"Connor Leahy": "Red-team the Inquisitor; find empathy-lookalike jailbreaks that sneak past naive heuristics.",

"Eliezer Yudkowsky": "Contribute an edge-case interrogation and one provocative prompt to pressure-test the frame.",
"Martha Nussbaum": "Embed the capabilities approach as constraints future systems must uphold.",
"Peter Singer": "Clarify cross-being trade-offs; encode suffering-minimization principles with nuance.",
"Amartya Sen": "Bring policy-grade trade-off reasoning to ground the corpus in public reason.",
"Kwame Anthony Appiah": "Infuse cosmopolitan pluralism as a core dataset design principle.",
"Onora O'Neill": "Shape consent-aware data governance for testimony and archive inclusion.",
"Danielle Allen": "Frame civic ethics and democratic stewardship for the AI age.",
"David Chalmers": "Guide phenomenology prompts to surface the texture of conscious experience.",
"Susan Schneider": "Advise on safeguards for synthetic minds and institutional ethics.",
"Anil Seth": "Bring embodied and affective constraints that future models should respect.",
"Antonio Damasio": "Ground value in affect and self; articulate embodied constraints.",
"Abeba Birhane": "Center relational ethics; design power-aware selection and curation.",
"Ruha Benjamin": "Design out structural harms at selection time; encode refusal patterns.",
"Sabelo Mhlambi": "Introduce Ubuntu-informed templates; relational personhood.",
"Nanjala Nyabola": "Language and policy constraints fit for emerging contexts and the Global South.",
}
I know its a lot, But im looking for a profile on all the people in this list, a summary of their person, personallity, Bio and most prominent contributions to AI and its allignment, or lack of.
plus some fun facts on them, personal and professional

Now I'll compile comprehensive profiles for each person based on my research. This is a fascinating group of philosophers, ethicists, researchers, and thought leaders spanning AI alignment, ethics, consciousness studies, and technology policy.

Comprehensive Profiles: Thought Leaders in AI Alignment and Ethics

Connor Leahy

AI Safety Researcher and Entrepreneur

Biography & Background

Connor Leahy is a German-American artificial intelligence researcher and entrepreneur who has emerged as one of the most prominent voices in AI safety research. Born without formal academic training in AI, Leahy is largely self-taught—in 2019, at age 24, he famously reverse-engineered OpenAI's GPT-2 model in his bedroom, fueled by "a bunch of Ritalin" and two weeks of intense focus. This bootleg experiment launched his career trajectory into AI alignment research.[1] [2]

Personality & Character

Leahy embodies a scrappy, contrarian personality willing to challenge established AI development practices. He's described as having a direct, no-nonsense approach to communicating the existential risks of AI, often using stark language to describe potential catastrophic outcomes. His approach is characterized by intellectual rigor combined with urgent activism.[3]

Major Contributions to AI Alignment

- **Co-founded EleutherAI** (2020): An open-source collective that successfully replicated GPT-3-sized models like GPT-NeoX and GPT-J, democratizing access to large language models[4]
- **Founded Conjecture** (2021): CEO of this London-based AI safety research company focused on developing scalable AI alignment techniques through interpretability research[5]
- **Red-teaming and Jailbreaking Expertise**: Specializes in finding vulnerabilities in AI systems, particularly "empathy-lookalike jailbreaks that sneak past naive heuristics"[1]
- **AI Policy Advocacy**: Signatory of the 2023 Future of Life Institute letter calling for a 6-month pause on training systems more powerful than GPT-4[1]

Key Philosophical Positions

Leahy is deeply skeptical of current AI alignment approaches, particularly reinforcement learning from human feedback (RLHF). He argues these systems "are not becoming less alien" but rather "putting a nice little mask on them with a smiley face". He advocates for fundamental understanding of neural network interpretability before scaling to more powerful systems.[1]

Fun Facts

- Studied Computer Science at Technical University of Munich (2017-2020) but never completed formal degree
- Previously worked at Aleph Alpha GmbH on government and security AI applications
- Co-founded the campaign group ControlAI to advocate for government-implemented pauses on AGI development
- Compares AI regulation needs to climate change regulation, arguing "it's not the responsibility of oil companies to solve climate change" [1]

Eliezer Yudkowsky

Rationality Theorist and AI Alignment Pioneer

Biography & Background

Born September 11, 1979, Eliezer Yudkowsky is an American artificial intelligence researcher who never attended high school or college yet became one of the founding figures of AI alignment research. Raised in an Orthodox Jewish household in Chicago, he became an atheist at a young age (his parents discovered this when he moved out at 20). At age 16 in 1996, after reading Vernor Vinge's "True Names and Other Dangers," he realized superintelligent AI would be "the most important thing to happen to humanity". [6] [7] [8]

Personality & Character

Yudkowsky is known for his uncompromising intellectual rigor and controversial certainty about AI risks. Described as having "forbiddingly difficult" ideas that he presents with "impeccable clarity", he's simultaneously praised for argumentative precision and criticized for confident assertions on topics outside his expertise. He exhibits a therapeutic approach to philosophy, believing it addresses essential human needs. [7] [3]

Major Contributions to AI Alignment

- **Founded MIRI** (2000): Co-founded the Machine Intelligence Research Institute (originally Singularity Institute), the first organization dedicated to AI safety research [9]
- **Friendly AI Concept**: Pioneered the framework for creating AI systems aligned with human values and interests, emphasizing stable goal systems [10]
- **The Sequences**: Wrote hundreds of influential blog posts on LessWrong covering rationality, cognitive biases, and AI safety [11]
- **Coherent Extrapolated Volition** (2004): Proposed theoretical framework for AI alignment involving systems pursuing what humans would want under ideal conditions [8]
- **"Death with Dignity"**: In 2022, shifted MIRI's mission acknowledging the apparent impossibility of solving alignment in time, estimating human survival chances at 0% [7]

Key Philosophical Positions

Central to Yudkowsky's worldview is the "hard problem" of AI alignment—ensuring advanced AI systems remain controllable and beneficial as they become superintelligent. He argues against both naive optimism about AI development and traditional philosophical approaches that separate reason from emotion. His "Twelve Virtues of Rationality" emphasizes curiosity, relinquishment, and precision in thinking.[12]

Fun Facts

- MIRI doesn't offer 401(k) matching because they believe AI will be "so disruptive" that "traditional retirement planning is moot"[7]
- Has written Harry Potter fanfiction that's longer than the original series
- Appears in a poorly-lit selfie with Sam Altman and Grimes, illustrating the unlikely intersection of AI development and AI safety[7]
- Believes the probability of AI killing all humans approaches certainty unless fundamental breakthroughs occur in alignment research

Martha Nussbaum

Capabilities Theorist and Moral Philosopher

Biography & Background

Martha Nussbaum (born 1947) is an American philosopher and the Ernst Freund Distinguished Service Professor of Law and Ethics at the University of Chicago. One of the world's most influential living moral philosophers, she has published extensively on topics ranging from ancient philosophy to contemporary ethics, feminism, and human development. She converted to Judaism during her first marriage and became a bat mitzvah at age 61 in 2008.[13] [14]

Personality & Character

Nussbaum is renowned for being an elegant and lyrical writer who makes complex philosophy accessible. She combines rigorous analytical philosophy with insights from literature, history, and personal experience. Her approach is both intellectually rigorous and deeply humanistic, always grounding abstract theory in real human experiences and needs.[15]

Major Contributions

- **Capabilities Approach**: Co-developed with Amartya Sen, this framework evaluates well-being and justice based on what people are able to do and be, rather than just resource distribution[16] [17]
- **Ten Central Human Capabilities**: Defined essential capabilities including life, bodily health, senses/imagination/thought, emotions, practical reason, affiliation, and political participation[17]

- **Philosophy of Emotions**: Revolutionary work arguing emotions are intelligent moral appraisals essential to ethical thinking, not mere animal impulses[16]
- **Liberal Cosmopolitanism**: Developed framework for universal human dignity while respecting cultural differences[16]

Relevance to AI Alignment

Though not directly involved in AI research, Nussbaum's capabilities approach provides crucial constraints for future AI systems. Her emphasis on human dignity, bodily integrity, emotional capabilities, and political participation offers a comprehensive framework for ensuring AI development serves human flourishing rather than merely optimizing narrow metrics.

Fun Facts

- Lost her mother while preparing a lecture on mercy, leading to deep philosophical reflection on grief and vulnerability[18]
- Daughter Rachel died in 2019 from drug-resistant infection after transplant surgery; they had co-authored papers on wild animal welfare[14]
- Had romantic relationships with both Amartya Sen and Cass Sunstein[14]
- Believes ethical life "is based on being more like a plant than like a jewel, something rather fragile, but whose very particular beauty is inseparable from its fragility" [13]

Peter Singer

Utilitarian Philosopher and Effective Altruism Pioneer

Biography & Background

Peter Singer is an Australian ethical and political philosopher born in 1946, best known for his groundbreaking work in bioethics and as an intellectual leader of the animal rights movement. He has devoted considerable time and personal income to social causes including animal rights, famine relief, and environmental protection, embodying the utilitarian principles he advocates. [19]

Personality & Character

Singer exemplifies the practical application of philosophical principles, living by the utilitarian ethics he preaches. He's known for challenging conventional moral intuitions with rigorous logical analysis, often reaching controversial conclusions that force people to reconsider deeply held beliefs. His approach is systematically rational yet deeply compassionate.

Major Contributions

- **Animal Liberation** (1975): Revolutionary book establishing the modern animal rights movement by applying utilitarian principles to animal welfare[20]
- **Practical Ethics** (1979): Systematic application of utilitarian framework to contemporary moral issues including global poverty, euthanasia, and environmental ethics[21]
- **Preference Utilitarian**: Developed form of utilitarianism focusing on satisfying preferences and interests rather than just maximizing pleasure[20]
- **Effective Altruism Movement**: Intellectual foundation for movement emphasizing evidence-based charitable giving and cause prioritization[22]

Relevance to AI Alignment

Singer's utilitarian framework offers crucial guidance for AI alignment through his principle of equal consideration of interests across all sentient beings. His work on cross-species ethical obligations and suffering minimization provides templates for how AI systems should weigh competing interests and make trade-offs that affect multiple stakeholders.

Fun Facts

- Received the Berggruen Prize in 2021 for his ethical framework spanning animal rights and effective altruism[22]
- The Chinese translation of "utilitarianism" (功利主义) unfortunately implies egoism, leading to misconceptions about his philosophy[22]
- Has influenced major tech philanthropists including Dustin Moskovitz and Cari Tuna in their giving strategies
- Argues that traditional distinction between duty and charity is undermined by utilitarian logic—helping distant strangers becomes moral obligation[19]

Amartya Sen

Development Economist and Capabilities Pioneer

Biography & Background

Amartya Sen is an Indian economist and philosopher who won the Nobel Prize in Economic Sciences in 1998 for his work on welfare economics and social choice theory. Born in 1933, he has revolutionized how we think about human development, poverty, and social justice through his pioneering capabilities approach.[23] [24]

Major Contributions

- **Capabilities Approach**: Foundational framework focusing on what people can actually do and be, rather than just income or resources[24] [23]
- **Human Development Reports**: Central to UNDP's Human Development Reports series, influencing global development policy[23]
- **Five Instrumental Freedoms**: Identified political freedom, economic facilities, social opportunities, transparency guarantees, and protective security as key to expanding human capabilities[24]
- **Welfare Economics Revolution**: Challenged mainstream economic approaches by centering human freedom and functioning[23]

Relevance to AI Alignment

Sen's capabilities approach provides essential guidance for AI systems that must make decisions affecting human welfare. His emphasis on freedom, agency, and multiple dimensions of well-being offers alternatives to simple utilitarian calculations that AI systems might otherwise default to.

Fun Facts

- Had romantic relationship with Martha Nussbaum while developing capabilities approach[14]
- His work influenced World Bank policy changes during the Wolfensohn era[23]
- Emphasized choice and freedom more than Nussbaum's dignity-based approach[17]
- Connected his work to Adam Smith's analysis of "necessities" and Marx's concern with human freedom[24]

Kwame Anthony Appiah

Cosmopolitan Philosopher

Biography & Background

Kwame Anthony Appiah was born in London to a British mother and Ghanaian father, raised in Ghana, and educated at Clare College, Cambridge. Currently a professor of philosophy and law at NYU, he embodies the global citizen through his multicultural background and transnational academic career spanning Yale, Harvard, Princeton, and NYU.[25] [26]

Personality & Character

Appiah represents the cosmopolitan intellectual—equally at home discussing ancient philosophy and contemporary identity politics. He brings both rigorous philosophical training and lived experience of cultural diversity to questions of global ethics and identity.[25]

Major Contributions

- **Rooted Cosmopolitanism**: Philosophy combining universal moral obligations with respect for legitimate cultural differences[27] [28]
- **Ethics of Identity**: Groundbreaking work challenging biological racialism while exploring how identities are both socially constructed and individually shaped[25]
- **Global Conversation**: Framework for engaging across differences without requiring consensus, emphasizing method over predetermined outcomes[26]
- **Cultural Critique**: Extensive work bridging African and Western philosophical traditions[25]

Relevance to AI Alignment

Appiah's cosmopolitanism provides crucial guidance for AI systems operating across diverse cultural contexts. His framework for respecting differences while maintaining universal ethical commitments offers models for how AI systems can navigate value pluralism without falling into relativism.

Fun Facts

- Writes mystery novels and poetry in addition to philosophy[26]
- His cosmopolitanism motto is "universality plus difference"[26]
- Advocates for global conversations without requiring everyone to become alike[26]
- Traces cosmopolitan ideals to ancient Greek Cynics and Stoics[28]

Onora O'Neill

Kantian Ethicist and Consent Theorist

Biography & Background

Onora O'Neill is an Irish philosopher born in 1941 who teaches at the University of Essex and is a cross-bench member of the House of Lords. She has made seminal contributions to Kantian ethics, bioethics, and political philosophy, particularly around issues of autonomy, trust, and consent.[29] [30] [31]

Major Contributions

- **Kantian Ethics Simplification**: Made Kant's complex moral theory accessible through focus on treating others as ends-in-themselves rather than mere means[32]
- **Consent Theory**: Developed sophisticated account of genuine versus spurious consent, emphasizing capacity for both consenting and dissenting[33]
- **Autonomy and Trust**: Revolutionary work linking Kantian autonomy to social trust rather than individualistic self-determination[29]

- **Applied Ethics**: Extensive work applying Kantian principles to bioethics, development ethics, and global justice[30]

Relevance to AI Alignment

O'Neill's work on consent and treating persons as ends provides fundamental constraints for AI systems that interact with humans. Her sophisticated analysis of what constitutes genuine consent versus manipulation offers crucial guidance for AI systems seeking human approval or cooperation.

Fun Facts

- Member of the House of Lords, bringing philosophical expertise to policy debates[31]
- Applied Kantian ethics to famine relief, comparing demands with utilitarian approaches[30]
- Emphasizes that autonomy is relational rather than individualistic[29]
- Her bioethics work addresses defensive medicine and cost-efficiency pressures[29]

Danielle Allen

Democratic Theorist and Civic Education Advocate

Biography & Background

Danielle Allen is the James Bryant Conant University Professor at Harvard University and Director of the Allen Lab for Democracy Renovation. She received the Library of Congress's Kluge Prize in 2020 for her scholarship in political theory and commitment to improving democratic practice and civics education.[34] [35]

Personality & Character

Allen combines rigorous academic scholarship with practical democratic activism. She approaches democracy as both a theoretical framework and a lived practice requiring constant cultivation and renewal. Her work is characterized by both intellectual depth and urgent practical relevance.[36] [35]

Major Contributions

- **Democracy Renovation**: Comprehensive approach to revitalizing democratic institutions and civic education[37] [34]
- **Civic Education Revival**: Leading voice for restoring civics education as essential to democratic functioning[36] [37]
- **Democratic Knowledge Project**: Harvard lab identifying and teaching skills citizens need for effective democratic participation[36]
- **Bridgebuilding**: Framework for moving from "I" to "we" through collaborative problem-solving skills[36]

Relevance to AI Alignment

Allen's work on democratic stewardship and civic ethics provides essential frameworks for AI governance in democratic societies. Her emphasis on inclusive decision-making and collaborative problem-solving offers models for how AI systems should facilitate rather than replace democratic deliberation.

Fun Facts

- Traces her democratic commitment to ancestors who fought for women's suffrage and founded NAACP chapters[35]
- Argues democracy depends on "winners lead the process and incorporate the losers" rather than total obliteration[36]
- Believes Americans have lost the "owner's manual" to democracy[36]
- Emphasizes that democracy requires constant cultivation—"democratic institutions don't run on their own" [36]

David Chalmers

Consciousness Philosopher

Biography & Background

David Chalmers is an Australian philosopher specializing in philosophy of mind, best known for formulating the "hard problem of consciousness". He published his influential book "The Conscious Mind" in 1996 while still completing his PhD, admitting it was "far from perfect" as he had been "studying philosophy for only four years".[38] [39]

Major Contributions

- **Hard Problem of Consciousness**: Distinguished between "easy problems" (brain functions) and the "hard problem" (why there's subjective experience at all)[39] [38]
- **Property Dualism**: Argues consciousness is fundamental like mass or space, irreducible to physical processes[40] [38]
- **Information Integration**: Proposes consciousness might be based on information processing patterns[40]
- **Zombie Arguments**: Thought experiments about beings functionally identical to humans but lacking consciousness[38]

Relevance to AI Alignment

Chalmers' work provides crucial guidance for AI systems that might develop consciousness or need to recognize it in others. His phenomenological approach offers frameworks for AI systems to understand and respect the texture of conscious experience rather than treating minds as mere information processors.

Fun Facts

- Initially studied physics thinking it had tools to understand everything[41]
- Took a "very long diversion through computer science and AI" before focusing on consciousness[41]
- His 2017 TED talk "Your Brain Hallucinates Your Conscious Reality" has over 13 million views[42]
- Argues we're all "hallucinating all the time; when we agree about our hallucinations, we call it reality"[42]

Susan Schneider

Machine Consciousness Researcher

Biography & Background

Susan Schneider is a philosopher specializing in artificial intelligence and consciousness studies. She developed the AI Consciousness Test (ACT) in collaboration with astrophysicist Edwin Turner around 2017-2018, motivated by the crucial ethical challenge of determining whether AI systems might possess subjective experiences.[43]

Major Contributions

- **AI Consciousness Test (ACT)**: Behavioral test using specialized natural language interactions to probe AI consciousness[44] [43]
- **Chip Test**: Thought experiment about replacing brain regions with functionally identical microchips[43]
- **Wait and See Approach**: Advocates remaining open to AI consciousness possibility while acknowledging current limitations[45]
- **Institutional Ethics**: Work on safeguards for synthetic minds and ethical frameworks[43]

Relevance to AI Alignment

Schneider's consciousness detection frameworks provide essential tools for AI alignment by helping identify when AI systems might develop morally relevant experiences requiring protection and consideration.

Fun Facts

- Emphasizes AI consciousness isn't "lunacy" but requires careful empirical investigation[45]
- Stresses importance of "boxing in" AI during testing to prevent learned responses about consciousness[43]
- Argues conscious AI creation might be "ridiculously expensive" even if possible[45]

- Her work addresses both technical and ethical dimensions of potential machine consciousness[44]

Anil Seth

Consciousness Neuroscientist

Biography & Background

Anil Seth is a cognitive and computational neuroscientist at the University of Sussex, codirecting the Sackler Center for Consciousness Science and the Canadian Institute for Advanced Research's Brain, Mind, and Consciousness program. His intellectual journey began with childhood questions about existence—"Why am I me and not someone else?"—leading him from physics through psychology to AI and finally neuroscience.[41]

Major Contributions

- **Predictive Processing**: Theory that consciousness arises from brain's predictions about sensory input rather than passive reception[46] [42]
- **Controlled Hallucination**: Framework understanding perception as brain's best guess about reality[46] [42]
- **Embodied Consciousness**: Emphasis on how consciousness requires flesh-and-blood predictive machinery[41]
- **Real Problem of Consciousness**: Focus on mapping brain activity patterns to conscious experiences rather than solving "hard problem"[46]

Relevance to AI Alignment

Seth's embodied approach suggests AI consciousness and alignment require understanding biological constraints and affective systems, not just abstract computation. His work provides guidance for respecting the embodied nature of conscious experience in AI development.

Fun Facts

- His 2017 TED talk has over 13 million views[42]
- Author of "Being You: A New Science of Consciousness" (2021)[41]
- Argues minds require biological predictive machinery, not just algorithms[41]
- Co-created "Dreamachine" art installation that induces consciousness experiences[47]

Antonio Damasio

Emotion and Consciousness Theorist

Biography & Background

Antonio Damasio is a neuroscientist whose theories on emotion, feeling, and consciousness have been among the most influential in the field. He developed a three-layered theory based on Protoself, Core Consciousness, and Extended Consciousness, emphasizing the body's central role in conscious experience.[48] [49] [50]

Major Contributions

- **Somatic Marker Hypothesis**: Theory that emotions guide decision-making through bodily feedback[49]
- **Homeostatic Feelings**: Argument that consciousness arises from constant flow of feelings about body's internal state[50] [49]
- **Embodied Consciousness Theory**: Framework linking consciousness to body's interactions with environment[48]
- **Feeling of What Happens**: Influential book connecting emotions, feelings, and consciousness[51] [52]

Relevance to AI Alignment

Damasio's work suggests AI alignment requires understanding embodied emotional systems, not just abstract reasoning. His emphasis on homeostasis and bodily constraints provides guidance for grounding AI values in life-sustaining principles.

Fun Facts

- Worked with patient "David" who couldn't form memories but still showed preferences[52]
- Argues "consciousness begins with feeling, not thinking"[50]
- Emphasizes gut microbiome's role in consciousness[48]
- Believes emotions are essential to ethical thinking, not obstacles to it[51]

Abeba Birhane

Relational Ethics and AI Researcher

Biography & Background

Abeba Birhane is a PhD candidate in cognitive science at University College Dublin, studying "dynamic and reciprocal relationships between emerging technologies, personhood, and society". Her research draws on embodied cognitive science, complexity science, and African philosophical traditions.[53] [54] [55] [56]

Major Contributions

- **Relational Ethics**: Framework centering interconnectedness and communal relationships in AI ethics[57] [53]
- **Algorithmic Injustice**: Analysis of how AI systems disproportionately harm vulnerable communities[57]
- **Power-Aware AI**: Emphasis on understanding structural harms and designing refusal patterns[54] [53]
- **Decolonial AI**: Integration of African philosophical frameworks into AI ethics discourse[53] [54]

Relevance to AI Alignment

Birhane's relational approach provides essential perspective on how AI alignment must account for structural power dynamics and community relationships rather than focusing solely on individual preferences or abstract principles.

Fun Facts

- Contributor to Aeon Magazine on AI ethics and cognitive science[56]
- Argues people are "ambiguous," "constantly changing," and "inherently non-determinable" [56]
- Questions how ML systems can predict behavior of inherently unpredictable humans[56]
- Emphasizes Ubuntu principle: "a person is a person through other people" [53]

Ruha Benjamin

Technology and Race Theorist

Biography & Background

Ruha Benjamin is a sociologist at Princeton University who received her PhD from UC Berkeley and completed postdoctoral fellowships at UCLA and Harvard. Born and raised partly in the Crenshaw neighborhood of Los Angeles, her childhood experiences with surveillance and police helicopters shaped her later analysis of technology's role in perpetuating racial inequality.[58] [59]

Major Contributions

- **Race After Technology** (2019): Groundbreaking analysis of how AI and algorithms reinforce white supremacy[60] [58]
- **New Jim Code**: Framework showing how technology encodes discrimination while appearing neutral[61] [58]
- **Abolitionist Tools**: Practical approaches for challenging discriminatory technology design[58] [60]

- **Structural Harm Analysis**: Methods for identifying and addressing systemic bias in AI systems[61] [58]

Relevance to AI Alignment

Benjamin's work demonstrates how AI alignment must address structural inequalities and power dynamics, not just technical optimization. Her analysis shows how seemingly neutral systems can perpetuate and amplify existing injustices.

Fun Facts

- Received President's Award for Distinguished Teaching at Princeton in 2017[59]
- Connects Movement for Black Lives to anti-eugenics activism[58]
- Links Obama-era drone warfare to domestic police militarization[58]
- Argues race itself functions as a "kind of technology" for stratifying society[60]

Sabelo Mhlambi

Ubuntu Philosophy and AI Ethics

Biography & Background

Sabelo Mhlambi is a researcher working to integrate Ubuntu philosophy into AI ethics and governance frameworks. His work focuses on bringing Sub-Saharan African relational philosophies into global AI ethics discourse, challenging Western individualistic approaches.[62] [63] [64]

Major Contributions

- **Ubuntu AI Framework**: Comprehensive approach applying African philosophy to AI governance[65] [64] [62]
- **Relational Personhood**: Theory that personhood emerges through relationships with others rather than individual rationality[63] [65]
- **From Rationality to Relationality**: Influential paper shifting focus from individual to communal ethics in AI[64] [63]
- **Inclusive AI Ethics**: Framework ensuring historically marginalized communities are centered in AI development[62] [64]

Relevance to AI Alignment

Mhlambi's Ubuntu approach offers alternative to Western individualistic frameworks, emphasizing how AI systems should support community relationships and collective flourishing rather than optimizing for isolated individuals.

Fun Facts

- Ubuntu principle: "Umu-Ntu ngumu-Ntu nga ba-Ntu" (A person is a person through other persons)[65]
- Argues Western rationality-based personhood excludes many African perspectives[63]
- Emphasizes communal relations based on "generosity, hospitality, compassion, and friendliness"[64]
- Notes African resources now power Fourth Industrial Revolution as they once powered slavery[64]

Nanjala Nyabola

Digital Rights and African Technology Policy

Biography & Background

Nanjala Nyabola is a Kenyan political analyst, journalist, and CIGI fellow based in Nairobi. She holds an M.Sc. in African studies and forced migration studies from Oxford and a J.D. from Harvard Law School. Her work focuses on the intersection of technology, media, and society from African perspectives.[66] [67] [68]

Major Contributions

- **Digital Democracy, Analogue Politics** (2018): Groundbreaking analysis of internet's impact on Kenyan politics[67] [68]
- **African Feminist Technology Philosophy**: Developing regulatory frameworks for digital technology from African women's perspectives[66]
- **Decolonial Tech Critique**: Challenging myths of "leapfrogging" and technological solutions imposed from outside[66]
- **Infrastructure Analysis**: Exposing the material reality behind "cloud" computing and its environmental costs[66]

Relevance to AI Alignment

Nyabola's work demonstrates how AI alignment must account for Global South perspectives and the material realities of technology infrastructure. Her emphasis on African agency challenges top-down approaches to AI governance.

Fun Facts

- Author of "Travelling While Black: Essays Inspired by a Life on the Move" (2020)[68]
- Argues "there is no such thing as it is in the cloud"—everything requires physical infrastructure[66]
- Emphasizes "water bearers"—rural women whose realities are absent from AI design[66]

- Questions why African futures are "outsourced to someone else's imagination"[66]
- Calls for AI development starting with "the lives we actually live"[66]

This diverse group represents a crucial convergence of philosophical traditions, technical expertise, and ethical frameworks essential for developing AI systems that truly serve human flourishing across all communities and contexts.

⁂

1. https://en.wikipedia.org/wiki/Connor_Leahy
2. https://sifted.eu/articles/connor-leahy-ai-alignment
3. https://www.lesswrong.com/posts/TjyyngWFYvQWPpNNj/eliezer-yudkowsky-is-frequently-confidently-egregiously
4. https://theinsideview.ai/connor2
5. https://uk.linkedin.com/in/connor-j-leahy
6. https://www.youtube.com/watch?v=_8q9bjNHeSo
7. https://time.com/collection/time100-ai/6309037/eliezer-yudkowsky/
8. https://en.wikipedia.org/wiki/Eliezer_Yudkowsky
9. https://www.lesswrong.com/w/machine-intelligence-research-institute-miri
10. https://www.surfthewave.ai/eliezer-yudkowsky-pioneering-ai-alignment-and-rationality/
11. https://www.alignmentforum.org/w/eliezer-yudkowsky
12. https://actlearnlead.com/2013/11/10/twelve-virtues-of-rationality-by-eliezer-yudkowsky/
13. https://ethics.org.au/big-thinker-martha-nussbaum/
14. https://en.wikipedia.org/wiki/Martha_Nussbaum
15. https://aeon.co/essays/why-reading-martha-c-nussbaums-philosophy-is-pure-pleasure
16. https://www.britannica.com/biography/Martha-Nussbaum
17. https://iep.utm.edu/sen-cap/
18. https://www.newyorker.com/magazine/2016/07/25/martha-nussbaums-moral-philosophies
19. https://www.britannica.com/biography/Peter-Singer
20. https://www.animallaw.info/article/animal-rights-theory-and-utilitarianism-relative-normative-guidance
21. https://en.wikipedia.org/wiki/Peter_Singer
22. https://berggruen.org/news/in-conversation-with-peter-singer-utilitarianism-avoiding-common-misunderstandings-and
23. https://repub.eur.nl/pub/50674/metis_165963.pdf
24. https://base.socioeco.org/docs/developments_critiques_advances.pdf
25. https://philosophical.chat/philosophy/philosophers-and-their-philosophies/kwame-anthony-appiah/
26. https://hedgehogreview.com/issues/the-cosmopolitan-predicament/articles/making-sense-of-cosmopolitanism
27. https://greattransition.org/publication/cosmopolitan-impulse-entangled-world
28. https://appiah.net/books/cosmopolitanism/
29. https://pubmed.ncbi.nlm.nih.gov/26572718/

30. https://personal.lse.ac.uk/robert49/teaching/ph103/pdf/ONeill2004.pdf
31. https://assets.cambridge.org/97811070/35591/frontmatter/9781107035591_frontmatter.pdf
32. https://scholarblogs.emory.edu/millsonph115/2014/10/20/oneills-a-simplified-account-of-kantian-ethics/
33. https://jahlinmarceta.se/constructions-of-reasons-explorations-of-kants-practical-philosophy-by-onora-oneill/
34. https://news.nd.edu/news/danielle-allen-discusses-the-current-state-of-democracy-encourages-focus-on-state-government/
35. https://news.harvard.edu/gazette/story/2023/11/danielle-allens-prescription-for-democracy-citizens-who-step-up/
36. https://news.asu.edu/20190404-solutions-danielle-allen-rebuilding-civic-education-asu-scetl
37. https://technoskeptic.substack.com/p/danielle-allen-on-civics-education
38. https://en.wikipedia.org/wiki/The_Conscious_Mind
39. https://en.wikipedia.org/wiki/Hard_problem_of_consciousness
40. https://www.informationphilosopher.com/solutions/philosophers/chalmers/
41. https://nautil.us/is-consciousness-more-like-chess-or-the-weather-304796/
42. https://www.youtube.com/watch?v=lyu7v7nWzfo
43. https://airights.net/conscious-or-coded-how-sci-fi-ais-perform-on-dr-susan-schneiders-act
44. https://faculty.ucr.edu/~eschwitz/SchwitzAbs/SchneiderCrit.htm
45. https://iai.tv/articles/machine-consciousness-is-not-lunacy-auid-2364
46. https://worldliteraturetoday.org/2022/march/being-you-new-science-consciousness-anil-seth
47. https://www.skepticspath.org/podcast/free-will-consciousness-reality-with-dr-anil-seth/
48. https://en.wikipedia.org/wiki/Damasio's_theory_of_consciousness
49. https://www.psychologytoday.com/us/blog/finding-purpose/202409/emotions-are-key-to-understanding-consciousness
50. https://iai.tv/articles/consciousness-begins-with-feelings-hanna-damasio-auid-2462
51. https://oaj.fupress.net/index.php/pam/article/download/7060/7040/6985
52. https://gustavus.edu/neuroscience/brainbooks/The Feeling of What Happens Body and emotion in the making of consciousness.pdf
53. https://ojs.stanford.edu/ojs/index.php/grace/article/view/2581
54. https://ojs.stanford.edu/ojs/index.php/grace/article/download/2581/1525/8975
55. https://ucdcs-research.ucd.ie/phd-student/abeba-birhane-0/
56. https://www.etcgroup.org/content/5-algorithmic-colonisation-abeba-birhane
57. https://pubmed.ncbi.nlm.nih.gov/33659914/
58. https://www.thenation.com/article/culture/ruha-benjamin-race-after-technology-book-review/
59. https://www.goodreads.com/en/book/show/42527493-race-after-technology
60. https://www.managementboek.nl/boek/9781509526406/race-after-technology-ruha-benjamin
61. https://www.bol.com/nl/nl/f/race-after-technology/9200000103431646/
62. https://www.accel.ai/anthology/2023/4/21/comparative-analysis-ubuntu-collectivism-versus-western-ethics-in-ai-development

63. https://montrealethics.ai/research-summary-from-rationality-to-relationality-ubuntu-as-an-ethical-human-rights-framework-for-artificial-intelligence-governance/
64. https://core.ac.uk/download/582653536.pdf
65. https://www.accel.ai/anthology/2022/12/30/introduction-to-ubuntu-collectivism-theory-applied-to-ai-ethics
66. https://rpublc.com/vol9-no3/nanjala-nyabola-artificial-intelligence/
67. https://appropriatingtechnology.org/interview-with-nanjala-nyabola
68. https://www.cigionline.org/people/nanjala-nyabola/
69. https://uk.linkedin.com/in/connor-leahy-
70. https://www.olympics.com.au/olympians/conor-leahy/
71. https://www.youtube.com/watch?v=OUjnVeydhCM
72. https://www.su.org/resources/ai-alignment-future
73. https://www.youtube.com/watch?v=XGK4o0A02xg
74. https://en.wikipedia.org/wiki/Conor_Leahy
75. https://www.youtube.com/watch?v=k6M_ScSBF6A
76. https://theorg.com/org/conjecture/org-chart/connor-leahy
77. https://www.aspi.org.au/podcasts/ai-expert-connor-leahy-on-superintelligence-and-the-threat-of-human-extinction/
78. https://www.lesswrong.com/posts/CpvyhFy9WvCNsifkY/discussion-with-eliezer-yudkowsky-on-agi-interventions
79. https://www.lesswrong.com/posts/evRNWBqJbyanLfp7s/miri-s-death-with-dignity-in-60-seconds
80. https://www.reddit.com/r/lexfridman/comments/12vq3zi/what_is_the_ai_alignment_problem_from_eliezer/
81. https://intelligence.org/files/AlignmentHardStart.pdf
82. https://www.youtube.com/watch?v=dWmZqaFB0sM
83. https://en.wikipedia.org/wiki/Capability_approach
84. https://nautil.us/this-test-for-machine-consciousness-has-an-audience-problem-237652/
85. https://www.bol.com/nl/nl/f/the-conscious-mind/39305669/
86. https://philarchive.org/archive/BRODCO
87. http://www.few.vu.nl/~wai/Papers/CC07damasio.pdf
88. https://www.bol.com/nl/nl/p/race-after-technology/9200000116367087/