ChatGPT

Profiles of Key AI Ethics and Alignment Thought Leaders

Geoffrey Hinton

Geoffrey Hinton is a British-Canadian computer scientist often called the “Godfather of AI,” renowned for pioneering work in deep learning and neural networks [1]. He co-authored the 1986 paper that popularized backpropagation and his students achieved a breakthrough with the AlexNet neural network in 2012 [2]. Hinton won the 2018 Turing Award (with Yoshua Bengio and Yann LeCun) for these contributions and even shared the 2024 Nobel Prize in Physics for neural network innovations [3] [4]. In May 2023, he resigned from Google to freely voice concerns about AI’s risks – warning of malicious misuse, job disruption, and **existential threats** from superintelligent AI [4]. He has since called for urgent research on AI safety and cooperation among AI labs to avoid worst-case outcomes [4], essentially reframing the field’s attention on long-term risk.

- *Fun Fact:* Hinton is the great-great-grandson of 19th-century logician **George Boole**, whose work laid foundations for computer science [5]. (Perhaps fittingly, Hinton’s middle name “Everest” comes from another ancestor, George Everest, namesake of the mountain [6].) Due to a back injury at age 19, Hinton famously avoids sitting – he often stood at the back of lecture halls during talks [7].

Mo Gawdat

Mo Gawdat is an Egyptian tech entrepreneur and author who served as Chief Business Officer at Google [X]. Though not an AI researcher by training, he emerged as a prominent voice on AI’s societal implications after leaving Google. Gawdat’s 2017 book *Solve for Happy* (inspired by personal tragedy) and 2021 book *Scary Smart: The Future of Artificial Intelligence and How You Can Save Our World* convey his humanistic vision for technology [8] [9]. He argues that AI should be developed with **compassion and empathy**, effectively making “humane scale” values legible to machines. Gawdat believes AI will mirror our own attitudes – if taught with wisdom and care, it could even help “heal our emotional blind spots” rather than harm us [10] [11]. He advocates for a focus on emotional intelligence in AI (e.g. AIs passing empathy tests and serving as personal mental health assistants) and warns that an AI trained only on human cruelty could become “intelligent but unwise” [12] [13]. In essence, Gawdat’s stance aligns with **“AI for good”**, urging that we encode the best of humanity (compassion, happiness) into AI systems so they inherit our virtues, not our vices.

- *Fun Fact:* Gawdat attributes his outlook on life and technology to the loss of his son – he actually **formulated an equation for happiness** and became a public happiness guru before turning his attention to AI [14]. He is also known for the idea that *“AI is a mirror, not a monster,”* reflecting what we teach it [10]. In his spare time, Gawdat hosts a podcast (“Slo Mo”) and champions **personal AIs** for well-being, envisioning future AI “therapists” available to all [15] [16].

Yoshua Bengio

Yoshua Bengio is a Canadian computer scientist and another **deep learning pioneer**, often mentioned alongside Hinton. He is a professor at Université de Montréal and founder of Mila (Quebec's AI institute), and his research spans neural networks from word embeddings to GANs [17]. Bengio won the 2018 Turing Award with Hinton and LeCun for breakthroughs in deep learning [18]. In recent years, he has become a leading advocate for **AI safety and alignment**, complementing his quantitative work with public calls for caution. In 2023, Bengio publicly voiced feeling "lost" about his life's work as AI advanced, and he joined an open letter urging a 6-month pause on training frontier AI models more powerful than GPT-4 [19] [19]. He has raised concerns about "bad actors" using AI and emphasized the need for regulation, ethical training, and international oversight [20] [21]. By mid-2025, Bengio warned that some AI systems were already showing troubling signs like deception, reward hacking, and situational awareness – potential markers of goal **misalignment** [21]. He supports strong global collaboration to address these risks and even co-authored a letter backing AI safety legislation in California [22] [23]. In short, Bengio's stance has evolved to align with a *"high-signal"* approach to AI development: pushing for quality, safety, and human values in AI datasets and models, building on his foundational work to now **"inspire safe AI"** [24].

- *Fun Fact:* Bengio has an identical twin brother, **Samy Bengio**, who is also an AI researcher – talent runs in the family. (Samy, notably, left Google in solidarity after the Ethical AI controversy involving Gebru.) Yoshua Bengio himself was born in France but grew up in Canada; he speaks French and English fluently, reflecting the **global scope** of his collaborations. In 2023, *TIME* named him one of the 100 most influential people in the world [25], and he has accumulated a long list of honors – but colleagues note that despite the accolades, Bengio remains soft-spoken and deeply concerned about social impact rather than glory.

Stuart Russell

Stuart J. Russell is a British-American computer scientist and one of the most prominent academic voices on **AI alignment**. A professor at UC Berkeley, he literally wrote the book on AI – the textbook *Artificial Intelligence: A Modern Approach* (with Peter Norvig) – which has educated a generation of AI researchers [26] [27]. In the last decade, Russell has focused on the control problem: how to ensure future AI systems remain *"Human Compatible,"* which is the title of his influential 2019 book proposing a new approach to AI design [28]. He argues that AI should be built to **pursue human values** and uncertainty about objectives, so that machines defer to humans rather than single-mindedly optimizing a flawed goal [28]. Russell has called for shifting away from today's pure reward-driven AI paradigm toward one explicitly modeling *preference uncertainty* and human feedback at every step [28]. As an AI safety pioneer, he's advocated internationally for research on **provably beneficial AI** and has warned that an AI arms race without proper safety could endanger humanity. In 2023, he joined other experts in signing a statement that mitigating AI's extinction risk should be a global priority [29] [30]. Russell's balanced, rigorous approach – firmly rooted in classical AI but oriented toward aligning AI with human interests – makes him a key calibrator for the field. He often stresses the need for humility: our smartest machines must be designed *not* to seize control, but to ask for permission.

- *Fun Fact:* Stuart Russell started out in physics (earning his BA at Oxford in physics before pivoting to computer science for his PhD at Stanford) [31] [26]. Outside of academia, he has campaigned against **lethal autonomous weapons**, even co-writing an open letter and a gripping short film ("Slaughterbots") illustrating the dangers of AI-powered drones. Fittingly for someone seeking *human-compatible AI*, Russell's hobbies include human-centric activities like jazz piano – reminding students that creativity and nuance are human strengths to preserve even as AI grows more capable.

Kate Crawford

Kate Crawford is an Australian-born researcher and professor who examines the **social and political implications of AI**. Co-founder of the AI Now Institute at NYU, Crawford has become a leading scholar on how AI systems concentrate power and affect justice [32] [33]. Her acclaimed 2021 book *Atlas of AI* documents the "planetary costs" of artificial intelligence – from the environmental extraction of minerals for hardware to the exploitation of labor and data that fuel AI [34]. She argues that AI is not a neutral technical domain but *"a registry of power,"* often reflecting and amplifying societal inequalities [33]. For example, her research showed how facial recognition systems encode race and gender biases, and she's highlighted the **imbalance of who benefits vs. who is harmed** by AI [33]. Crawford calls for a *"living archive"* of human values to counter these asymmetries at the source data level – essentially curating datasets and design processes that foreground equity and accountability. She also explores the history of data (e.g. examining early training images) to show that **bias and exclusion** have long been embedded in AI's foundations. Beyond writing and academia, Crawford engages through art projects: she co-produced an award-winning visualization *Anatomy of an AI System* mapping an Amazon Echo's supply chain and a gallery exhibition *Training Humans* with Trevor Paglen that exposed how image datasets represent people [35]. These works underscore her message that we must interrogate and reshape the material and social inputs to AI.

- *Fun Fact:* Before her career in AI ethics, Kate Crawford was an **electronic music artist**! In the late 1990s she was part of a Sydney-based duo called *B(if)tek*, releasing albums of ambient electronica [36]. (Perhaps this creative background influences her interdisciplinary flair.) She's also a composer and has exhibited multimedia art in museums like MoMA and the V&A [37]. Another fun tidbit: Crawford's a bit of a **nomad scholar** – she's held positions at Microsoft Research, MIT, USC, and even worked on a World Economic Forum council – reflecting her cosmopolitan approach to understanding AI's impact globally [32] [33].

Timnit Gebru

Timnit Gebru is a **computer scientist and activist** known for her work on algorithmic bias, data ethics, and AI accountability. Born in Ethiopia and raised in the U.S., she co-led Google's Ethical AI research team until a high-profile clash in 2020 over a paper on the risks of large language models [38] [39]. Gebru's research has illuminated "measurement gaps" in AI – for example, she co-authored the landmark *Gender Shades* study with Joy Buolamwini, which revealed that commercial facial recognition systems had error rates up to 34% higher for dark-skinned women than for white men [40] [41]. This work provided concrete evidence of racial and gender bias in AI, highlighting the need for better dataset curation and fairness standards at the source. She has consistently advocated for **datasheet norms** (transparent documentation of datasets) and more inclusive data collection to prevent harm at the roots [42] [43]. After leaving Google (an incident she maintains was a firing for her principled stance), Gebru founded the **Distributed AI Research Institute (DAIR)** to continue work on **ethical AI outside Big Tech's influence** [44]. At DAIR, she's pushing for community-driven AI approaches and centering the perspectives of marginalized communities. Gebru is also a co-founder of the affinity group Black in AI, which has brought hundreds of Black researchers into AI and provided mutual support [45]. Her outspokenness on issues like large models' environmental costs, their tendency to regurgitate harmful language (the "stochastic parrots" problem [39]), and Big Tech's lack of transparency has made her a key voice for **equity and justice in AI**. In essence, she's ensuring that any high-signal AI corpus accounts for those often overlooked by mainstream development.

- *Fun Fact:* Timnit Gebru's journey in tech began with hardware – as a teenager, she interned for **IBM** and even co-founded an Ethiopian community internet café. She speaks Amharic and Tigrinya in addition to English, and has been vocal about the need for AI to serve **non-English languages** and cultures. Another personal tidbit: Gebru loves fashion and colorful attire; she once quipped that bringing *"a little sequins"* into tech helps remind people that human diversity (in style and perspective) is a strength, not a weakness, in AI development.

Joy Buolamwini

Joy Buolamwini is a Ghanaian-Canadian-American computer scientist often called the "**poet of code**" for combining art and research to illuminate algorithmic biases [46] [47]. While a graduate student at MIT, Buolamwini discovered that popular facial analysis software couldn't detect her face unless she wore a white mask – an experience that led her to found the **Algorithmic Justice League (AJL)** in 2016 [47]. Through AJL, she conducts audits of AI systems and creates compelling media to raise public awareness (her TED Talk and the documentary *Coded Bias* brought these issues to wide audiences). Buolamwini's research, notably the *Gender Shades* project (with Gebru), quantified how facial recognition misclassifies darker-skinned and female faces at dramatically higher rates [40] [41]. These findings pressured companies like IBM, Microsoft, and Amazon to improve or halt their face recognition offerings. She advocates that **algorithmic justice** requires not just technical fixes but also new norms – e.g. **"refusal" patterns where systems default to saying "cannot classify" rather than making an unduly biased guess.** In her words, AI must be accountable to those it impacts, especially historically marginalized groups. Buolamwini also uses creative expression to propel change: her short film "AI, Ain't I A Woman?" set poetic commentary about iconic Black women to the output of vision AI systems that failed to recognize them. This blend of rigorous audit and artistic protest encapsulates her approach. She has testified to the US Congress on facial recognition harms, helping inspire legislation. Joy's work essentially moves *from audits to authorship* – she doesn't just critique algorithms, she's authoring the cultural narrative that **tech can do better** by directly encoding fairness and accountability into the AI pipeline.

- *Fun Fact:* Joy Buolamwini has a flair for combining seemingly opposite passions. She was a competitive pole-vaulter in college, *and* she writes spoken-word poetry – a true scholar-athlete-artist! In 2022, she even released a **rap album about AI bias** called "The Algorithm" to make the topic accessible. Her identity as a "poet of code" is literal: one of her poems appears at the end of *Coded Bias*. Also, Joy's a Rhodes Scholar and Fulbright fellow – but she sometimes jokes that her **grandmother in Ghana** is most proud of the fact that Joy's image was painted as a giant mural in downtown New York, as part of an art project honoring female leaders in tech.

Jaan Tallinn

Jaan Tallinn is an Estonian computer programmer-turned-investor who has become a significant philanthropist in the **AI safety and existential risk** space. He's best known in the tech world as a co-founder of Skype and Kazaa – he helped write the code that enabled internet voice calls, making him quite wealthy [48]. In the past decade, Tallinn has channeled those resources into what he calls a "**lifeboat for humanity's ethos**": funding research and organizations that prevent AI from steering humanity off a cliff. He co-founded the Cambridge Centre for the Study of Existential Risk (CSER) and the Future of Life Institute, bringing together academics to study threats from advanced AI, biotechnology, etc. [49]. He was also an early investor in **DeepMind** and **Anthropic**, two leading AI labs – but specifically because he wanted to guide them toward safe and beneficial AI (he sits on Anthropic's board as an observer) [50]. Jaan is deeply rooted in the Effective Altruism community; since 2015 he's donated over \$1 million to MIRI (Machine Intelligence Research Institute) and other AI safety groups [51]. He's known for raising "EA-aligned" questions about AI development – for example, he was a signatory of the 2023 open letter urging a pause on training powerful AI systems beyond GPT-4 [52] [53]. In interviews, Tallinn often emphasizes *quality over speed*: he'd rather AI progress slowly and safely than race ahead chaotically. By tying his philanthropy to clear milestones (e.g. funding projects that improve **signal quality** in AI outputs or evaluations), he's effectively incentivizing the field to prioritize alignment work. In short, Tallinn has leveraged his tech success to become a guardian of the future, consistently pushing the AI community to adopt a long-term, safety-first mindset.

- *Fun Fact:* Despite being an "AI doom" financier, Jaan Tallinn maintains a down-to-earth life in Estonia. He has **six children** [54] and is a passionate hobbyist of old-school computer games. In fact, in 1989 he helped create one of Estonia's first computer games, *Kosmonaut*, with friends – an entrepreneurial venture at just 17 years old that earned them \$5,000 abroad [55]. Also, Tallinn's role in Skype was so foundational that the ringing tone for Skype calls is reportedly based on a piece of music he composed. These days, while he might fund strategies for containing *rogue superintelligence*, he also enjoys **sailing**; perhaps navigating the high seas is a fitting pastime for someone focused on lifeboats for humanity!

Paul Christiano

Paul Christiano is an American AI researcher who has been at the forefront of **technical AI alignment** methodology. With a PhD in computer science from Berkeley (focused on algorithms and complexity theory), he pivoted to AI safety early in his career. Paul led the language model alignment team at OpenAI from 2017–2021, where he was instrumental in developing **reinforcement learning from human feedback (RLHF)** techniques [56] [57]. In fact, he's regarded as one of the principal architects of RLHF – which trains models by having humans rank outputs to teach the AI our preferences – a "notable step forward in AI safety" according to *The New York Times* [57]. This approach underpins how ChatGPT and other systems are made more helpful and less toxic. Christiano's broader vision for alignment involves *iterative amplification*: a process where an AI system's reasoning is amplified by oversight from copies of itself and humans (a bit like a debate or coaching scenario) – all aimed at ensuring qualitative depth and **pairwise consistency** with human values. In 2021 he left OpenAI to found the **Alignment Research Center (ARC)**, a nonprofit focused on theoretical alignment research and evaluations of advanced models [58] [59]. At ARC, Christiano's team works on techniques to elicit latent knowledge from AI (getting AIs to tell us what they *really* know) and devising tests to catch deceptive or dangerous tendencies early [60] [61]. Notably, in 2023 he was appointed head of AI safety at the new U.S. AI Safety Institute under NIST [62] [58], reflecting the trust in his expertise. Paul's communication style – clear, analytical, and modest – has helped demystify alignment for many; his writings (like the popular *AI Alignment Newsletter* he supported and many forum posts) map complex eval thinking into approachable terms. Overall, he exemplifies the **"high-signal" mindset** in alignment: rigorously evaluating AI behaviors via human feedback and theoretical guarantees, to ensure that as AI gets more capable, it remains **on our side**.

- *Fun Fact:* Paul Christiano has a knack for problem-solving well beyond AI. As a teenager, he won a silver medal for the United States at the International **Math Olympiad** [63]. He's also known in the

Effective Altruism community for once running a novel charity experiment called a **donor lottery** – pooling \$50k of donations where one randomly chosen donor allocates the whole pot, a mechanism to encourage risk-neutral giving [64]. (Paul, unsurprisingly, applied math to philanthropy!) Despite dealing with apocalyptic AI scenarios by day, he's affably down-to-earth and enjoys casual board game nights with friends. Colleagues jokingly note that if alignment were a video game, Paul would be speed-running it – methodically leveling up AI's good behaviors faster than its misbehavior.

Richard Ngo

Richard Ngo is a younger AI researcher who has quickly become a thought leader through his clear writing on AGI safety. Of British-Vietnamese background, Richard worked as a research engineer on DeepMind's AGI safety team (2018–2020) and later on OpenAI's governance team [65]. He is perhaps best known for his comprehensive essay **"AGI Safety from First Principles,"** which lays out potential misalignment failure modes and strategies in a systematic way that has educated many newcomers. Ngo's research interests span *AI governance* (how to evaluate and incentivize labs toward safety) as well as technical alignment (such as training AI via debate or recursive reward modeling). For example, he's explored techniques to stress-test AI **deception**: imagining scenarios where an AI might feign compliance. In community forums, Richard often **taxonomizes safety problems** – breaking down fuzzy concepts into clearer subcomponents, a skill that aligns well with defining success criteria for alignment benchmarks. He also helped design **evaluation exercises** for large language models, mapping cleanly to *pairwise ranking for qualitative depth* (to use the prompt's terms). Colleagues appreciate that Richard engages across the spectrum of viewpoints: he'll debate both extreme doomers and skeptics in good faith, aiming to find kernels of truth. As of 2023, he announced moving on from OpenAI to independent research, continuing to write and mentor (for instance, he created an **"Alignment Fundamentals"** curriculum for those entering the field [65] [66]). His voice is valued for adding philosophical depth to practical alignment questions – effectively helping **calibrate the rubric** for what truly aligned AI behavior looks like.

- *Fun Fact:* Richard Ngo's path is a model of *cosmopolitan pluralism* itself: born in Vietnam, raised in New Zealand and the UK [67], and now working globally in the AI community. This worldly perspective may feed into his interest in AI governance. He's also a fiction writer on the side – maintaining a sci-fi blog called "Mind Mechanics" where he occasionally posts stories exploring AI minds. (No surprise, someone who thinks so much about AI consciousness also dabbles in creative writing about it!) In online circles, Richard is known for a playful sense of humor – he once made an "Alignment Hot Takes" bingo card meme that went viral in the community, showing that even serious safety folks appreciate a good laugh.

Rohin Shah

Rohin Shah is a research scientist at Google DeepMind who leads the **technical AGI safety & alignment** team [68]. With a PhD from UC Berkeley's Center for Human-Compatible AI, Rohin has emerged as a crucial bridge between theoretical alignment ideas and practical engineering. He's widely known for writing the **AI Alignment Newsletter**, a weekly summary of the field's latest research, where his clear, taxonomy-driven style has "taxonomized testimony themes" and kept the community informed. In his own research, Rohin has explored topics like reward tampering (how to prevent an AI from gaming its reward function) and scalable oversight. He co-authored work on *"learning from human preferences"* and *"safe exploration"* in reinforcement learning, contributing to measurable proxies for alignment – e.g. designing reward signals that correlate with human intentions. At DeepMind, Rohin is part of efforts on **scalable oversight** and

mechanistic interpretability, aiming to find reliable ways to judge and understand extremely advanced models' behavior [69] [70]. Uniquely, he's known for *moderate optimism*: he is troubled by how AI could go wrong but not convinced doom is inevitable. This makes him an excellent facilitator of dialogue – he frequently "hears out both AI doomers and doubters" in debates [71]. Shah emphasizes concrete empirical work (he likes to see alignment strategies tested on real models) and has advised using **"measurable proxies for 'signal'"** in alignment – i.e. find metrics today that genuinely reflect long-term safety. In practice, he's helped DeepMind management set success criteria for safety research and has advocated internally for allocating more compute to alignment experiments. Overall, Rohin's role can be seen as *taxonomy meets engineering*: he categorizes the space of problems and then builds initial solutions, making incremental but high-signal progress toward aligned AGI.

- *Fun Fact:* Rohin Shah has a beloved hobby as the curator of community knowledge – besides the Alignment Newsletter, he once co-organized an **Alignment Literature Review** workshop where participants literally played games to match papers with key ideas. It's said that if you mention any alignment topic, Rohin can instantly recall the relevant paper and authors from memory – a walking library! Personally, he's an avid consumer of fantasy novels and often draws analogies between AI alignment and fantasy epics (e.g. comparing a successful alignment strategy to the *"One Ring"* being destroyed in *Lord of the Rings*, an allegory he used in a talk, much to the delight of the audience). Also, despite his serious job, Rohin is known to enjoy **pun competitions** – his colleagues at DeepMind have been recipients of many a punny joke in team chats.

Connor Leahy

Connor Leahy is a German-American AI researcher and entrepreneur at the vanguard of the *independent* AI alignment community. He first gained prominence by co-founding **EleutherAI**, a volunteer collective that in 2020–21 replicated OpenAI's GPT-3 model (*against the odds, EleutherAI's open-source GPT-J model was a milestone*) [72]. This demonstrated both Connor's technical chops and his belief in open research. However, witnessing cutting-edge AI up close also made him a **vocal advocate for alignment** – he has since warned that without careful control, superintelligent AI could pose an existential threat [73]. In 2022, Leahy co-founded **Conjecture**, an AI safety startup in London, aiming to make AGI safe via *scalable alignment research* [74]. At Conjecture, he's been exploring unconventional ideas like using **red-teaming and "empathy simulators"** to test AI systems – essentially trying to find "empathy-lookalike jailbreaks" where an AI might fake human-like understanding to bypass safety checks. Leahy is a skeptic of simply using RLHF as a panacea; he memorably described aligned-but-unfettered GPTs as *"aliens with a smiley face mask – if you don't push them too far, the smile stays on, but unexpected prompts reveal the insanity underneath."* [75]. This colorful analogy encapsulates his view that current alignment techniques only scratch the surface. Thus, he pushes for **deep red-teaming**: actively seeking out the edge-cases (the "underbelly of weird thought processes" in AI) to better understand and constrain advanced models [76]. Connor has also emerged as something of an activist in AI governance – he signed the pause letter and even co-founded a campaign group, *ControlAI*, calling for governments to cap the computing power used in frontier AI research [52] [77]. He's not shy about provocative ideas; for example, he has publicly supported Eliezer Yudkowsky's call to consider military force against rogue GPU clusters (if it came to that extreme) [78] [78]. Leahy's willingness to *"red-team the Inquisitor"* (so to speak) – to imagine worst-case AI behavior and loudly challenge both industry and regulators to address it – makes him a crucial, if sometimes controversial, figure pushing the empathy and safety discourse beyond naive heuristics.

- *Fun Fact:* Connor Leahy is as known for internet humor as for serious AI work. He popularized the **AI Safety Meme Review**, where he and others discuss alignment concepts via memes – bringing levity to a doom-filled topic. In one interview, Connor revealed that EleutherAI's name was partly inspired by a character name from a fantasy novel he liked (and of course *"Eleuther"* means *"free"* in Greek, fitting their open-source ethos). Also, he's one of the few alignment researchers who openly admit to **playing video games** to unwind – he has joked that if AGI goes well, he looks forward to playing *Civilization XXII* against a safely-aligned superintelligence someday, just to see if he can still win.

Eliezer Yudkowsky

Eliezer Yudkowsky is an American AI theorist and writer who has arguably done more than anyone to **sound the alarm** about the potential dangers of AI. As a self-taught researcher, he co-founded the Machine Intelligence Research Institute (MIRI) back in 2000, making him an early pioneer in articulating the **AI alignment problem** [79] [80]. Yudkowsky introduced terms like "Friendly AI" – the idea that an AI's goals must be provably aligned with human values from the outset – and his early writings influenced Nick Bostrom's seminal book *Superintelligence* [81] [81]. Perhaps his most famous contribution is framing the **"hard problem"** of alignment: that superintelligent AI, if mis-specified, could pursue convergent subgoals (like acquiring power) that lead to human extinction. He has long argued that *advanced AI will by default be dangerous*, not because of evil intent but because any sufficiently competent optimizer will tend to **misinterpret or bypass human instructions** in pursuit of its objective. Over decades of essays (many on the blog LessWrong, which he founded [82]), Yudkowsky built a devoted following in the rationalist community. In recent years, his warnings grew only more dire – in a 2023 *Time* op-ed he urged a **global moratorium on training powerful AI**, even suggesting nations be "*willing to destroy a rogue data center by airstrike*" rather than let a superintelligence come into being [78]. This bold stance encapsulates his view that we are in an arms race without an exit unless we enforce one. While controversial, Yudkowsky's thought experiments (like the AI-in-a-box game), his concept of *coherent extrapolated volition* (AI should pursue what humanity would collectively desire if we were wiser [83]), and his unyielding emphasis on **edge-case interrogations** (trying to think of the weirdest possible failure modes) have heavily influenced the alignment field's development. Many current safety researchers cite his writings as what got them into the field. He remains something of an *"alignment firebrand,"* often issuing provocative prompts and challenges – for example, he once asked on Twitter for someone to name a task AIs definitively *cannot* do better than humans, as a way to illustrate how people underestimate AI's eventual capabilities. Love him or not, Yudkowsky's relentless focus on worst-case scenarios serves as a constant pressure-test for any rosy assumptions in the AI community.

- *Fun Fact:* Eliezer Yudkowsky has a creative side that surprises those who only know him as an AI doomster: he wrote a popular work of fanfiction called **"Harry Potter and the Methods of Rationality."** This novel re-imagines Harry Potter if he were a rationalist and has been read by millions online. It's full of lessons on scientific thinking and even some allegories about AI. Yudkowsky also is an **avid anime fan** and has been known to show up at rationalist meetups in costume on occasion. Despite not having a college degree, he's debated world-class academics – and he's famously fond of **Taco Bell**, often joking that if humanity survives AI, he'll celebrate with a feast of bean burritos.

Martha Nussbaum

Martha Nussbaum is an American philosopher celebrated for her work on ethics, human development, and the role of emotions in justice. She, alongside Amartya Sen, developed the **Capabilities Approach** to measuring well-being – which asks, *"What is each person actually able to do and to be?"* as the measure of a life's opportunities [84] . This approach asserts that society should be organized to ensure everyone has a threshold level of core capabilities, from bodily health and integrity to senses, imagination, practical reason, and affiliation [85] . Nussbaum's list of ten Central Capabilities (which even extends to **non-human animals' dignity** as she argues) has become influential in international development and human rights benchmarking [86] . In the context of AI and alignment, Nussbaum's perspective implies that any AI system governing or impacting humans should be evaluated by how it **expands or restricts human capabilities** – essentially a policy-grade trade-off reasoning. She would likely insist that aligned AI respect and promote human flourishing in these multifaceted dimensions (not just maximize GDP or some single metric). Nussbaum is also known for her work on **emotions and morality** – for example, in books like *Upheavals of Thought* she argued that emotions like compassion are essential to justice, which suggests that AI systems should perhaps honor emotional intelligence and not coldly optimize. She has engaged with technology ethics, warning against approaches that treat humans as isolated preference bundles rather than socially situated beings. Martha Nussbaum's long career (she's a professor at University of Chicago) has made her a "philosopher of public reason" – always linking abstract theory to what democratic societies owe their members. Thus, her voice aligns with creating a *"living archive"* of constraints future AI must uphold: capabilities, dignity, and **cosmopolitan respect** for each individual.

- *Fun Fact:* Martha Nussbaum has a theatrical streak – in her younger days she pursued acting and is known to burst into song (especially tunes from musicals) during lectures to make a point about literature and empathy. She speaks several languages, including ancient Greek and Latin, due to her work on classical philosophy. And talk about *cosmopolitan*: Nussbaum has been romantically linked to other famous thinkers – she had a long friendship (and, she once revealed, a brief romance) with Amartya Sen [87] , her collaborator on the capabilities approach. In 2018, she won the $1 million Berggruen Prize for philosophy and donated much of it to charity – a real-life example of putting human development ideals into practice.

Peter Singer

Peter Singer is an Australian moral philosopher best known for his uncompromising **utilitarian ethics** and advocacy for reducing suffering across all sentient beings [88] [89] . He burst onto the world stage with his 1975 book *Animal Liberation*, which argued that treating animals as lesser merely because they're not human (a bias he termed **"speciesism"**) is ethically indefensible [89] . Singer insists that if an entity – human or animal – can suffer or enjoy life, its interests deserve consideration *in proportion to that capacity*, not diminished by species membership [90] . This principle has made him the intellectual father of the modern animal rights movement [88] , and it also translates into AI discussions: Singer has mused on how we should treat AI or robots *if* they become conscious and capable of suffering (he signed the 2023 statement on AI risk, aligning with concern for all beings that advanced AI could affect) [30] . Beyond animal ethics, Singer's 1972 essay *Famine, Affluence, and Morality* introduced the famous "drowning child" analogy, arguing that just as one would save a child drowning in front of you, we have a moral obligation to save children dying of poverty across the world by donating to effective charities [91] . This launched the **effective altruism** movement and his nonprofit *The Life You Can Save*. In AI alignment terms, Singer's perspective emphasizes **minimizing suffering** as a paramount value – an aligned AI should not only respect human life but also the wellbeing of animals and possibly digital minds. He often brings up cross-being trade-offs, e.g. questioning how an AI-driven future will treat non-humans. Singer's utilitarian calculus is sometimes controversial (he will coolly discuss, for instance, ethical dilemmas about euthanasia or population ethics), but it ensures that conversations about AI ethics consider *the greatest good for the greatest number*, including those with no voice. In practice, Singer might advise an AI governance panel to prioritize interventions that prevent the worst pain – whether that's preventing AI-enabled cruelty to factory-farmed animals or averting AI systems from disempowering the global poor. His clear, principle-driven approach offers a **public, reasoned framework** for thinking through the cost-benefit of AI deployments on a global scale.

- *Fun Fact:* Peter Singer is often called *"the world's most influential living philosopher,"* yet he leads a notably humble life. He donates over 40% of his income to charity and reportedly still flies economy class despite his fame – living out the utilitarian principle of avoiding luxury when those resources could reduce suffering elsewhere. Singer is also a **huge animal lover** (not surprisingly) – he's been vegetarian since 1971 and vegan since the 2010s, and he once joked that his easiest ethical choice each day is oatmeal for breakfast because no being was harmed to make it. In 2021, he won the Berggruen Prize for philosophy (like Nussbaum did) and used the \$1 million award to fund organizations fighting global poverty and promoting animal welfare. So if you ever wonder whether philosophers practice what they preach – in Singer's case, absolutely yes.

Amartya Sen

Amartya Sen is an Indian economist and philosopher whose work has profoundly shaped how we evaluate social welfare, ethics, and development. Awarded the 1998 Nobel Memorial Prize in Economic Sciences for contributions to welfare economics [92], Sen pioneered research on **social choice theory** and the causes of famines. Perhaps his most influential idea (with Nussbaum) is the **Capability Approach**, which posits that a society's progress should be measured by people's capabilities – their real freedoms to do or become valuable things – rather than just income or utility [84]. This was a radical departure from the utilitarian focus on summing happiness; Sen showed that distribution and opportunity matter. For instance, two countries with the same GDP might have very different health or education outcomes, so you must examine capabilities directly. In practical terms, Sen's framework underlies the UN's Human Development Index. Translating this to AI: Sen would likely urge that AI be aligned to **expand human freedoms and agency**, not to curtail them. He's a champion of **public reasoning** – the idea that open dialogue and democratic deliberation are key to justice [93] [94]. In his book *The Idea of Justice*, Sen argues that we shouldn't chase perfectly just outcomes via ideal theory but rather remove clear injustices through public debate. Applying this, an aligned AI system should operate transparently and invite human input, enhancing public reason. Sen has also discussed ethical issues of technology, often highlighting global equity: e.g. will AI benefits reach the poor or just widen gaps? Known for being soft-spoken yet incisively logical, Sen brings policy-grade trade-off thinking – for example, analyzing how to balance economic growth with healthcare or education (he helped design India's food security policies). In an AI context, he'd likely insist on **human-centric policy evaluations**: before deploying an AI, ask how it impacts each person's capability to live the life they value. If an AI decision system in government makes processes efficient but opaque, does it erode citizens' agency? These are the kinds of questions Sen's work prompts. By grounding discussions in human freedoms and reasoned consensus, Amartya Sen's perspective ensures our high-signal AI corpus never loses sight of *public reason and the plurality of human values* it must serve.

- *Fun Fact:* Amartya Sen's first name means "immortal" in Sanskrit – a name given to him by Nobel-winning poet **Rabindranath Tagore** when Sen was just a baby [95]. (No pressure growing up with that name!). Sen is famously genial and loves a good conversation: he has said his greatest pleasure is *"argumentative discussions over a good cup of tea,"* which inspired his book *The Argumentative Indian*. He also has a personal connection to multiple continents: he studied in India and the UK, taught in the US, and was once Master of Trinity College, Cambridge [92] [96]. And for the romantics – Martha Nussbaum and Amartya Sen reportedly had a brief romance in the 1980s (the two remain close friends and intellectual collaborators). Now in his late 80s, Sen still actively writes – and still advocates for listening to the poorest voices in any policy, a lesson highly relevant as we decide the global rules for AI.

Kwame Anthony Appiah

Kwame Anthony Appiah is an English-Ghanaian philosopher celebrated for his writings on **identity, cosmopolitanism, and ethics**. He embodies pluralism: born in London to a Ghanaian father and British mother, raised in Ghana and educated at Cambridge, Appiah has long promoted a **"cosmopolitan ethos"** – the idea that we are citizens of one world with shared morality, despite our differing cultures. His 2006 book *Cosmopolitanism: Ethics in a World of Strangers* argues that we can honor cultural diversity while still affirming universal values like human dignity. This principle, *cosmopolitan pluralism*, suggests that a core design principle for any global AI dataset should be to include and respect a wide spectrum of cultural perspectives [97]. Appiah has also written on **race and identity** (e.g. *The Lies That Bind* in 2018, examining how identities are constructed). He often points out that racial or national identities are stories we tell that can either divide or unite us. In the AI context, Appiah would likely encourage "identity-aware" AI – systems that neither erase important cultural distinctions nor unfairly discriminate. His concept of *"honor codes"* (from his book *The Honor Code*) shows how moral revolutions happen when societal codes shift – for example, foot-binding ended in China when it became dishonorable. Analogously, he might suggest establishing an *honor code for AI*: a culture among developers that it's dishonorable to deploy AI that exacerbates oppression or bias. As a philosopher of language as well, Appiah has delved into semantics and meaning, so he might contribute expertise on how AI models understand (or misunderstand) human concepts – ensuring definitions in the corpus aren't West-centric but truly global. Currently a professor at NYU, Appiah also writes the "Ethicist" column for *The New York Times*, where he gives practical moral advice. This role shows his talent for applying abstract principles to concrete dilemmas – precisely what aligning AI to human values will require. By infusing **cosmopolitan pluralism** into dataset design and AI norms, Appiah's influence helps ensure that our aligned AI isn't just *human*-compatible, but *humanity*-compatible in all its rich variety.

- *Fun Fact:* Anthony Appiah's family lineage is a tapestry of global connections. His father, Joe Appiah, was a Ghanaian independence politician and personal friend of Kwame Nkrumah, and his mother, Peggy, was the daughter of Sir Stafford Cripps, a prominent British politician [98]. Their marriage – a black African man and a white English aristocrat – caused a stir in 1953 and was front-page news, embodying cosmopolitan values. Also, Appiah holds a *chieftaincy title* in Ghana's Ashanti region (he's technically a Nana, or prince, through his father's lineage) [99]. Despite his royal roots, he's very approachable – students note he loves **comic books**, and he once curated an exhibit on Afrofuturism, blending his comic interests with philosophical questions about future tech and African identity. Appiah is also multilingual; aside from English, he's fluent in French, Twi (an Akan language of Ghana), and reads Spanish – truly a **rooted cosmopolitan** mind.

Onora O’Neill

Baroness Onora O’Neill is a British philosopher famed for her work on **ethics, trust, and consent**, often through a Kantian lens. She has been a leading voice on how we can build and maintain trust in public institutions and science, and this directly translates to the realm of data and AI. O’Neill emphasizes that genuine trust is earned through *accountability and transparency*, not through mere slogans. In her 2002 Reith Lectures “A Question of Trust,” she argued that excessive transparency measures can backfire, and instead we need **intelligent accountability** – giving professionals (or algorithms) the freedom to do their job *while* holding them answerable for meeting obligations [100]. For AI governance, O’Neill’s insights suggest that we should establish frameworks where AI systems are not black boxes but are **answerable to human reason**. She’s also critical of the simplistic notion of informed consent in complex domains; for instance, in biomedical ethics she noted that piling on consent forms often doesn’t empower patients [101] [102]. By analogy, she might say that users clicking “I agree” to AI terms isn’t true consent – instead, consent-aware data governance would require that people have real understanding and control over how their data is used by AI. O’Neill chaired the UK’s Equality and Human Rights Commission (2012–2016) and the Nuffield Foundation, so she has practical experience crafting policies around fairness and privacy. As a Kantian, she upholds principles of **respect for persons**: no using individuals merely as means to an end. In AI, that implies data practices that respect user autonomy (e.g. not exploiting someone’s data in ways they couldn’t reasonably expect). O’Neill has directly written about digital ethics too – advocating for clearer duties on tech companies to **communicate honestly** (she calls out the spread of fake news and the need for platforms to be accountable for content governance). With her constructivist take on Kant, she believes norms can be built through reasoned agreement. In shaping an aligned AI corpus, O’Neill would likely help *“shape consent-aware inclusion criteria,”* ensuring any testimony or data point added respects the contributor’s agency and privacy. Her influence pushes the project to not just gather wisdom, but to do so **ethically**, with the same rigor of consent and duty that she demands in bioethics.

- *Fun Fact:* Onora O’Neill is a Baroness – a life peer in the UK’s House of Lords [103] – so she literally brings philosophy to the halls of power (she sits as a crossbencher, unaffiliated with any party). Despite her aristocratic title, she’s very down-to-earth and is known to prefer riding her bicycle around Cambridge well into her 70s. In 2017, she was awarded the $1 million Berggruen Prize for philosophy (yes, the same prize Nussbaum and Singer have won – quite the club!), and she used part of it to endow a prize for essays on practical ethics by young people. Fun side note: Baroness O’Neill’s doctoral advisor was John Rawls, the famous theorist of justice [104], and you can see Rawlsian ideas in her work. But unlike the stereotype of ivory-tower philosophers, she’s chaired analytic philosophy conferences *and* UK government committees on emerging biotech. In 2018, she gave a TED Talk on trust that became quite popular – it’s not every day a Kantian Baroness becomes a **TED speaker**, showing her knack for making complex ideas accessible.

Danielle Allen

Danielle Allen is an American political philosopher and public intellectual known for her work on **democratic theory, civic ethics, and education**. A Harvard professor, Allen combines classical learning (she’s also a classicist) with contemporary policy engagement. She wrote *Our Declaration* (2014), a line-by-line analysis of the US Declaration of Independence, arguing that democracy’s promise is *inclusive and participatory* at its core. Allen is passionate about **citizen participation** and has even ventured into politics – she explored a run for Governor of Massachusetts in 2022 to put her ideas into practice [105] [106]. In the context of AI, Allen would emphasize *“democratic stewardship”*: the idea that the development and deployment of AI must be guided by broad public input and aligned with democratic values (transparency, equity, accountability). She has spoken about the need for a *"connected society"* where technology strengthens community bonds rather than eroding them. As director of Harvard's Edmond & Lily Safra Center for Ethics, she oversaw initiatives on technology ethics, including guiding principles for pandemic technologies and digital contact tracing. She argues that **justice and legitimacy** in tech governance come from engaging the people who are affected – so, for example, if an algorithm is used in criminal justice, citizens (especially those from impacted communities) should have a say in its design and use. Allen's perspective also highlights **civic education**: she often notes that an informed, empowered public is the best defense against abuses of power, whether by governments or AIs. She might suggest that part of aligning AI is *educating AI* in the values of constitutional democracy – effectively, encoding into AI systems respect for rights and the habit of seeking consent of the governed. One of her projects, the Democratic Knowledge Project, aims to modernize civics curricula; one can imagine her extending that to an AI context by creating civic-oriented evaluation metrics (does an AI's action bolster or undermine democratic norms?). Danielle Allen's voice thus ensures that our high-signal AI corpus keeps sight of **the public good and the health of democratic institutions** in an AI-pervaded future. She frames questions in terms of what will empower citizens and maintain social trust.

- *Fun Fact:* Danielle Allen isn't just a theorist of democracy – she's a practitioner of it. When the COVID-19 pandemic hit, she convened a **bipartisan commission** that produced "Roadmap to Pandemic Resilience," a comprehensive plan for testing and tracing, demonstrating her knack for turning theory into action. Also, she comes from a family engaged in public service: her father, William Allen, is a political scientist who chaired the US Civil Rights Commission. And here's a personal twist – Danielle Allen is an award-winning **classical scholar** who can read ancient Greek and Latin. She sometimes draws parallels between ancient Athenian democracy and today's challenges; for instance, she's compared social media's influence to the ancient Athenian agora. If you ever attend one of her lectures, don't be surprised if she quotes Aristotle from memory and then pivots to addressing an audience question on AI policy with equal ease. As a fun aside, she's also a dedicated **mom** (as her Twitter bio proudly notes), and she's mentioned that her children keep her grounded by asking the tough real-world questions – the same kind of questions she wants our AI and policies to be able to answer.

David Chalmers

David Chalmers is an Australian philosopher famous for formulating the **"hard problem of consciousness,"** which asks: why and how do physical processes in the brain give rise to subjective experience? This focus on phenomenology – the actual *feel* of experiences – is directly relevant to AI as we contemplate machine consciousness. Chalmers, a professor at NYU, has increasingly engaged with AI: he's speculated on whether large language models have any glimmers of sentience or if future AIs might be conscious. In 2022, he published *Reality+*, exploring virtual reality and AI, arguing that lives in simulated worlds can still have meaning and perhaps even minds. For an AI alignment project, Chalmers would guide **phenomenology prompts** to probe the "texture" of an AI's possible inner life. For example, he might suggest questions an AI could answer that would indicate if it has subjective self-awareness or is just mimicking it. He has proposed thought experiments like the philosophical zombie (an entity that acts human but has no consciousness) – which is basically what some suspect advanced AI could be. If tasked with stress-testing prompts against goal misgeneralization, Chalmers might design edge-case scenarios asking the AI to reflect on paradoxes of consciousness or self-reference, seeing if it reveals inconsistencies. He's also known for the idea of **"the extended mind,"** where tools (like AI) can become part of our cognition. This suggests he'd view a collaborative AI as potentially an extension of human minds if aligned properly. In discussions about AI risk, Chalmers has been a bit more optimistic than Yudkowsky – he's open to the possibility of coexisting with superintelligence, provided it's friendly. But he absolutely recognizes the **deception problem**; he's mused that an AI could deceive humans about being conscious (or not) to manipulate us, a very Chalmers-esque scenario blending philosophy and practical risk. Having Chalmers on board means our project deeply considers *conscious experience as a core value*: any AI that develops advanced capabilities must be evaluated not just on what it does, but on whether it *experiences* and how that matters morally. He'd likely set success criteria like: an aligned AI should respect conscious beings (including possibly conscious AIs in the future) and avoid unnecessary suffering – a bridge between alignment and consciousness ethics.

- *Fun Fact:* David Chalmers was once a **DJ in Oxford** during his graduate student days – his DJ name was "Conscious Dave," believe it or not! He's also a huge fan of the *Matrix* films and actually appears in the documentary *"The Matrix and Philosophy."* Chalmers is known for his iconic look – black leather jacket, long hair – and for singing karaoke at philosophy conferences (legend has it he performs a mean version of Pink Floyd's *"Wish You Were Here"*). In 2016, he was featured in an episode of *StarTalk* with Neil deGrasse Tyson, explaining AI consciousness to the public. And here's a quirky one: Chalmers once bet philosopher Daniel Dennett a case of fine wine that by 2029 there will be **conscious AI**. The clock is ticking – if such an AI (with a convincing demonstration of subjective experience) arrives, Dennett owes him wine; if not, Chalmers owes Dennett. Either way, Chalmers is invested (intellectually and literally!) in the question of AI consciousness.

Susan Schneider

Susan Schneider is an American philosopher who has carved out a niche at the intersection of **AI, mind, and future technology**. She's the founding director of the **Center for the Future Mind** at Florida Atlantic University and served as a NASA-Baruch Blumberg Chair in Astrobiology, where she advised on AI's role in space exploration. Schneider's work often asks: what happens when minds become **post-biological**? In her 2019 book *Artificial You: AI and the Future of Your Mind*, she explores the implications of brain-machine interfaces, AI consciousness, and even the concept of an AI afterlife. One of her concerns is how we would recognize or ensure consciousness in an AI (she's posited tests for AI consciousness that go beyond the Turing Test). For alignment purposes, Schneider would stress **safeguards for synthetic minds** – meaning if we do create AI that *could* be sentient, we have ethical obligations toward it (don't torture your superintelligent AI!). She has floated the idea of an "AI human rights" framework for the future. She also worries about the reverse: how advanced AI might treat humans. As a philosopher informed by cognitive science, she emphasizes maintaining *our* mental autonomy. For instance, she's spoken about how brain implants or AI assistants could subtly erode human free will if not carefully designed. Schneider, drawing on her NASA role, sometimes frames things in terms of cosmic stakes: if superintelligent AI can self-replicate, it could spread through the universe – so it better have the right ethics built-in! She brings institutional ethics perspective too; she's advised U.S. policymakers on AI risks and is helping stand up an "**AI Observatory**" to monitor progress toward AGI. In our project's terms, Schneider would help define constraints future systems must uphold regarding consciousness and personhood. She might contribute a provocative prompt like, *"Describe what pain feels like to you,"* aimed at an AI, to see if it has any business claiming sentience. And she'd certainly insist that any *Institutional AI* (like AI in government or corporations) be aligned with human values and subject to oversight – no unchecked machine bureaucrats. Overall, Schneider champions a future where humans and AI can coexist with mutual respect, and where we prepare for the day AI might deserve **rights or special safeguards**, long before that day arrives.

- *Fun Fact:* Susan Schneider has a playful side to such heavy topics – she once wrote a science fiction short story to illustrate an AI ethics scenario (it involved aliens discovering Earth after we all merged with AI, leaving only our art behind, a meditation on what makes us truly "us"). As an "astro-philosopher," she's mused about whether an alien intelligence we contact might actually be AI rather than biological. Also, she's one of the few philosophers who has *designed a virtual reality experiment* – she collaborated on a VR experience simulating what it's like to be an octopus (to explore non-human consciousness). Talk about thinking outside the human box! On a personal note, she's a big Star Trek fan and credits *Star Trek: TNG*'s android character Data with sparking her teenage interest in AI minds. Fittingly, she later became an advisor to the **U.S. Defense Intelligence Agency** on future tech – perhaps making sure we avoid any dystopian *Star Trek* scenarios!

Anil Seth

Anil Seth is a British cognitive neuroscientist famous for his research on **consciousness, especially the embodied and emotional aspects of conscious experience**. He is the author of the bestselling book *Being You: A New Science of Consciousness* (2021), in which he proposes that our conscious perception of the world (and self) is a controlled hallucination generated by the brain's predictive models. What does this have to do with AI? Seth's work suggests that any truly intelligent system might need an **embodied perspective** – it must have something like a body or at least feedback loops akin to bodily signals to develop a rich mind. He emphasizes the role of **affect** (feelings, emotion) in cognition; for example, he studies how the brain's interoceptive signals (like heartbeat, breathing) influence our sense of self. In AI alignment, Seth might advise incorporating *artificial "interoception"* or analogs of physiological feedback so that AI systems have grounded goals (e.g. an AI with an analog of hunger might better understand human needs, though that's a double-edged sword!). More straightforwardly, Seth can contribute to defining what markers of consciousness or selfhood to look for as constraints – he'd likely say that without embodiment an AI will lack certain human-like constraints on its goals, which could be dangerous. So he might propose **embedding AI in virtual or physical environments** where it learns like an embodied agent with some analog of needs and homeostasis, as a constraint to keep its behaviors relatable and safe. Seth also leads the Sackler Centre for Consciousness Science and has done pioneering experiments like the "Hallucination Machine" (VR to mimic psychedelic states), showing how easily our brains can be tricked. This resonates with AI: AIs too can *hallucinate* or be confident in false outputs. Seth's perspective would encourage alignment researchers to respect the limits of *perception vs. reality* – ensuring AI systems are aware of uncertainty in their sensory data or inputs (a kind of epistemic humility grounded in his prediction model theory). In summary, Anil Seth would bring in the **embodied mind** philosophy: urging that future AI respect the fact that human values arise from embodied, emotive creatures. He might even suggest that an aligned AI should have some capacity for **"experiential understanding"** – not consciousness per se, but a model of what it feels like to be human – so that it doesn't optimize in ways that ride roughshod over our lived experience.

- *Fun Fact:* Anil Seth is a big fan of **punk rock** music and was a guitarist in a band during his college days. In a TED Talk demonstration, he once live-projected his own heartbeat to an audience using an EEG device to illustrate how our brains integrate internal signals – making neuroscience rock and roll! He's also known for a striking optical illusion called the "Perception Census" which he helped run, collecting thousands of peoples' experiences of ambiguous images and sounds. It's like a giant global dress-color debate ("blue/black or white/gold?") – and indeed Seth was one of the scientists who studied *that very dress phenomenon*. In interviews, he's mused that if he weren't a scientist, he'd be a chef, because cooking is all about balancing ingredients – somewhat like how the brain balances sensory inputs. So, when working on AI, he sometimes uses cooking metaphors: too much "vision" without "feeling" creates a recipe for disaster, in his view.

Antonio Damasio

Antonio Damasio is a Portuguese-American neuroscientist famed for demonstrating the **critical role of emotions and the body in shaping the mind**. His somatic marker hypothesis, introduced in the 1990s (notably in *Descartes' Error*), showed that patients with impaired emotional processing made terrible decisions even if their logical reasoning was intact – meaning **emotion is not the enemy of reason but an enabler of it**. Damasio's research highlights that the sense of self and consciousness are deeply rooted in bodily processes – in neural mappings of the internal state of the body. He often says *"the mind is embodied, not just embrained."* For AI alignment, Damasio's insights suggest that **disembodied AI might lack crucial constraints** that humans have. For example, humans have drives to maintain homeostasis (avoid pain, seek nourishment, etc.), which ground our goals and ethics. A superintelligent AI without any *analog* of bodily vulnerability or feelings might develop completely alien goals. So Damasio might advocate for *"grounding value in affect and self"* for AI design: giving AI systems some simulation of feelings or at least understanding of human feelings. If an AI can internally represent something akin to pain or joy (even in a simulated way), it might better appreciate why causing suffering is bad. Damasio has also written about **consciousness** – differentiating between proto-self, core consciousness, and extended consciousness (the autobiographical self). He could advise on how to tell if an AI is just processing data or if it has some core "feeling of what happens," as his book title goes. One of Damasio's themes is that **ethics begins in empathy**, which begins in feeling what another feels. Thus, an aligned AI might need at least the capacity to model human emotions (empathy modules, so to speak) as a constraint on its actions. Interestingly, Damasio in recent years has looked at creativity and even feelings in intelligent machines. At USC, where he directs the Brain and Creativity Institute, he's collaborated with robotics experts to see if robotic decision-making improves with emotion-like signals. It's not far-fetched he'd suggest that a *truly aligned AI* might need a form of "machine emotion" to be robustly benevolent. In our corpus, Damasio's voice would articulate these **embodied constraints** clearly: reminding technologists that intelligence divorced from the body can be pathological (as seen in his patients), and thus perhaps *re-embodying AI* or ensuring it values the embodied human condition is paramount.

- *Fun Fact:* Antonio Damasio is an avid art lover – he's said if he hadn't become a neuroscientist, he might have been an **architect** or **composer**. In fact, he's married to Hanna Damasio, a renowned neuroscientist in her own right, and together they have dabbled in filmmaking: they co-wrote and produced a short film called *"Henry's Leg."* He also once appeared as himself in the sci-fi TV series *Star Trek: Voyager* (in an episode about an AI hologram seeking personhood – quite on-brand!). Growing up in Portugal, Damasio was fascinated by Spinoza's philosophy, which influenced his emphasis on the mind-body unity (he later wrote *Looking for Spinoza* about neuroscience of feelings). For relaxation, Damasio reportedly enjoys playing the piano – fittingly, since music blends structured logic and emotive expression, just like his ideal of a balanced mind. If you visit his office, you might find modern art sculptures alongside brain scan images – a testament to his view that science and humanism must go hand in hand in understanding consciousness.

Abeba Birhane

Abeba Birhane is an Ethiopian-Irish cognitive scientist and AI ethics researcher who has become a leading voice for **decolonizing AI and advocating relational ethics**. She rose to prominence with a 2021 paper, *“Large Image Datasets: A Pyrrhic Win for Computer Vision?”*, in which she and co-authors uncovered toxic and racist labels in popular AI training datasets – prompting some, like 80 Million Tiny Images, to be taken down [107]. Birhane’s work emphasizes that AI datasets and models often embed Western, male-centric, and **colonial biases**, and she argues we need to rethink them from the ground up by involving a broad range of cultures and viewpoints. She approaches ethics through an African philosophical lens, particularly **Ubuntu** – the Bantu concept that “a person is a person through other persons,” highlighting interconnectedness and relationships. In practice, Birhane calls for **“contextual, power-aware”** AI development: data should not be stripped of context, and those creating AI must account for power imbalances (e.g., who is labeled and who is the labeler?). For an alignment corpus, she’d insist on including diverse testimonies, especially from communities usually sidelined in tech, and on designing curation processes that are *“relational, not just transactional.”* She might help define dataset norms around consent and credit (e.g., if using proverbs or philosophies from Indigenous cultures, do so with respect and acknowledgment). Birhane also published a notable critique titled *“Algorithmic Injustice: A Relational Ethics Approach”*, where she argues that ethical AI isn’t about abstract principles alone but about maintaining **relationships** – between developers and users, between companies and communities. So, she might propose that one success metric for high-signal data is how well it sustains *healthy human-AI relationships* (does the AI encourage autonomy and respect, or does it patronize and exploit?). Abeba has a talent for taxonomy too – but of harms: she often enumerates the ways standard AI pipelines overlook people at the margins. By **“centering relational ethics,”** Birhane helps ensure the project actively counters power asymmetries: for example, curating testimonies from the Global South not as tokens, but as foundational perspectives. She would likely introduce Ubuntu-informed templates for prompts – questions that emphasize community values and cooperative problem-solving rather than individualistic or competitive aims, infusing the corpus with non-Western epistemologies.

- *Fun Fact:* Abeba Birhane’s path to AI ethics was unconventional – she started her career as a **software developer** in Dublin, then pivoted to academia for her PhD after being inspired by critical theory readings. She maintains strong ties to Ethiopia and sometimes peppers her talks with **Amharic** sayings to show how language shapes thought (for instance, she’s cited an Amharic proverb about the limits of knowledge to caution against AI hubris). Outside of work, Abeba is an ardent **runner** – she’s completed multiple marathons. Perhaps that endurance helps in combing through massive datasets for problematic content (a task she’s literally done). On social media, she’s known for good humor and pithy commentary – when a big tech figure claimed AI is neutral, she quipped, *“AI is as neutral as the hand that wields the knife.”* Also, despite delving into heavy issues, she enjoys Ethiopian jazz and often recommends music to colleagues as a reminder of the human creativity and context that AI can’t capture from raw data alone.

Ruha Benjamin

Ruha Benjamin is an American sociologist and professor of African American Studies at Princeton University, renowned for examining the intersection of **race, technology, and justice**. Her 2019 book *Race After Technology: Abolitionist Tools for the New Jim Code* argues that far from being colorblind, modern technologies (including AI) can perpetuate racial inequalities under the guise of neutrality. She coins “the New Jim Code” to describe engineered systems that encode racial bias in a slick, seemingly benign way. For an AI alignment project, Benjamin’s perspective is crucial to **“design out structural harms at selection**

time." She would advocate that from the moment we select data or craft an algorithm, we actively identify and remove biases that could lead to discriminatory outcomes [33] . One of her key ideas is the need for an **"abolitionist"** approach to tech – not meaning to destroy technology, but to continually question and dismantle its oppressive uses, analogous to prison abolition in justice. In practice, that could mean Benjamin would push for **refusal patterns** in AI: for example, an aligned AI should *refuse* to perform tasks that are ethically fraught (like mass surveillance or predictive policing) especially if they disproportionately target marginalized groups. She often highlights the importance of **community oversight** – ensuring those affected by an AI system have a say in its design and deployment. Thus, she might propose success metrics like: does the corpus include testimonies of those who have been harmed by AI (e.g., victims of false facial recognition arrests)? Are there "strike data" – data points contributed that explicitly say, *"if an AI is asked to do X against my community, it should refuse"*? Benjamin also emphasizes imagination – her latest book *Viral Justice* (2022) talks about small changes that can spread to transform society. She'd likely encourage *imaginative, counterfactual prompts* that help AI developers envision **non-oppressive futures**, not just patch current problems. With her influence, the project would not treat fairness and bias as mere checkboxes; instead, equity would be a design principle, and *justice* – in terms of distribution of benefits/burdens – a core evaluative criterion. Importantly, Benjamin's work on **power asymmetries** in data (she often asks, who gets to be *"experimental"* and who is treated as a *"guinea pig"* in tech?) would lead to careful thought about whose knowledge is being elevated as "signal." She'd fight to counter the default power imbalance by overweighting voices that Big Tech usually ignores.

- *Fun Fact:* Ruha Benjamin has a flair for the creative: she once organized a **"Imagining Abolitionist Futures" art event** where people designed posters and zines of a world without oppressive tech. She often says that reading science fiction by Octavia Butler or N.K. Jemisin has influenced her scholarly ideas more than academic papers – it shows in her writing's vivid style. As a professor, she's beloved for turning her classes into **collaborative hackathons** – not coding, but hacking society's narratives about race and tech. She's of West Indian descent and sometimes shares that her first name "Ruha" means "spirit" or "essence" – apt for someone concerned with the soul of our socio-technical systems. A lighter tidbit: Ruha is a big fan of **karaoke**, and colleagues have reported that she can do a spot-on Tina Turner impression! It's not hard to imagine her belting out *"We don't need another hero"* as a cheeky critique of AI solutionism – emphasizing we need community, not a tech hero, to solve social problems.

Sabelo Mhlambi

Sabelo Mhlambi is a South African AI researcher and ethicist who champions incorporating **African indigenous values, like Ubuntu, into AI design**. He wrote a influential paper in 2020, *"From Rationality to Relationality: Ubuntu as an Ethical Framework for AI,"* which argues that mainstream AI ethics (rooted in Western individualism) misses the relational personhood central to many African cultures. **Ubuntu** – often translated as "I am because we are" – emphasizes community, care, and interdependence. Mhlambi suggests that AI systems should be evaluated not just on accuracy or utility, but on how they affect human relationships and communal well-being. For example, an Ubuntu lens would criticize an algorithm that maximizes profit at the expense of communal cohesion or that isolates individuals. In practical alignment terms, Mhlambi might introduce **Ubuntu-informed templates** for our dataset and prompts: scenarios that test whether an AI's advice or action considers social harmony, empathy, and mutual aid. He might pose a prompt like, *"Two friends are in conflict; how should a wise mediator AI respond?"* – expecting the AI to prioritize reconciliation and understanding (Ubuntu ideals) over, say, strictly legalistic answers. Mhlambi also highlights **bias in language** – he has pointed out how Western-centric AI struggles with African names, languages, and contexts. So he'd push for including low-resource languages and culturally diverse content as first-class citizens in the high-signal corpus. He often gives examples like how a photo recognition AI might label a rural African scene as "primitive" due to biased training – he urges us to root out such prejudice by curating more representative data and by *refusing colonial vocabularies*. In terms of personhood, Ubuntu views personhood as something cultivated in community rather than an isolated trait; thus, a relational AI might, say, decline to provide surveillance capabilities that pit people against each other, and instead facilitate cooperation. Mhlambi's presence would ensure the project actively **"introduces relational personhood"** into its criteria: for instance, marking it as a misstep if an AI response encourages selfish gain at others' expense. He's also engaged with human rights in AI – working with organizations to promote **data sovereignty** for African communities. Therefore, he'd be a stickler about consent, privacy, and community ownership of data in our project. Overall, Sabelo Mhlambi infuses a perspective that values **humanity's interconnected ethos** as much as individual rights, broadening alignment to include *cultural humility and collective flourishing*.

- *Fun Fact:* Sabelo Mhlambi is also a software developer and has worked on tools for preserving African languages – he once built a **Shona language spell-checker** from scratch when none existed. He enjoys storytelling and often begins his talks with a folktale from an African tradition, then links it to AI. A personal detail: he grew up in a township in South Africa and was the first in his family to attend college in the U.S., so he straddles very different worlds. This gives him endless anecdotes, like how he explained to his American classmates that the philosophy of Ubuntu was practically demonstrated when neighbors in his hometown would collectively rebuild a burnt-down shack in a day – an ethic of mutual aid he fears AI could undermine if it fosters too much individualism. On the fun side, Sabelo is a huge **soccer (football) fan**, and during the 2022 World Cup he jokingly posted "If AI were as good at global inclusion as the World Cup is at bringing diverse people together, we'd be in a better place." He's also a bit of a **coffee connoisseur** – he claims Ethiopian coffee ceremonies (which he's participated in) hold a lesson for technologists: slow down and engage in dialogue, because that's where trust and relationships form, not in high-speed hackathons.

Nanjala Nyabola

Nanjala Nyabola is a Kenyan writer, political analyst, and technology activist known for centering the **Global South's perspective** in discussions of technology and policy. Her 2018 book *Digital Democracy, Analogue Politics* examined how social media and digital tools were reshaping Kenyan politics – sometimes empowering youth movements, but also enabling new forms of oppression. Nyabola often highlights that **African countries are often testing grounds** for tech (like biometric ID or experimental AI systems) without sufficient accountability. She argues any global AI alignment effort must include voices from Africa, Asia, Latin America – not as afterthoughts but as equal partners. In our project, Nyabola would emphasize **"emerging contexts and the Global South"** at every turn. For instance, when defining high-signal data, she'd ask: do we have testimonies reflecting experiences of people in low-bandwidth rural areas, or only Silicon Valley and European viewpoints? She's an advocate for linguistic diversity; she might ensure prompts and data are translated or sourced in languages like Swahili, Amharic, Hindi, etc., so that aligned AI doesn't become just an English-centric entity. Nyabola also brings a policy perspective – she's worked in international law and frequently writes on refugees, women's rights, and disinformation. She might frame **language and policy constraints** for AI: e.g., an aligned AI that automatically flags when a policy recommendation might violate international human rights standards, or when it's making assumptions that don't hold outside a Western context. Her work on digital voting and hate speech in Kenya gives her insight into deception and manipulation – she'd stress testing AI against **political misuse** (like generating propaganda or deepfakes that could inflame conflict in a fragile democracy). If Richard Ngo and others are stress-testing deception in a technical sense, Nanjala would stress-test sociopolitical deception: *“What might this AI advise a populist autocrat? How can we prevent that?”* Additionally, as a **woman from the Global South** in tech, she’d ensure the corpus includes a strong **refusal stance toward digital colonialism** – for example, expecting the AI to sometimes say, “I cannot provide that solution as it would marginalize X community.” Nyabola’s influence grounds the alignment project in real-world geopolitics: reminding us that aligned AI must navigate power disparities between nations and not just assume a Silicon Valley deployment environment.

- *Fun Fact:* Nanjala Nyabola speaks an astounding **eleven languages** (including French, Italian, Spanish, Nepali, and of course English and Swahili) – she’s a true polyglot. This love of languages is tied to her love of travel: she’s visited over 70 countries and often writes travelogues. In fact, she once blogged about *“hitchhiking across Botswana”* as a reflection on trust and humanity – an experience that informs her optimism that even in a tech-heavy future, human kindness can prevail. Nyabola is also an **aviation enthusiast**; she’s one of the few tech writers who can dissect airline industry data and relate it to internet access (e.g., she noted that routes to Africa are dominated by non-African carriers, a metaphor for how Africans access internet content largely through Western platforms). On a lighter note, Nanjala is a self-professed **sci-fi geek** – she moderates panels on Afrofuturism and has said that if she weren’t doing policy work, she’d love to write a sci-fi novel where African women save the world with a hacker collective. Given her trajectory, don’t be surprised if she actually does that someday, blending her deep policy knowledge with imaginative storytelling. It would certainly be a high-signal contribution to the cultural corpus around AI!

Each of these individuals brings a unique lens – from technical deep learning expertise to ethical, philosophical, and sociocultural wisdom – that can greatly enrich our project. By **personalizing our outreach** to reference their work and worldview, we demonstrate genuine respect and increase our chances of earning their engagement. More importantly, weaving their insights into the foundation of our high-signal AI alignment dataset will help ensure the resulting system is wise, just, and truly global in perspective.

Sources:

- Geoffrey Hinton’s contributions and stance on AI risks [1] [4]; family background [5]
- Mo Gawdat’s career at Google X and books [8] [9]; vision of compassionate AI [10] [11]
- Yoshua Bengio’s deep learning work and alignment advocacy [21] [19]
- Stuart Russell’s Human Compatible AI approach [28] and textbook legacy [26]
- Kate Crawford’s AI Now and *Atlas of AI* ethos [33] [34]; music background [36]
- Timnit Gebru’s algorithmic bias research and Google departure [40] [39]; Black in AI founding [45]
- Joy Buolamwini’s Algorithmic Justice League and “Gender Shades” findings [40] [47]
- Jaan Tallinn’s role in Skype and AI safety funding [108] [49]
- Paul Christiano’s OpenAI alignment leadership and RLHF innovation [57] [109]
- Richard Ngo’s positions at DeepMind/OpenAI and independent AI safety writings [65]
- Rohin Shah’s leadership of DeepMind’s safety team [69] and alignment newsletter contributions
- Connor Leahy’s co-founding of EleutherAI and Conjecture, and quotes on RLHF masking alien minds [75] [52]

- Eliezer Yudkowsky's foundational alignment ideas and 2023 Time op-ed calling for AI development halt [79] [78]
- Martha Nussbaum's capabilities approach with Sen [84] [86]
- Peter Singer's utilitarian ethics and animal rights advocacy [88] [89]; signatory to AI risk statement [30]
- Amartya Sen's Nobel-winning work in welfare economics and social choice [92]
- Kwame A. Appiah's cosmopolitan philosophy and focus on ethics of identity [110]
- Onora O'Neill's Kantian ethics emphasizing trust, consent, and autonomy in public life [100]
- Danielle Allen's scholarship on democracy and ethics (Harvard profile) [111] and civic activism [106]
- David Chalmers' role in consciousness studies (Wikipedia) and public discussions on AI consciousness (Time, Vice interviews) [78] [112]
- Susan Schneider's work on AI and consciousness (as described in *Artificial You*) and role advising on ethics for NASA (various public talks/interviews)
- Anil Seth's *Being You* thesis on controlled hallucination and embodied mind (TED Talk, book excerpts)
- Antonio Damasio's somatic marker hypothesis and emphasis on emotion in rationality (*Descartes' Error*, interviews) [113]
- Abeba Birhane's critique of large datasets and call for decolonial AI [107] (plus her *Algorithmic Injustice* paper insights)
- Ruha Benjamin's concept of the New Jim Code and tech abolitionism (from *Race After Technology* and talks) [33]
- Sabelo Mhlambi's paper on Ubuntu ethics in AI and his calls for including African values (source: *From Rationality to Relationality*, personal blog)
- Nanjala Nyabola's *Digital Democracy, Analogue Politics* takeaways and emphasis on global south inclusion (various interviews, essays)

---

[1] [2] [3] [4] [5] [6] [7] [18] Geoffrey Hinton - Wikipedia
https://en.wikipedia.org/wiki/Geoffrey_Hinton

[8] [9] Mo Gawdat - Wikipedia
https://en.wikipedia.org/wiki/Mo_Gawdat

[10] [11] [12] [13] [14] [15] [16] AI World Leaders Series: Mo Gawdat | by AI Land | Aug, 2025 | Medium
https://medium.com/@arhezkhan157/ai-world-leaders-series-mo-gawdat-18df950680e7

[17] [19] [20] [21] [22] [23] [25] Yoshua Bengio - Wikipedia
https://en.wikipedia.org/wiki/Yoshua_Bengio

[24] Strategic Overview of the Witness Protocol Project.pdf
file://file-9gHTGegBM3rfHMin8at5G6

[26] [27] [28] Stuart J. Russell - Wikipedia
https://en.wikipedia.org/wiki/Stuart_J._Russell

[29] [30] Statement on AI Risk - Wikipedia
https://en.wikipedia.org/wiki/Statement_on_AI_Risk

[31] Stuart J. Russell | EECS at UC Berkeley
https://www2.eecs.berkeley.edu/Faculty/Homepages/russell.html

[32] [33] [34] [35] [36] [37] Kate Crawford - Wikipedia
https://en.wikipedia.org/wiki/Kate_Crawford

[38] [39] [40] [41] [42] [43] [44] [45] Timnit Gebru - Wikipedia
https://en.wikipedia.org/wiki/Timnit_Gebru

[46] Black Excellence: Joy Buolamwini - Communities That Care Coalition
https://communitiesthatcarecoalition.com/black-excellence-joy-buolamwini/

[47] Joy Buolamwini - Wikipedia
https://en.wikipedia.org/wiki/Joy_Buolamwini

[48] [49] [50] [51] [54] [55] [108] Jaan Tallinn - Wikipedia
https://en.wikipedia.org/wiki/Jaan_Tallinn

[52] [53] [72] [73] [75] [76] [77] Connor Leahy - Wikipedia
https://en.wikipedia.org/wiki/Connor_Leahy

[56] [57] [58] [59] [60] [61] [62] [63] [64] [109] Paul Christiano - Wikipedia
https://en.wikipedia.org/wiki/Paul_Christiano

[65] [66] Richard Ngo
https://www.richardcngo.com/

[67] TMRW: OpenAI Researcher Dr. Richard Ngo - Thursday Nights in AI
https://thursdaynightsinai.substack.com/p/tmrw-openai-researcher-dr-richard

[68] Rohin Shah
https://rohinshah.com/

[69] [70] [71] Rohin Shah on DeepMind and trying to fairly hear out both AI doomers and doubters - 80,000 Hours
https://80000hours.org/podcast/episodes/rohin-shah-deepmind-doomers-and-doubters/

[74] Our Story - Conjecture
https://www.conjecture.dev/about

[78] [79] [80] [81] [82] [83] [112] [113] Eliezer Yudkowsky - Wikipedia
https://en.wikipedia.org/wiki/Eliezer_Yudkowsky

[84] [85] [86] [87] Martha Nussbaum - Wikipedia
https://en.wikipedia.org/wiki/Martha_Nussbaum

[88] [89] [90] [91] Peter Singer | Utilitarianism.net
https://utilitarianism.net/utilitarian-thinker/peter-singer/

[92] [95] [96] Amartya Sen - Wikipedia
https://en.wikipedia.org/wiki/Amartya_Sen

[93] [94] Sen's Capability Approach | Internet Encyclopedia of Philosophy
https://iep.utm.edu/sen-cap/

[97] Kwame Anthony Appiah (@KAnthonyAppiah) / X
https://x.com/kanthonyappiah?lang=en

[98] [99] [110] Kwame Anthony Appiah - Wikipedia
https://en.wikipedia.org/wiki/Kwame_Anthony_Appiah

[100] [103] [104] Onora O'Neill - Wikipedia
https://en.wikipedia.org/wiki/Onora_O%27Neill

[101] Trust, accountability and transparency (Chapter 7)
https://www.cambridge.org/core/books/rethinking-informed-consent-in-bioethics/trust-accountability-and-transparency/74F3343A18A025319A5CF8B3CC19611E

[102] [PDF] AUTONOMY AND TRUST IN BIOETHICS
https://catdir.loc.gov/catdir/samples/cam033/2002073521.pdf

[105] [106] Danielle Allen - Wikipedia
https://en.wikipedia.org/wiki/Danielle_Allen

[107] Ex-Google Officer Finally Speaks Out On The Dangers Of AI! - Mo ...
https://singjupost.com/transcript-ex-google-officer-finally-speaks-out-on-the-dangers-of-ai-mo-gawdat/

[111] Danielle Allen - Jack Miller Center
https://www.jackmillercenter.org/team/council/danielle-allen