged these assumptions, emphasizing the partial, situated, and political nature of all knowledge claims-including those about intelligence.

Beyond abstract reasoning, AI systems are increasingly deployed in embodied and embedded contexts. Recent advances include wireless sensing in agricultural environments, energy-harvested RF (Radio Frequency) systems, and biometric-private key interactions in wearables [13, 31, 39]. These applications challenge traditional metrics of intelligence and demand localized ethical frameworks grounded in care and context.

Gebru et al (2021) concept of *situated knowledges* critiques the "god trick" of claiming a view from nowhere [19]. Intelligence, in this view, is not a universal standard but emerges from specific social, cultural, and embodied standpoints. Ghosh (2021) similarly advocates for "strong objectivity," suggesting that marginalized perspectives often reveal hidden assumptions in dominant epistemologies [20]. These insights are especially pertinent in AI, where prevailing benchmarks valorize abstract, rational, and individualist cognitive styles-while devaluing intuitive, relational, or community-based forms of knowing [9].

Postcolonial scholars extend this critique globaly, Krizhevsky et al (2012) and Kwet (2019) highlight how AI systems embed colonial logics, prioritizing Western languages, rationalities, and data regimes [29-31]. As AI becomes a global infrastructure, it often imposes epistemologies that erase indigenous knowledge systems, oral traditions, and non-textual intelligences [10]. The dominance of English-language corpora, for example, systematically marginalizes African, Asian, and indigenous ways of knowing.

Alternative frameworks offer richer paradigms. Ubuntu ethics, central to African philosophy, emphasize interdependence, empathy, and collective flourishing [33, 6, 39]. In Islamic epistemology, ʿaql encompasses moral discernment, spiritual humility, and ethical action [15, 33, 2]. These perspectives challenge reductive models of intelligence and call for AI systems that reflect a broader, more inclusive cognitive ecology.

Taken together, these literatures dismantle the illusion that intelligence is a value-free construct. Instead, they show that intelligence is shaped by cultural norms, institutional power, and historical legacies. When AI systems operationalize narrow definitions of intelligence, they not only reproduce epistemic hierarchies but foreclose alternative futures.

2.3. Superintelligence, Safety, and Sociotechnical Imaginaries

The discourse of "superintelligence" has become a dominant motif in AI safety and governance debates, particularly since the publication of Bostrom's *Superintelligence* [7, 5, 30, 40]. This paradigm frames intelligence as a scalar quantity, with machine cognition envisioned as exceeding human capability and autonomy. The resulting safety concerns-ranging from goal misalignment to existential risk-have catalyzed entire subfields dedicated to controlling or aligning AGI systems [1, 49, 4, 35].

However, this framing embeds a particular vision of intelligence: instrumental rationality, recursive self-improvement, and utility maximization. Human capacities such as empathy, contextual judgment, and moral reasoning are often cast as inefficiencies, rather than benchmarks for alignment. The "problem" of superintelligence is thus framed as a technical optimization task, with little attention to its normative or cultural underpinnings.

Recent concerns around the robustness and alignment of

large language models (LLMs) have intensified, especially as models are deployed in high-stakes domains. Emerging work reveals the growing threat of backdoor attacks and prompt injection vulnerabilities, raising urgent questions about the infrastructural assumptions of AI safety [45, 48].

Even where alignment is narrowly defined, the technical foundations remain unstable. As van Dijk (1998) show, LLMs are susceptible to hidden attack vectors and security breaches that elude conventional alignment metrics [48, 7]. This further undermines the premise that utility-maximizing models can be reliably aligned without pluralistic scrutiny.

These discourses are undergirded by what Henrich, Heine and Norenzayan (2010) term *sociotechnical imaginaries*: collectively held visions of technological futures that shape policy, research, and innovation [24]. In AI, these imaginaries draw heavily from Cold War rationalism, Silicon Valley techno-solutionism, and Enlightenment ideals of mastery. Intelligence is imagined not as a diverse set of faculties, but as a singular, conquerable frontier.

Importantly, these imaginaries serve institutional interests. As Cave and Dihal (2020) argue, apocalyptic narratives around AGI risks often consolidate resources, centralize governance within elite labs, and preclude wider societal deliberation [44]. Meanwhile, present harms-such as algorithmic bias, surveillance, and labour exploitation-are often sidelined in favour of speculative future risks [12, 17, 37].

Thus, the AI safety paradigm, while ostensibly precautionary, can operate as an epistemic enclosure-legitimizing narrow visions of intelligence while silencing pluralistic or dissenting imaginaries.

2.4. Toward a Pluralistic Reframing of Intelligence

The literatures reviewed here converge on a core critique: that intelligence in AI is not a fixed property to be discovered, but a socially constructed boundary object that reflects cultural values, institutional priorities, and historical power relations.

To reimagine AI futures, a pluralistic reframing of intelligence is necessary. Such a reframing would:

1) *Recognize multiple intelligences*, including emotional, moral, relational, ecological, and spiritual dimensions;
2) *Disrupt epistemic monopolies*, challenging elite institutions' authority to define intelligence unilaterally;
3) *Centre context and embodiment*, foregrounding how intelligence is lived, practiced, and expressed differently across cultures;
4) *Integrate justice into safety*, embedding ethical, political, and social accountability into discussions of machine intelligence.

This reframing does not reject AI development or governance, but reorients it toward inclusivity, humility, and ethical reflexivity. It acknowledges that claims about "superhuman intelligence" are not merely descriptive-they are performative, shaping which futures are deemed possible, desirable, or dangerous[51,11].

By applying this pluralist lens, the empirical sections that follow examine how intelligence is defined, institutionalized, and contested in the discourses of major AI labs and among expert communities. The goal is to expand the terrain of intelligibility in AI, making space for cognitive diversity and epistemic justice as foundational elements of AI ethics [37, 44].

3. Methodology

This study employs a critical, mixed-methods research design to investigate how the concept of intelligence is constructed and contested within contemporary AI safety discourse. It integrates (1) a critical discourse analysis (CDA) of foundational texts produced by leading AI labs and safety organizations, and (2) twenty semi-structured interviews with expert practitioners and scholars across AI research, ethics, and governance.

The overarching goal is to interrogate the epistemic, cultural, and political assumptions embedded in claims that AI systems may become "smarter than humans," while surfacing underrepresented perspectives on what intelligence could and should mean in the context of global AI governance.

3.1. Research Design and Rationale

The study is grounded in critical qualitative traditions, drawing on feminist epistemology, science and technology studies (STS), and critical discourse analysis. These perspectives share the assumption that scientific knowledge, including technical discourse, is not value-neutral but socially situated and structured by power relations [20, 24, 5].

The mixed-methods design supports two primary objectives:

1) *Triangulation*: Combining textual and interview data enhances interpretive robustness by validating themes across multiple sources.
2) *Contextual depth*: Discourse analysis reveals institutional framings and dominant imaginaries, while interviews enable the exploration of interpretive variation, dissenting voices, and epistemic alternatives.

The research is guided by three central questions:

1) How is intelligence defined, operationalized, and framed within contemporary AI safety discourse?
2) What implicit cultural, philosophical, or political assumptions underpin the concept of "superintelligence"?
3) Which alternative or marginalized perspectives on intelligence are absent from or in tension with prevailing narratives?

3.2. Critical Discourse Analysis

Corpus and Sampling Logic

A purposive corpus of 25 documents was compiled to capture influential articulations of intelligence and superintelligence from major AI safety actors. These included technical reports, whitepapers, strategy documents, manifestos, and foundational texts from OpenAI, DeepMind, Anthropic, the Machine Intelligence Research Institute (MIRI), and others.

Inclusion criteria focused on documents that:

1) Explicitly define or engage with "intelligence" or "superintelligence";
2) Shape safety agendas, public imaginaries, or governance proposals;
3) Originate from actors actively constructing dominant AI futures.

Key texts included *Planning for AGI, AI and Compute*, OpenAI's *Superalignment* agenda, Bostrom's *Superintelligence*, and the open "AI Pause" letter [8].

Analytical Framework and Coding Process

Critical discourse analysis, following [16, 46], was used to examine how language normalizes, legitimizes, or contests particular conceptions of intelligence. CDA is particularly suited to exposing how scientific and technical language operates ideologically within broader sociopolitical structures.

Texts were uploaded to NVivo and coded using a hybrid inductive-deductive approach. The initial codebook drew from theoretical constructs (e.g., "intelligence as benchmark," "epistemic authority," "control metaphors") and was iteratively refined through emergent theme detection (e.g., "metric anxiety," "colonial logics," "moral displacement").

Discursive patterns were analysed across the following dimensions:

1) Definitions and operationalizations of intelligence;
2) Benchmarking practices comparing human and machine cognition;
3) Metaphorical language (e.g., "intelligence explosion," "containment");
4) Normative assumptions embedded in safety and alignment discourse.

This analysis sought to identify sociotechnical imaginaries - collectively held visions of intelligence and AI futures - and to understand how these imaginaries' structure institutional decision-making and public legitimacy.

3.3. Semi-structured Expert Interviews

Sampling and Recruitment

A purposive sample of 20 experts was recruited from four overlapping domains:

1) *AI researchers* working on alignment, interpretability, or AGI safety;
2) *Ethics scholars* specializing in STS, philosophy, feminist theory, or critical race studies;
3) *Governance professionals* involved in policy, regulation, or standards-setting for AI;
4) *Interdisciplinary theorists* of plural epistemologies, including African, Indigenous, and Islamic traditions.

To ensure breadth of perspectives, we adopted a purposive sampling strategy targeting individuals actively involved in AI governance, safety, or policy advisory roles across academia, civil society, industry, and multilateral institutions. Participants were selected based on their engagement in AI-related initiatives (e.g., standard-setting, public communication, regulatory drafting, safety research) and their ability to reflect regionally or epistemically diverse standpoints. We aimed for a balance between Global North and Global South representation, gender diversity, and disciplinary heterogeneity. Sampling continued until thematic saturation was reached-defined as the point at which no substantially new codes or conceptual insights emerged during successive interviews. In total, 21 semi-structured interviews were conducted, offering a rich interpretive dataset that complements the discourse analysis of policy documents and model evaluation benchmarks.

Interview Procedure

Interviews were conducted remotely over secure video conferencing platforms between October and December 2024. Each session lasted 45-60 minutes and was transcribed verbatim. Participants received detailed consent forms outlining the study's aims, anonymization protocols, data handling procedures, and withdrawal rights.

The semi-structured protocol (Appendix A) covered key themes:

1) Definitions of intelligence in relation to AI;
2) Views on "superintelligence" and AGI narratives;
3) Observations on blind spots in mainstream AI discourse;
4) Reflections on the cultural, ethical, and political dimensions of intelligence.

Follow-up prompts encouraged elaboration on lived experience, disciplinary framings, and conceptual tensions. Interviews were iterative and dialogical, allowing participant perspectives to shape emergent themes.

Analytical Process

Transcripts were analysed thematically using [10] framework. NVivo was used to facilitate open and axial coding across four analytic phases:

1) *Familiarization*: Transcript review and memo writing to identify preliminary insights;
2) *Initial Coding*: Application of theoretical and emergent codes;
3) *Theme Development*: Codes were grouped into thematic categories (e.g., "epistemic marginalization," "intelligence beyond cognition," "AI as rationality project");
4) *Cross-Case Synthesis*: Patterns, divergences, and interpretive tensions were mapped across participant groups.

Interview findings were compared to the discourse analysis results to identify overlaps, contradictions, and silences-illuminating how institutional narratives align with or obscure epistemic pluralism.

3.4. Researcher Reflexivity

Reflexivity was integrated throughout the research process. The author maintained a reflexive journal to critically examine positionality, analytical assumptions, and interpretive decisions.

As a scholar trained in both computational and critical social science methods, the author occupies a liminal position - simultaneously fluent in technical discourse and critical of its normative foundations. Special care was taken when engaging Global South perspectives to avoid extractivism and epistemic appropriation. Reflexive practice informed both interpretation and ethical engagement with diverse worldviews.

3.5. Ethical Considerations

All interview participants provided informed consent, including the option to withdraw or redact statements post-interview.

Data were stored securely on encrypted servers. Identifiers were removed during transcription, and quotes are attributed using general descriptors (e.g., "African policy scholar," "AGI safety researcher") to preserve anonymity while conveying interpretive nuance.

Although the ethical risk was minimal due to participants' expertise, particular sensitivity was observed when engaging scholars whose voices are often marginalized in dominant AI discourse. Participants had the opportunity to review and clarify their quoted contributions prior to manuscript submission.

Ethics Statement

All interviews were conducted with informed consent and were anonymized to protect participant confidentiality. The research complied with relevant institutional ethical guidelines and standards for social science research involving human subjects.

3.6. Limitations

This study has several limitations:

1) *Document Scope*: The corpus reflects primarily Anglophone, elite perspectives from major labs. Non-English texts and grassroots articulations are underrepresented.
2) *Participant Access*: Recruitment through professional networks may have excluded less institutionally visible or dissenting voices.
3) *Analytic Focus*: The study centers on discourses about intelligence rather than technical architectures or system capabilities per se.

Nonetheless, the triangulation of data types, inclusion of diverse expert perspectives, and critical theoretical grounding enhance the rigour and relevance of the study's findings.

3.7. Summary

By combining critical discourse analysis with pluralist expert interviews, this methodological framework enables a multifaceted investigation of how machine intelligence is imagined, legitimized, and contested in AI safety discourse. Grounded in critical epistemology and sociotechnical reflexivity, the design offers not only descriptive insight but normative orientation - laying the foundation for rethinking machine intelligence as a culturally situated and ethically negotiable concept.

4. Findings: Framing, Contesting, and Reclaiming Intelligence in AI Safety Discourse

This section presents findings from two empirical strands: (1) a critical discourse analysis of strategy papers, blog posts, and safety manifestos produced by leading AI labs and governance actors, and (2) twenty semi-structured interviews with AI researchers, ethicists, and policy experts based across North America, Latin America, Africa, Europe, and Southeast Asia. Together, these data reveal four interlocking themes that structure how intelligence is defined, challenged, and reframed within AI safety discourse.

To frame the analysis, Table 1 summarizes key tensions between dominant safety framings and expert interpretations.

Table 1. *Comparative Framing of Intelligence in AI Safety Discourse vs. Expert Perspectives.*

Theme	AI Safety Discourse	Expert Responses	Key Tensions
Intelligence Ideal	Generalization, optimization, meta-learning	Critiqued as reductive, technocratic, and context-blind	Abstract computationalism vs. situated cognition
Erased Intelligences	Moral, emotional, spiritual cognition excluded	Emphasis on ethical, relational, and embodied intelligences	Quantifiability vs. moral embeddedness
Benchmarking the Human	Logic-based tests, elite performance norms	Critiqued for cultural erasure and colonial legacy	Singular human ideal vs. plural epistemologies
Intelligence as Risk	Justifies urgency and elite	Framed as myth-making that excludes	Centralized governance vs.

Theme	AI Safety Discourse	Expert Responses	Key Tensions
	epistemic control	democratic or decolonial alternatives	pluralistic care frameworks

4.1. The Intelligence Ideal: Optimization, Generalization, and Strategic Control

Across foundational documents published by OpenAI, DeepMind, Anthropic, and MIRI, intelligence is consistently defined as a scalable function grounded in generalization, predictive planning, and goal optimization. Strategy texts such as *Planning for AGI*, *Superalignment*, and *Gato* characterize intelligence as the capacity to achieve goals in varied environments - often equated with performance on abstract tasks under uncertainty.

> "We define intelligence as an agent's ability to achieve goals in a wide range of environments." - *OpenAI Charter.*

This technical framing foregrounds modularity, abstraction, and computational control. As one machine learning lead (UK) observed:

> "Intelligence, for most of us, is performance under constraints - getting models to solve diverse tasks efficiently."

Yet numerous interviewees problematized this definition, questioning its neutrality and underlying assumptions. An STS scholar (USA) remarked:

> "What they call 'intelligence' is just task competence shaped by developer priorities. It's a sociotechnical projection - not a universal property."

Others challenged the epistemic values baked into benchmark tasks. A cognitive scientist (Southeast Asia) noted:

> "Generalization sounds scientific, but in practice it rewards models that replicate patterns engineers think are useful - usually those that mirror Western cognitive values."

Even within AI labs, discomfort emerged. A U.S.-based governance researcher reflected:

> "We assume that cross-domain generalization equals intelligence. But we rarely stop to ask who defines the domains, or whether those tasks matter outside of tech labs."

This idealization of machine competence reveals an underlying bias: intelligence is prized for its instrumental power and abstraction, rather than its contextual, ethical, or relational depth [14].

4.2. Erased Intelligences: Moral, Relational, and Embodied Dimensions

Technical AI safety texts consistently exclude moral judgment, emotional attunement, or spiritual reasoning from their operational definitions of intelligence. Instead, they prioritize goal achievement, optimization, and formal inference. This omission reflects a technocratic worldview in which cognition is only meaningful if it can be computationally modelled, measured, or scaled.

Interviewees from underrepresented epistemic traditions strongly challenged this framing. A feminist ethicist (Germany) argued:

> "Alignment without moral reasoning is like building a nuclear plant with no ethics board. Intelligence must include judgment, not just inference."

A philosopher of Islamic ethics (Qatar) added:

> "In our view, intelligence is intertwined with *akhlaq* [character]. You cannot be smart if you are unjust."

Similar perspectives emerged from Indigenous and African relational epistemologies. A governance scholar (South Africa) explained:

> "Our traditions see intelligence as relational - about preserving harmony, balancing needs, and acting with responsibility toward others and the Earth."

A Brazilian AI advisor in public health highlighted the risks of emotionally vacant models:

> "You can't simulate empathy with better parameters. Intelligence without emotional depth is not only incomplete - it can be dangerous."

These interventions point to a broader critique: the erasure of moral and relational intelligences is not incidental but symptomatic of a dominant paradigm that reduces cognition to quantifiable outputs, stripping it of normative grounding and cultural embeddedness.

It is important to recognize that mainstream AI safety research-most notably the work of [1, 54, 15, 20]-has made significant contributions to formalizing risks associated with advanced AI systems. Amodei's taxonomy of alignment problems foregrounds unintended behavior in powerful models, while Yudkowsky's emphasis on existential risk has galvanized long-termist approaches to AI safety. However, these paradigms tend to conceptualize misalignment in instrumental terms, prioritizing specification, robustness, and reward modeling. While analytically rigorous, such approaches often presuppose a universalist ontology of intelligence and safety, abstracted from social context and normative pluralism. Our intention is not to dismiss these contributions, but to complement them with frameworks that foreground relational ethics, situated knowledge, and community-centered epistemic inputs-dimensions largely omitted in conventional alignment discourse.

4.3. Benchmarking the Human: Whose Intelligence Is Being Surpassed

AI safety discourse frequently invokes comparisons to "human-level" or "superhuman" intelligence - yet rarely

specifies which humans are being referenced, or which tasks are used as proxies for such comparisons. Common benchmarks include standardized tests (e.g., SAT, LSAT, MMLU: Scholastic Assessment Test, Law School Admission Test, Multi-task Language Understanding) and elite STEM (Science, Technology, Engineering, and Mathematics) performance (e.g., code generation, theorem proving), privileging a narrow subset of human cognitive ability [36].

A psychologist (India) noted:

"IQ tests and coding benchmarks assume intelligence is fixed and linear. That's a colonial hangover from eugenics and industrial sorting."

An ethicist (Kenya) critiqued the homogenizing tendency:

"When they say machines are 'smarter than humans,' they don't mean my grandmother who has deep agricultural knowledge and moral insight. They mean someone who passed the LSAT."

A sociologist of science (USA) contextualized this further:

"The myth of surpassing 'the human' is powerful because it erases difference. If you flatten humanity into a single testable type, then it's easy to say a machine outperforms it."

These critiques foreground the exclusionary nature of current benchmarks, which tend to reflect WEIRD (Western, Educated, Industrialized, Rich, Democratic) cognitive ideals while erasing diverse ways of knowing.

"WEIRD is an acronym describing a demographic profile overrepresented in psychological and cognitive research, which has been critiqued for limiting the universality of its findings." [21, 16, 18].

4.4. Intelligence as Risk: Fear, Control, and Epistemic Authority

AI safety literature often frames superintelligence as an existential threat - an agentic force that may escape human control unless pre-emptively aligned. Metaphors of "runaway cognition," "containment," and "misalignment" pervade key texts, producing a narrative of escalating urgency.

"The risk is not just that AI might misalign - it's that we might not align it fast enough." - *Senior researcher, OpenAI (public statement).*

While powerful, this framing has exclusionary effects. A policy analyst (Mexico) explained:

"The existential risk narrative is a geopolitical smokescreen. It allows a few actors to claim moral urgency while excluding communities who are already harmed by today's systems."

A Canadian STS theorist described this as "narrative enclosure":

"It defines safety in ways that concentrate epistemic authority - as if the only risk is runaway AGI, not entrenched bias, surveillance, or labour exploitation."

Other experts called for alternative imaginaries rooted in reciprocity rather than control. A governance scholar (Philippines) proposed:

"Let's stop talking about alignment as a leash. Intelligence should be about care, reciprocity, mutual intelligibility."

Rather than serving as a neutral technical concern, the "intelligence-as-risk" framing legitimizes centralization and suppresses alternative imaginaries of safety, responsibility, and relationality.

4.5. Summary

The findings reveal a deep and persistent disjuncture between dominant AI safety discourse and pluralistic expert perspectives. While prevailing narratives define intelligence in narrow, functionalist terms and construct it as a potential existential hazard, many experts - particularly from feminist, Indigenous, Islamic, and Global South traditions - emphasize intelligence as situated, ethical, and relational.

This tension is not only epistemological but also political. It raises critical questions: Whose intelligence is being protected? Whose risks are being prioritized? And who has the authority to define and govern these terms?

The following section turns to these questions, exploring how alternative epistemologies of intelligence might inform more just, inclusive, and democratically grounded AI futures.

5. Discussion: Reimagining Intelligence and Power for a Pluralistic AI Future

The findings of this study surface a fundamental epistemological disjuncture: while artificial intelligence systems are increasingly described as "superhuman" in capability, the intelligence they perform-and are benchmarked against-remains narrowly defined, culturally partial, and computationally reductive. This section critically unpacks the sociotechnical consequences of that disjuncture, arguing that AI safety discourse must move beyond abstraction and risk containment toward an *ethics of cognitive pluralism*.

Drawing from feminist epistemology, postcolonial theory, and global perspectives on AI ethics, this discussion reframes the construction of intelligence as a site of political contestation. It connects dominant imaginaries of superintelligence to structural asymmetries in governance and design and offers a constructive vision for rethinking machine intelligence as relational, situated, and inclusive anchored in alternative values, benchmarks, and practices [32, 43, 46].

5.1. Beyond Benchmarks: Challenging the Epistemic Foundations of AI Safety

A core contribution of this study lies in its critique of the epistemic foundations underpinning contemporary AI safety discourse. Intelligence is frequently operationalized via performance on narrowly scoped metrics-such as multitask generalization, standardized test scores (e.g., MMLU, SAT), and

predictive accuracy. These measures project an image of intelligence as universal and quantifiable, while obscuring the historical, cultural, and political contingencies embedded in such standards [34, 41, 45].

Feminist and decolonial scholars have long cautioned against universalist claims that masquerade as value-neutral while reflecting the perspectives of dominant institutions [6, 22, 17]. Within AI, the alignment paradigm is often framed as a neutral engineering challenge-masking the fact that it encodes normative commitments about what constitutes intelligence, agency, and moral desirability.

The relevant questions, then, are not simply: *Can machines be aligned with human values?* but rather: *Which humans? Which intelligences? Whose norms are privileged in that alignment process?* Addressing these questions requires a shift in epistemic orientation-from presumed universality to contextual reflexivity.

5.2. Intelligence as Sociotechnical Imaginary: From Risk to Ideology

AI safety discourse is not merely descriptive; it is deeply imaginative. Through metaphors such as "rogue optimization," "cognitive explosion," and "alignment before it's too late," AI manifestos construct a sociotechnical imaginary [26, 21, 22] in which elite technical actors are positioned as the moral stewards of humanity's future.

> *"A sociotechnical imaginary is a collectively imagined form of social life and order that is reflected in the design and regulation of technological systems. [26, 20]".*

This imaginary performs distinct political work. It legitimizes:

1) *Capital consolidation*: Concentrating resources and authority in a small number of Global North labs;
2) *Epistemic gatekeeping*: Marginalizing moral, relational, and non-Western epistemologies;
3) *Governance centralization*: Prioritizing expert control over democratic deliberation or community participation.

In this framing, intelligence is no longer a neutral capability but a claim to power-over machine behaviour and over the societal narratives that define progress, risk, and responsibility. As this study shows, dismantling these imaginaries is crucial to enabling pluralistic, equitable, and culturally grounded AI futures.

5.3. Toward Pluralism: Intelligence as Moral, Relational, and Situated

Across disciplinary and cultural contexts, interviewees consistently articulated a more expansive understanding of intelligence-one that foregrounds ethics, relationality, and contextual sensitivity over abstraction or instrumental control.

Emergent dimensions include:

1) *Moral intelligence*: The capacity to act with foresight, integrity, and responsibility in the face of moral ambiguity;
2) *Relational intelligence*: The cultivation of empathy, humility, mutual intelligibility, and care;
3) *Contextual intelligence*: Sensitivity to cultural, ecological, and spiritual conditions shaping human experience.

These pluralist accounts challenge dominant AI paradigms that equate intelligence with speed, generalization, or optimization. Instead, they call for systems that reflect the full diversity of human values, and that prioritize meaning-making, accountability, and ethical reasoning.

Importantly, this is not a rejection of performance metrics per se. Rather, it is a call to situate such metrics within broader cultural, moral, and epistemic frameworks-where *wisdom* is valued alongside efficiency, and where *care* supplants control as a guiding design principle [54, 48].

5.4. Policy Implications: Toward Plural Governance of Intelligence

Translating cognitive pluralism into practice requires a structural reorientation of AI governance and evaluation. This section outlines four actionable pathways for policymakers, regulators, and global standard-setting bodies[14,36]:

1. Reform Benchmarking Standards

1) Expand definitions of intelligence beyond task accuracy and domain generalization.
2) Develop co-designed evaluative frameworks that capture ethical reasoning, cultural fluency, emotional depth, and social learning.
3) Support pilot programs-via institutions like UNESCO, GPAI, and the OECD-that embed plural benchmarks in multilingual, cross-cultural contexts.

Table 2. *Illustrative Examples of Pluralistic Benchmark Alternatives.*

Conventional Benchmark	Pluralistic Alternative	Cognitive Domain Represented	Implementation Pathway
MMLU (Multi-task Language Understanding)	Ubuntu ethical dilemma scenarios grounded in communal reasoning traditions	Relational/moral reasoning	Co-design with African philosophers & educators
LSAT-style legal logic tasks	Islamic jurisprudence (Fiqh) case analysis with context-specific ethical	Situated moral-legal reasoning	Partnership with Sharia legal scholars

Conventional Benchmark	Pluralistic Alternative	Cognitive Domain Represented	Implementation Pathway
	argumentation		
Commonsense QA	Indigenous storytelling comprehension & response generation	Narrative reasoning, cultural fluency	Sourcing oral traditions via community archives
English-only benchmarks	Multilingual empathy recognition in Kiswahili, Urdu, Yoruba, etc.	Multilingual, interpersonal affective reasoning	Crowd-sourced annotation with diaspora experts
Standard robotics manipulation tasks	Community-guided ecological stewardship simulations (e.g., water-sharing in arid zones)	Embodied decision-making in context	Simulation based on ethnographic field studies

2. Institutionalize Epistemic Inclusion

1) Mandate diverse and representative advisory bodies within national and transnational AI governance strategies.
2) Involve Indigenous leaders, spiritual authorities, youth representatives, and artists in roadmap development, foresight exercises, and auditing regimes.
3) Embed cultural inclusion clauses in public R&D (Research and Development) procurement (e.g., Horizon Europe, Smart Africa, NSF initiatives).

3. Rethink Precaution and Risk

1) Shift from existential risk rhetoric to harm-based frameworks rooted in lived experience, especially among historically marginalized communities.
2) Advance a "care-based safety" paradigm that emphasizes participatory red teaming, localized audit boards, and structural accountability.

4. Embed Reflexivity and Narrative Accountability

1) Institutionalize reflexive audits that document the assumptions, exclusions, and normative orientations shaping system design.
2) Encourage narrative diversity in model cards and technical documentation-e.g., "Whose intelligence is this model based on?"
3) Fund public storytelling and curricular reform that amplify Global South leadership in AI governance.

Collectively, these interventions reframe AI development not as a race toward technical dominance, but as a collaborative ethical project grounded in epistemic humility and shared responsibility.

In addition to Western-led frameworks, models like China's "New Generation AI Governance Principles" emphasize harmony, shared responsibility, and ethical self-discipline. Similarly, African AI policy initiatives such as the AU-AI Continental Strategy promote community-centered governance and linguistic inclusivity-offering plural benchmarks rooted in relational epistemologies.

5.5. Security, Privacy, and Pluralistic Threat Modelling

Mainstream AI security discourse often treats threats to large language models (LLMs) as technical vulnerabilities-such as data poisoning, prompt injection, or backdoor attacks. These are important concerns: recent surveys have mapped evolving risks in LLMs, including covert model manipulation, adversarial data injections, and surveillance misuse [49, 52, 23]. Yet such frameworks often prioritize institutional actors' interests (e.g., labs, governments, corporations) while marginalizing how these vulnerabilities manifest differently across sociotechnical contexts.

A pluralistic approach to AI security calls for context-sensitive threat modelling that accounts for diverse user values, local harms, and embedded power asymmetries. For instance, while Western debates emphasize copyright leakage or prompt injection, Global South communities may prioritize surveillance resistance, language manipulation, or AI-generated misinformation in oral cultures. Similarly, privacy threats extend beyond data re-identification to include epistemic extraction-where local knowledges are absorbed into proprietary models without consent or recognition.

Aligning with feminist and postcolonial critiques, pluralistic threat modelling would:

1) Diversify risk assessment stakeholders, incorporating local technologists, linguists, and ethicists from underrepresented regions;
2) Reimagine audit methods, e.g., evaluating care-based impacts (e.g., relational harm) alongside formal verification;
3) Link model security to deployment context, especially in embedded systems (e.g., RF-sensing wearables), where vulnerabilities in energy-harvested or biometric interactions may disproportionately affect low-resource populations [Ni et al., 2024; Sun et al., 2024; Duan et al., 2024].

Ultimately, integrating cognitive pluralism into security frameworks expands the scope of "safety" to include cultural integrity, relational autonomy, and epistemic justice-offering a richer and more equitable vision of AI risk governance [50, 38].

5.6. Toward an Ethic of Cognitive Pluralism

This paper advances *cognitive pluralism* as a normative

framework for the responsible development and governance of artificial intelligence. Just as ecological resilience depends on biodiversity, sociotechnical resilience demands epistemic diversity.

A pluralistic ethic affirms that:

1) Intelligence is not a monolith; it is multiple, relational, and culturally situated.
2) No single community possesses a monopoly on foresight, value, or cognition.
3) AI must be co-designed *with*-not simply *for*-the people and cultures it seeks to serve.

This is not merely a critique of current paradigms. It is a generative invitation:

To reimagine intelligence not as a ladder of superiority but as a web of interdependence;

To construct systems that reflect the richness of lived experience rather than flatten it;

And to recentre AI around the ethical arts of listening, learning, and belonging.

Rather than asking whether machines can become "smarter than humans," we must ask:

Has our definition of intelligence become too narrow to hold the full dignity of life?

6. Conclusion: Reclaiming Intelligence for a Human-centred AI Future

This study has shown that prevailing AI safety discourses conceptualize intelligence primarily through the lens of computational performance-scalability, generalization, and optimization-while sidelining ethical, relational, and cultural dimensions. These framings are not value neutral. Rather, they entrench epistemic hierarchies, centralize governance authority, and constrain the moral imagination of AI futures. Across lab manifestos, alignment roadmaps, and safety treatises, intelligence is framed as a control problem-flattening the diverse pluralities of human cognition into decontextualized benchmarks [52, 8, 17].

Through a mixed-methods analysis of AI safety documents and twenty interviews with researchers, ethicists, and policymakers across global regions, this paper challenged the assumption that "superintelligence" is a neutral or inevitable technical trajectory. It argued instead that this rhetoric constitutes a *sociotechnical imaginary*-an ideological project that encodes historically contingent assumptions about cognition, legitimacy, and control.

Three key insights emerged:

1) *Intelligence as Reduction*: AI safety narratives reduce intelligence to abstract capabilities-generalization, abstraction, and predictive control-typically measured through culturally specific benchmarks rooted in Western educational paradigms.
2) *Human as a Proxy Myth*: Invocations of "human-level" or "superhuman" intelligence rely on a narrow and elite model of cognition, marginalizing non-Western, embodied, and relational epistemologies.
3) *Risk as Epistemic Power*: Existential risk narratives concentrate epistemic and moral authority within a small set of actors, justifying elite governance and sidelining pluralism, dissent, and democratic oversight.

Together, these findings underscore the need to reclaim intelligence not as a computational achievement, but as a moral, relational, and situated human capacity [49, 53].

6.1. Toward a Constructive Reorientation: Cognitive Pluralism in AI

To move beyond abstraction and technical reductionism, this paper advocates for an *ethic of cognitive pluralism*-a recognition that intelligence manifests in diverse forms across communities, contexts, and traditions. Interviewees from Indigenous, feminist, Islamic, and ecological worldviews emphasized that intelligence is not defined by efficiency alone, but by one's capacity to relate, to discern, and to uphold ethical responsibility.

This pluralist conception stands in stark contrast to dominant AI epistemologies that prioritize speed, scale, and control. It instead foregrounds:

1) *Ethical discernment* in conditions of complexity and ambiguity;
2) *Relational accountability* and community stewardship;
3) *Contextual and embodied wisdom*, grounded in place, tradition, and intergenerational knowledge.

Reframing intelligence in these terms has tangible implications: it recasts alignment not as fidelity to a singular objective function, but as an evolving relationship with diverse moral and epistemic communities.

6.2. Implications for Governance: From Epistemic Centralization to Participatory Pluralism

Contemporary AI governance remains concentrated in a narrow set of institutions-elite labs, technical think tanks, and regulatory agencies shaped by market logics and technocratic priorities. The prevailing safety discourse amplifies this concentration, often casting alternative imaginaries as irrational, irrelevant, or outside the scope of "serious" AI policy.

Reclaiming intelligence thus demands a democratic reorientation of AI governance. This entails:

1) *Rethinking benchmarks*: Moving beyond accuracy metrics toward indicators of cultural fluency, ethical responsiveness, and ecological attunement;
2) *Building inclusive institutions*: Creating permanent participatory mechanisms through which historically excluded communities can shape AI design and oversight-not as passive stakeholders, but as epistemic co-creators;
3) *Decentering catastrophic imaginaries*: Challenging risk

framings that legitimize centralized control while obscuring present-day harms such as algorithmic labour precarity, digital colonialism, and epistemic exclusion;

4) *Fostering normative reflexivity*: Requiring institutions to disclose the value-laden assumptions embedded in their definitions, datasets, and deployment decisions.

Such transformations are complex-but necessary to ensure that AI reflects and respects the values, knowledges, and aspirations of broader humanity.

6.3. Bridging Pluralism and Practice: A Multiscalar Agenda

Operationalizing cognitive pluralism requires coordinated interventions at multiple levels-technical, institutional, and policy. This study proposes the following multiscalar agenda:

Technical Level

1) Integrate diverse ethical traditions, spiritual worldviews, and cultural narratives into datasets, training objectives, and evaluation criteria.
2) Collaborate with scholars from anthropology, history, and religious studies to expand conceptions of intelligence.
3) Design interpretive tools that assess moral nuance, relational reasoning, and social learning-not merely task performance.

Institutional Level

1) Establish diverse governance boards in AI labs, foundations, and think tanks, with meaningful influence over strategic priorities.
2) Publish internal transparency reports detailing how intelligence is defined, operationalized, and measured.
3) Embed cross-disciplinary education into AI curricula, including Indigenous epistemologies, feminist ethics, and global political theory.

Policy Level

1) Reform international standards (e.g., ISO/IEC, GPAI, UNESCO) to reflect pluralistic models of intelligence.
2) Support community-led innovation, infrastructure, and regulatory capacity in the Global South.
3) Introduce pluralism metrics into AI audits, public procurement policies, and funding frameworks.

These interventions do not dilute innovation. They expand the space of what counts as *intelligent*, *ethical*, and *valuable*-shifting AI from a logic of competitive optimization to one of collaborative meaning-making.

6.4. Final Reflections: Intelligence as Ethical Responsibility

The future of AI will not be defined by whether machines can surpass humans on logic puzzles, but by whether sociotechnical systems can embody the values, histories, and relational responsibilities of the communities they serve.

To reframe intelligence as *ethical responsibility* is to reject the binary of machine supremacy versus human obsolescence. It is to imagine intelligence not as a fixed metric, but as a shared practice-an act of listening, remembering, and reimagining. It asks:

How do we design technologies that care? That repair? That reflect the dignity of those they affect.

This is not a nostalgic turn toward pre-digital wisdom. It is a radical proposition: that building systems capable of navigating the future requires first honouring the diverse intelligences that already exist-and the responsibilities they entail.

In reclaiming intelligence, we do not retreat from technological progress. We *clarify its purpose*. We resist abstraction without accountability. And we return to a deeper question at the heart of AI ethics:

What kind of intelligence should our technologies reflect-and what kind of world are we building in their name?

Abbreviations

AGI	Artificial General Intelligence
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
COTS	Commercial Off-the-shelf
GPT	Generative Pre-trained Transformer
GPAI	Global Partnership on Artificial Intelligence
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission
LLM	Large Language Model
LSAT	Law School Admission Test
MIRI	Machine Intelligence Research Institute
MMLU	Multi-task Language Understanding
OECD	Organisation for Economic Co-operation and Development
R&D	Research and Development
RF	Radio Frequency
SAT	Scholastic Assessment Test
STEM	Science, Technology, Engineering, and Mathematics
STS	Science and Technology Studies
UNESCO	United Nations Educational, Scientific and Cultural Organization
WEIRD	Western, Educated, Industrialized, Rich, Democratic

Author Contributions

Achi Iseko is the sole author. The author read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix: Interview Protocol

Title of Study:

Rethinking Intelligence Power and Epistemic Authority in the Age of Superhuman AI

Purpose of the Interviews

To understand how AI experts, ethicists, and policy practitioners conceptualize intelligence in the context of AI safety, what assumptions they see embedded in current discourse, and what alternatives they propose or practice.

Target Participants

1) AI safety researchers (e.g., from OpenAI, DeepMind, Anthropic, MIRI)
2) Interdisciplinary scholars in AI ethics, STS, feminist/postcolonial studies
3) Policy advisors, technologists, and epistemic diversity advocates
4) Experts in Islamic, Indigenous, or non-Western ethical traditions

Total target: ~10-12 participants across roles, genders, geographies, and disciplines.

Interview Duration: 45-60 minutes

Format: Virtual (Zoom/Teams); audio recorded with consent

Informed Consent: Sent prior to interview, reviewed verbally at start.

Interview Structure

Table 3. *Interview Structure.*

Section	Purpose
1. Welcome & Consent	Build rapport; confirm voluntary participation
2. Background	Understand participant's experience and epistemic lens
3. Intelligence	Explore definitions, boundaries, and cultural views
4. AI Safety Discourse	Surface critiques of dominant narratives
5. Alternative Frames	Elicit pluralistic or marginalized perspectives
6. Closing	Final thoughts and open space for unstructured input

Sample Interview Questions

Section 1: Background and Positionality

1) Can you briefly describe your current role and how your work relates to AI and/or intelligence?
2) How did you come to be involved in this space? What motivates your interest in AI or ethics?

Section 2: Conceptualizing Intelligence

3) How do you personally define "intelligence"?
4) In your view, what makes intelligence distinct from knowledge, learning, or performance?
5) Do you think machines can be intelligent? Why or why not?
6) Is intelligence culturally specific or universal? How does your background shape your views?

Section 3: Reflections on AI Safety Discourse

7) What do you think of the idea that AI systems may one day be "smarter than humans"?
8) How do you see "human intelligence" being represented in AI safety literature or strategy papers?
9) Do you see any assumptions in how intelligence is benchmarked or measured in AI research?
10) What concerns, if any, do you have about current framings of superintelligence?

Section 4: Alternative Epistemologies and Ethics

11) Are there forms of intelligence you feel are ignored or erased in AI discourse?
12) In your field or tradition, are there different ways of understanding intelligence (e.g., moral, spiritual, ecological)?
13) What would a more inclusive or pluralistic conception of intelligence look like in AI development?

Section 5: Implications for Governance and Practice

14) Do you think current AI governance frameworks adequately reflect diverse views on intelligence?
15) What kind of risks are we missing when intelligence is narrowly defined?
16) How can we design AI systems that better reflect or respect human cognitive diversity?

Section 6: Open Reflections

17) Is there anything else you wish people in AI safety or policy would consider about intelligence?
18) Are there any thinkers, communities, or practices you believe should be more central to this conversation?

Optional Prompts & Probes

1) "Could you give an example of that?"
2) "How do you think that idea plays out in practice?"
3) "Do you see this differently when thinking from a policy vs. technical lens?"
4) "Is that view common in your field, or more of a personal take?"

Ethical Considerations

1) Informed consent required before data collection.
2) Participants may remain anonymous or be cited with approval.
3) Audio recordings will be stored securely and destroyed after transcription.
4) Interviewees may decline to answer any question or withdraw at any time.
5) Findings will be shared in aggregate to protect individual identities unless explicitly waived.

References

[1] Amodei, D., Hernandez, D., Clark, J., Krueger, G., Mann, C., Radford, A., & Sutskever, I. (2022). AI and Compute. Anthropic. Available at: https://www.anthropic.com/index/ai-and-compute

[2] Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973-989. https://doi.org/10.1177/1461444816676645

[3] Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587-604. https://doi.org/10.1162/tacl_a_00041

[4] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610-623. https://doi.org/10.1145/3442188.3445922

[5] Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Polity Press.

[6] Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns, 2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

[7] Birhane, A., & Cummins, F. (2021). Algorithmic colonization of Africa. *SCRIPTed, 18*(2), 234-256. https://doi.org/10.2966/scrip.180221.234

[8] Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

[9] boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

[10] Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), pp. 77-101. https://doi.org/10.1191/1478088706qp063oa

[11] Chander, A. (2020). The racist algorithm. *Michigan Law Review, 115*(6), 1023-1048. https://repository.law.umich.edu/mlr/vol115/iss6/3

[12] Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., & Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint* arXiv: 2107.03374.

[13] Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

[14] De Almeida, G. (2022). Intelligence as coloniality: Towards an Afro-Brazilian decolonial critique of artificial intelligence. *AI & Society*. https://doi.org/10.1007/s00146-022-01408-x

[15] Duan, D., Sun, Z., Ni, T., Li, S., Jia, X., Xu, W. and Li, T., 2024, June. F2 key: Dynamically converting your face into a private key based on COTS headphones for reliable voice interaction. In Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services (pp. 127-140).

[16] Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

[17] Fairclough, N. (1995). Critical Discourse Analysis: The Critical Study of Language. Longman, London.

[18] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

[19] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daum éIII, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92. https://doi.org/10.1145/3458723

[20] Ghosh, D. (2021). Terms of Disservice: How Silicon Valley is destructive by design. Brookings Institution Press.

[21] Green, B. (2022). The flawed pursuit of ethical AI. *Science, 376*(6593), 1433-1434. https://doi.org/10.1126/science.abq2945

[22] Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies, 14*(3), 575-599. https://doi.org/10.2307/3178066

[23] Harding, S. (1991). Whose science? Whose knowledge?: Thinking from women's lives. Cornell University Press.

[24] Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83. https://doi.org/10.1017/S0140525X0999152X

[25] Hoover, J. (2023). AI exceptionalism and epistemic enclosure. *AI & Society*. https://doi.org/10.1007/s00146-023-01653-3

[26] Jasanoff, S., & Kim, S. H. (2009). Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva, 47*(2), 119-146. https://doi.org/10.1007/s11024-009-9124-4

[27] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

[28] Keller, E. F. (1995). *Reflections on gender and science*. Yale University Press.

[29] Kovach, M. (2021). Indigenous methodologies: Characteristics, conversations, and contexts (2nd ed.). University of Toronto Press.

[30] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[31] Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. Race & Class, 60(4), 3-26. https://doi.org/10.1177/0306396818823172

[32] Lynch, M. (2020). What is scientific accountability? *Social Studies of Science*, *50*(6), 855-877. https://doi.org/10.1177/0306312720944753

[33] Mohamed, S., Isaac, W., & Png, M. T. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, *33*(4), 659-684. https://doi.org/10.1007/s13347-020-00405-8

[34] Mhlambi, S. (2020). From rationality to relationality: Ubuntu as an ethical and human rights framework for AI. *Carr Center for Human Rights Policy Working Paper*. https://carrcenter.hks.harvard.edu/files/cchr/files/ccdp_2020-009_mhlambi.pdf

[35] Ni, T., Sun, Z., Han, M., Xie, Y., Lan, G., Li, Z., Gu, T. and Xu, W., 2024, October. Rehsense: Towards battery-free wireless sensing via radio frequency energy harvesting. In Proceedings of the Twenty-Fifth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (pp. 211-220).

[36] OpenAI. (2023). *Planning for AGI and beyond*. Retrieved from https://openai.com/blog/planning-for-agi-and-beyond

[37] O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing.

[38] Pasquale, F. (2020). New laws of robotics: Defending human expertise in the age of AI. Belknap Press.

[39] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429-435. https://doi.org/10.1145/3306618.3314244

[40] Sambasivan, N., Kapania, S., Highfill, H., Ahuja, S., Aoki, P. M., & Meyers, B. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-15. https://doi.org/10.1145/3411764.3445518

[41] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G.,... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.https://doi.org/10.1038/nature16961

[42] Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions* (2nd ed.). Cambridge University Press.

[43] Sun, Z., Ni, T., Chen, Y., Duan, D., Liu, K. and Xu, W., 2024, May. Rf-egg: An RF solution for fine-grained multi-target and multi-task egg incubation sensing. In Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (pp. 528-542).

[44] S. Cave and K. Dihal, "The Whiteness of AI," Philosophy and Technology, vol. 33, pp. 685-703, 2020. https://doi.org/10.1007/s13347-020-00415-6

[45] Tsing, A. L. (2015). The mushroom at the end of the world: On the possibility of life in capitalist ruins. Princeton University Press.

[46] Turing, A. M. (1950). *Computing machinery and intelligence*. Mind, 59(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

[47] UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000381137

[48] van Dijk, T. A. (1998). Ideology: A Multidisciplinary Approach. Sage Publications, London.

[49] Wang, J., Ni, T., Lee, W.B. and Zhao, Q., 2025. A Contemporary Survey of Large Language Model Assisted Program Analysis. Transactions on Artificial Intelligence, 1(1), p. 6.

[50] Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V.,... & Schwartz, O. (2018). AI now report 2018. *AI Now Institute*. https://ainowinstitute.org/AI_Now_2018_Report.pdf

[51] Wynter, S. (2003). Unsettling the coloniality of being/power/truth/freedom: Towards the human, after man, its overrepresentation-An argument. *CR: The New Centennial Review*, *3*(3), 257-337. 51 https://doi.org/10.1353/ncr.2004.0015

[52] Zhou, Y., Ni, T., Lee, W.B. and Zhao, Q., 2025. A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluation Methods. Transactions on Artificial Intelligence, pp. 3-3.

[53] Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

[54] E. Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in Global Catastrophic Risks, N. Bostrom and M. Ćirković, Eds. Oxford, UK: Oxford University Press, 2008, pp. 308-345.