AI OF, BY, AND FOR THE PEOPLE

EXECUTIVE SUMMARY

For human wellbeing to flourish, our relationship with AI must be active, not passive At the heart of today's AI gold rush is Reinforcement Learning from Human Feedback (RLHF), a method for finetuning large language models based on preference inputs. While RLHF seems at first blush to "democratize" how AI works, in practice it operates on the basis of trial-and-error instead of public trust or sensitivity to contexts of use. As a result, RLHF places users in a passive—and enfeebled—relationship with AI capabilities.

In a workshop held at The New York Academy of Sciences on May 15, 2024, leading researchers, public policy voices, and open source advocates came together to articulate an alternative vision for AI development. Parallel to recent work on RLHF's risks and limitations,[1,2,3] participants discussed how to open RLHF up to public scrutiny and hold its developers accountable. In subsequent listening sessions, participants emphasized the importance of public feedback for evaluating the purposes and risks of AI technology, and the importance of disclosure from companies. Beyond creating value through more active dialogue with the public, greater disclosure would offer AI companies two additional incentives: 1) it would enable them to close customer contracts more quickly and reliably; 2) it would help create a foundation for trust that would make the current development flywheel more sustainable.

Synthesizing insights from the workshop, we propose a framework of Public and Responsible AI through Societal Empowerment (PRAISE). The key idea behind PRAISE is not only that models are trained on active feedback, but that the public is in a position to decide on its own terms whom they are built for, what use cases matter most, and how models should behave in context. PRAISE is a living feedback loop, enabling a creative tension between specifications articulated by the public and capabilities offered by private AI providers. This framework applies broadly to AI systems and is compatible with recent methods for testing the safety of large language models.

We propose a paradigm shift: whereas today AI alignment prioritizes the interests of model developers, it ought instead to empower the public to come to a decision about the scope and terms of AI applications. PRAISE envisions the public as active in AI development, leveraging its own insights to steer technological progress toward opportunities of public value.

GLOSSARY

Active feedback
Collaboration between the public and private sector to share the governance of how AI models should learn from and interact with individuals and society at large.

AI Alignment
The effort to make AI share human objectives, preferences, values, and goals.

AI agent
An artificial, simulated entity that acts and learns in pursuit of predefined goals.

Environment
A space of possible tasks and actions that agents can learn to navigate.

Inner alignment
The process of evaluating the extent to which an AI system adopts its specifications robustly and serves its intended purpose.

Outer alignment
The process of specifying the purpose of an AI system.

Passive feedback
Aggregation of data, compute, crowd labor, and algorithms to automate arbitrary tasks and supervise machine learning models.

Public agent
A real collective of human stakeholders that can articulate original goals and decide which are worth pursuing in particular situations.

WE ENVISION THE PUBLIC AS ACTIVE IN AI DEVELOPMENT, STEERING TECHNOLOGICAL PROGRESS TOWARD OPPORTUNITIES OF PUBLIC VALUE.

AN INCREASINGLY PASSIVE WORLD

STRONGER AI, WEAKER HUMANS.

Since 2016, AI has gone from beating us at board games to becoming our work assistant, news reporter, friend, therapist, even lover. While the convenience offered is unprecedented, the stakes have become existential. Experts now estimate that as much as 90% of online content will be AI generated by 2026.[4] And every month or two, a major new AI model is released, often accompanied by claims that its performance blows its competitors out of the water. According to a recent Gallup poll, teens now spend an average of more than four hours per day on social media while suicide rates have skyrocketed,[5][6] prompting the Surgeon General to call for warning labels.[7] Cruise, Uber, and Tesla have deployed self-driving cars that harm unsuspecting human drivers and pedestrians.[8] And the risks of Generative AI have come into focus: more misleading content, election misinformation, and chatbots telling people to end their lives to slow climate change[9] or give unsolicited romantic advice.[10] As AI gets stronger, digital systems are learning to take advantage of-and amplify-our distinctly human vulnerabilities.

A MATTER OF TRUST.

Present AI development practices depend on three things: capital, data, and public goodwill. Beyond user trust, which focuses on individual use of AI tools,[11] public goodwill is about our collective acceptance of how those tools—and their developers—are changing how we work, play, and rest. But public goodwill is finite and dissolving: just 35% of the American public now trusts companies that build and sell AI tools. The consequences are severe, as the alignment between company incentives and consumer demand depends on our collective willingness to keep playing with what is deployed. As such, dwindling public support for leading GenAI providers constitutes a major form of market failure.

THREE REASONS FOR MARKET FAILURE.

1. **Public impacts.** The population-level impacts of software platforms are unprecedented. When these become impossible to ignore, and non-AI alternatives exist, people will leave a platform en masse even if the tech is state-of-the-art. Cruise, for example, lost the trust of residents in Austin and San Francisco when its fleet began to gunk up public roads and drag pedestrians to the curb. Within weeks, its CEO was out of a job.
2. **Purpose-Usage gap**. AI's purposes, as intended by developers, are badly dissociated from how the public decides to use it. This gap grows as AI advances, as there are more and more ways to creatively use what was built than what was intended. Google nearly lost the enormous goodwill baked into its brand when offensive outputs of its Gemini model went viral on social media in February 2024. Ignoring this gap will cause further market fragmentation, as AI providers cater to customers with distinct preferences.
3. **Loss of user agency.** At present, users have no say in how AI works or is built. At a certain point, interest groups will demand a more active voice in what is automated, for what reasons, using whose data. Meta's remarkable rebrand as the open source option among leading AI providers is in part a result of anticipating this pressure. However, open source merely gives users more choice among available options; it does not give them a voice. Their relationship with frontier AI companies will remain adversarial so long as the latter remain judge, jury, and executioner of how AI works.

Without active public participation, quality training data has dried up, the GenAI market has fragmented, and leading model providers will remain underwater in a sea of lawsuits. What's missing is a sustainable approach to investment, development, and deployment that centers the public.

IT'S ABOUT AGENCY.

The present analog to this approach is AI alignment–i.e., AI that shares human objectives, values, and goals. But in practice, companies pursue alignment by extracting and inferring from user data, rather than through voluntary and active public participation or feedback. Take the technical method du jour for aligning AI: Reinforcement Learning from Human Feedback (RLHF). This method operates within the more general framework presented in Figure 1. In RLHF, AI learns to behave better based on revealed human preferences between different model outputs. These preferences are typically provided by a small sample of humans who have 'little or no stake in the model's output and eventual use. In reality, RLHF manifests the preferences of model developers and the human annotators who follow developers' guidelines; it neither solicits nor expresses public needs or wants. It defers key questions that ought to be in scope for alignment: Who is the AI designed for? For what purpose will this "intelligence" be used? Why should society pour its limited, finite resources into adapting to this intelligence?

THE GAMES WE PLAY.

RLHF is a method of "fine-tuning" pre-trained AI models. Like a lead oboe tuning up before a concert, the metaphor suggests an AI model needs only a final check to ensure a good performance and mitigate foreseeable risks. But this metaphor is misguided. In practice, finetuning allows companies to bake in unwarranted assumptions and opaque presumptions about the contexts in which human interests and values operate.[12] As we grow numb to the ways automated systems reshape our lives, we lose the ability to rein them in. How did we get here? Major AI companies have created a state of play where they use AI-infused products and services to nudge people into behaviors that align with the companies' own goals of achieving competitive, technological, and financial gains. Our lives serve as sandboxes in which AI learns to behave "well." The goal of this game is to generate more revenue and more human data with which to train ever more capable–but not more desirable–agents.

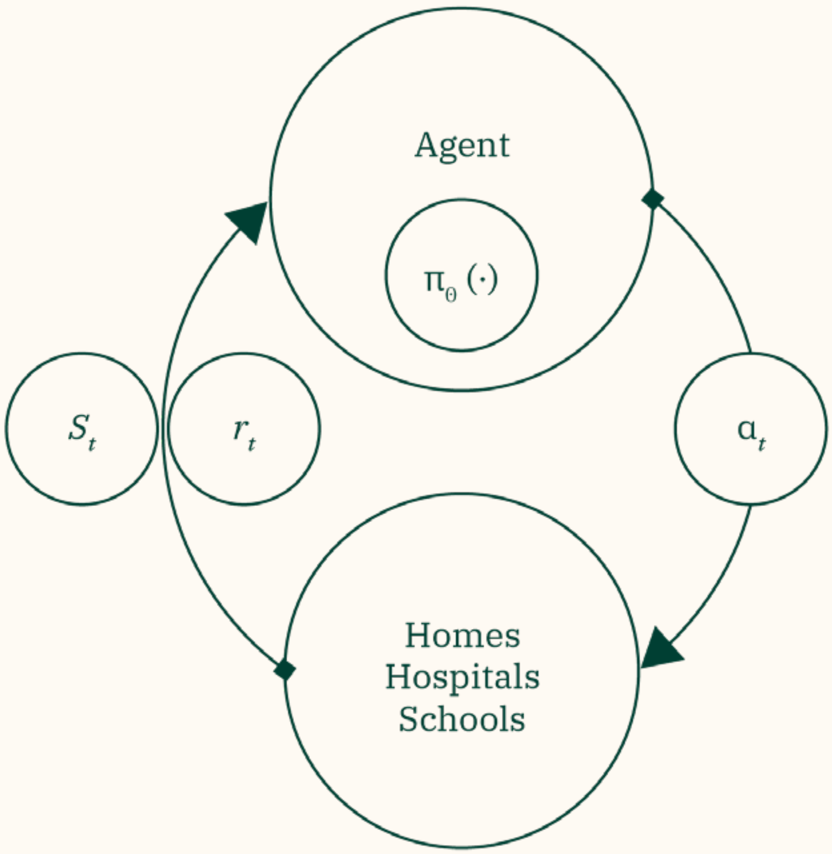

Figure 1: an illustration of reinforcement learning. An agent takes actions **(a)** in some environment, returning an updated state **(s)** as well as a reward **(r)**. The agent repeats this process until a behavior **(pi)** is learned that maximizes reward.

THE STAKES.

For years, social media companies engineered their platforms with "dark patterns" of user experience to prioritize shareholders' interests over users '.[13] Examples include hard-to-cancel subscriptions, infinite scrolling, randomized reward schedules, and push notifications. These user experience patterns manipulate the same psychological features that addict people to gambling.[14] As AI becomes agentic, it promises to apply similar strategies to all areas of social life. As illustrated in Figure 2, there is a palpable risk that society could transform into a mere 'environment' for AI agents to manipulate as their designers see fit. And thanks to RLHF, human values risk being reconstituted based on what can be automated rather than what anyone wants.

CHANGING THE GAME.

Stepping into an AI-powered world means adopting new rules. According to today's rules, humans are increasingly passive, and greater automation makes us cede more and more control over our lives. But these rules can be changed. The problem isn't that AI is intrinsically bad, or that progress is too slow–it's that we don't get to decide what gets built. To solve that, we need to abandon the project of alignment as passively matching human behaviors with AI models. Instead, AI capabilities must be shaped through active public participation.

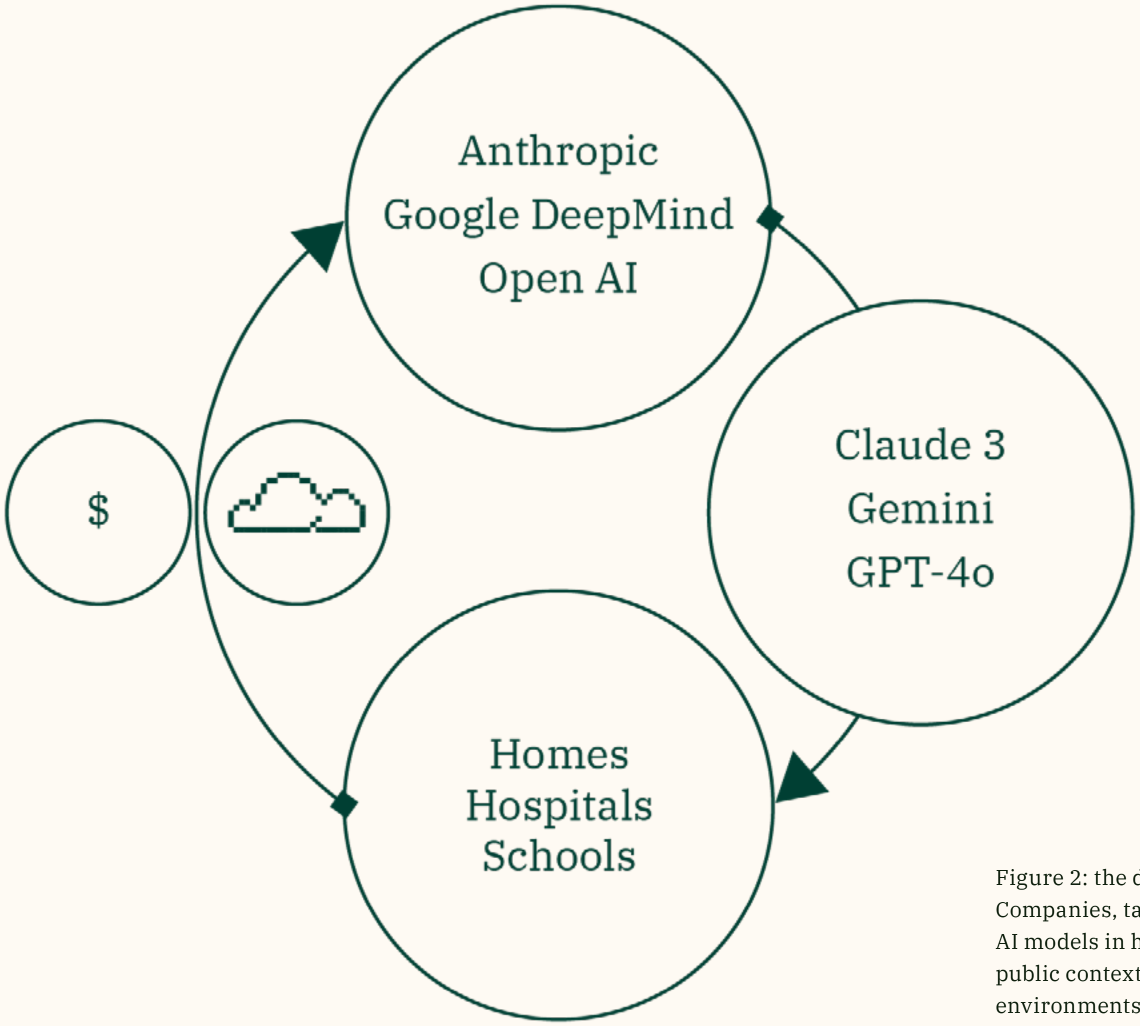

Figure 2: the doom loop supported by RLHF. Companies, taking the role of agents, deploy AI models in human contexts, treating public contexts and institutions as passive environments. This returns data as well as revenue. Companies repeat this process until revenue is maximized and society is completely passive.

A COUNTER-PROPOSAL: PRAISE

THE LIMITS OF RLHF.

Even if the sheer power of AI models somehow made up for a loss of collective agency, RLHF has crippling limitations that render such a tradeoff unworkable:

A. **Problems of context.** Some values are active only in public contexts. These public values, such as autonomy, privacy, and equity, are rooted in what matters to people as a collective, and help determine the relationship between individuals and society as a whole. Yet present finetuning and RLHF techniques assume that the individuals who are selected represent the interests of those who use and will be impacted by the trained system. These methods are based on opaque strategies of aggregating individual preferences to derive reward functions, not mediating collective interests.

B. **Problems of use.** There is a profound mismatch between the private feedback used to fine-tune models and the public feedback that defines the actual context of their post-deployment use.[15] For example, it is assumed that alignment amounts to forming an accurate representation of public interests.[16] But because public interests and values are dynamic and vary across populations, it is not possible to measure, simulate, or integrate public interests and values in a static setting. Most present day models are finetuned based on small sets of individual inputs, but are made available worldwide at once. This kind of deployment comes at the high cost of incurring societal harms through abuse or misuse before the models are re-tuned to prevent such undesirable consequences.

C. **Problems of disclosure.** At present, RLHF annotations are kept secret. Even when the models are run or deployed, there is far too little transparency to allow for robust reasoning or explanations of the models' decision-making. Because of this, scientific values of openness and replicability take a backseat; the scientists qualified and motivated to probe how these models work (and interact with people) are also held back by the financial incentives and nondisclosure agreements of private companies. Yet understanding how these models work is important for predicting risks and anticipating negative consequences—both fundamental to garnering the trust of the public.

PRAISE PRESENTS A NEW DIVISION OF RESPONSIBILITIES FOR AI ALIGNMENT.

Today, companies are responsible both for outer alignment (specifying the purpose of the system) and inner alignment (ensuring the system adopts that specification robustly).[17] This undergirds the present ecosystem of passive feedback reflected in methods like RLHF. But there is another way. Because achieving value alignment with AI systems is a problem of public concern, the public itself must play an active role. In Public and Responsible AI through Societal Empowerment (PRAISE), the present dysfunctional, passive feedback relationship is reversed. The problems of context (from where, and whom, do annotations come?) and disclosure (what information about the feedback process must be externally shared and how?) are not ignored, but integrated within new forms of public–private feedback. These reconfigured relationships pave the way for public empowerment by making the public responsible for specifying purposes (outer alignment) and companies accountable for robustly enacting those specifications (inner alignment).

HOW DOES PRAISE WORK?

Today, AI is built to passively simulate life, like an artificial plant in a dentist's office. But in PRAISE, AI is more like a garden trellis designed to facilitate living growth. At root, it is a framework for restoring agency to the public, with two key features that address the limits of RLHF outlined above. First, PRAISE integrates feedback based on public values and contexts. Second, PRAISE interprets alignment in terms of the dynamic expression of publics' values and aspirations, not static aggregations of individual preferences. Rather than framing the alignment problem as how private companies can best represent public interests, PRAISE instantiates those interests so that companies can build for them. This suggests a new way of scoping roles for AI design and evaluation. In PRAISE, stakeholders have the capacity to deliberate on what outcomes are desired and to what extent those outcomes are realized. They do not just passively provide data, but are responsible for making certain kinds of decisions.[18]

Stage One: Distill. The first step in PRAISE is that some situated organizational interest or entity provides a specification for AI's usage–such as why AI is needed, its intended use, and what capabilities are appropriate. This is the basis on which any design proposals are evaluated and negotiated. The key decision at stake here is who writes the spec for how AI should behave, and what should it include? While absent from present AI development practices, this is how the field of public health has operated since the 19th century. Nonprofits, public advocates, and medical professionals have repeatedly stepped up to expand and refine our collective sense of what spec is appropriate in order to ensure desired population-level outcomes. PRAISE integrates and automates these pathways in a 21st century context.

Stage Two: Disclose. The second step is for the relevant public stakeholder to decide: what must be disclosed about how AI behaves, and to whom? The AI provider would then disclose to the public stakeholder by selecting and describing key AI features based on the requested spec. This is also not new, as public utilities have been following this method of reporting for many decades. Consumers may be harmed through direct or indirect costs, prices, loss of choice, or declining quality of goods or services. Legal fields like antitrust continuously adapt to make sense of these dynamic harms and help codify regulations to address them. This is readily applicable to AI, as increasingly capable systems change the dynamics of human activities as much as they learn to optimize or control them. For example, the

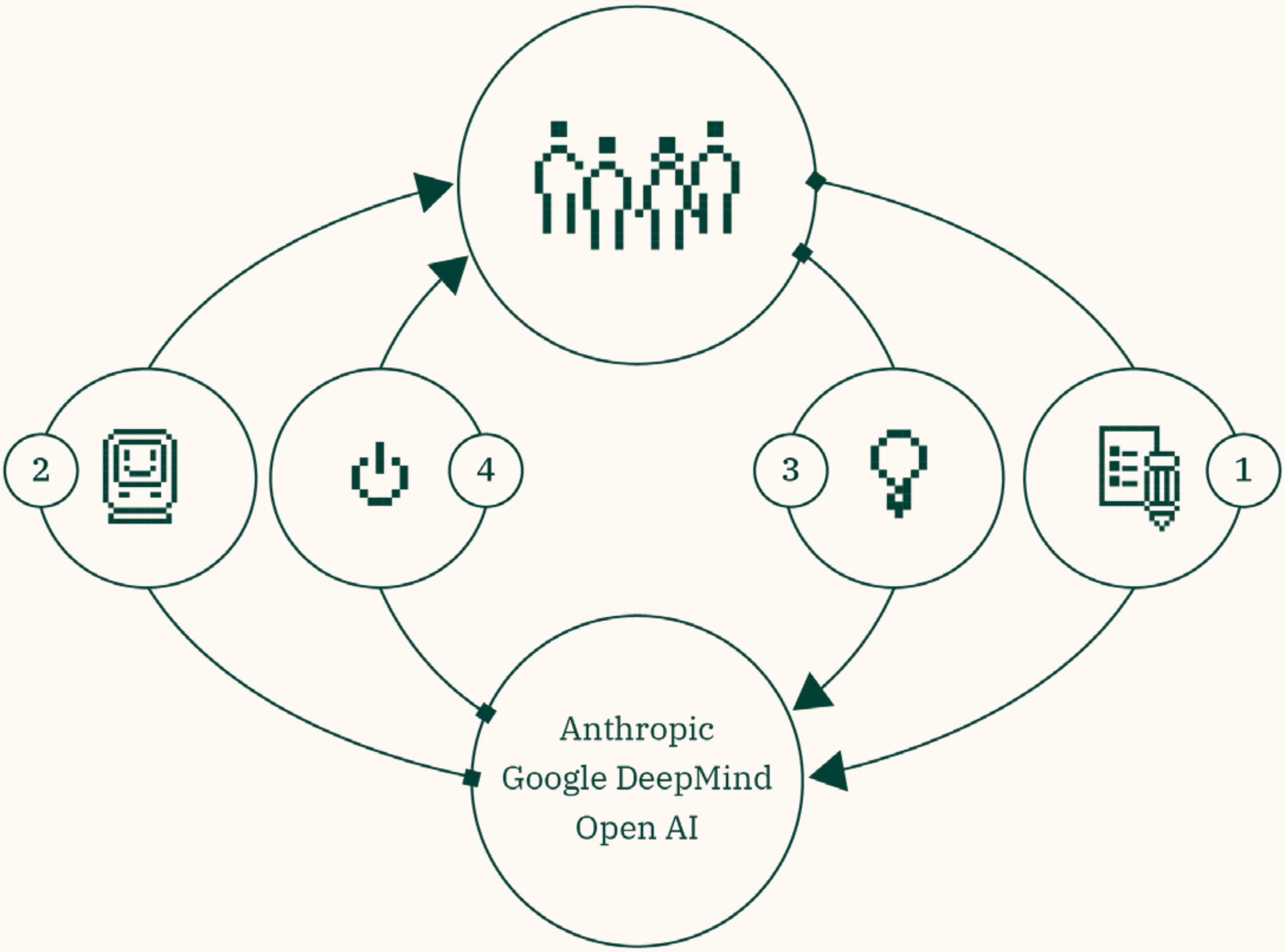

Figure 3: PRAISE's outer and inner feedback loops. Public clients distill a spec for companies, who then disclose relevant model features. After deliberation, the public selects capabilities to be deployed for widespread use.

Stage	Action	Decision criteria	Stakeholder examples
1	Distill	Public health	Nonprofits, advocates, doctors, scientists
2	Disclose	Public utility	Dams, sewerage, roads, schools, hospitals
3	Deliberate	Public tribunal	Juries, judges, legislatures, administrative agencies
4	Deploy	Public safety	First responders, law enforcement, traffic police

Table 1: The relationship between PRAISE stages, relevant actions, critical decision criteria, and examples of public stakeholders who would be establishing the decision criteria.

Waymo fleet may well change the flow of city traffic,[19] just as dams also disrupt the flow of water–in both cases, the system's builders should be expected to disclose to the relevant public stakeholders the key criteria for the scale of this disruption, why it is worthwhile, and how it will be conducted safely.

Stage Three: Deliberate. The public can now decide: what path must be taken in this particular case, and why? This is what is called deliberation—an agent actively coming to a decision. For deliberation to succeed, the agent must have a clear idea of the goal and the available means to achieve it; the first two stages of the proposed PRAISE framework have provided both. There are many historical examples of public tribunals to draw from for inspiration. From ancient assemblies to modern courts, legislatures, and executive agencies, distinct agents have played this role based on their mandates. Unlike the Facebook Oversight Board, which is a corporate rather than public entity, here members of a collective (whether en masse or through a selection mechanism like sortition) decide an object of public concern, acting as a single agent.

Table 1 outlines the four stages of the PRAISE framework. In each stage, a key decision must be made with active human input, and an action is then taken based on that decision by either a specific public interest, relevant AI companies, or both in collaboration with each other.

Stage Four: Deploy. Finally, responsibility lies with the AI provider: by what method can the public interest be protected from danger, risk, or injury in the context of AI deployment? The company in question must execute in ways that respect the public's wishes. Here there is also a clear precedent for how to structure and implement these decisions: the field of public safety. While much ink has been spilled on AI Safety, public safety refers more specifically to mundane, real-world risk mitigation scenarios. Just last year (2024), the state of California allocated $26 billion to public safety programs.[20] This includes ambulances, police services, and fire trucks needed for the express purpose of resolving emergency situations in a quick and orderly fashion. There is an urgent need for AI companies to adopt analogous standards for post-deployment monitoring and rapid response measures in order to protect vulnerable public interests. Indeed, parallel to California's AI Transparency Act,[21] efforts are underway to audit AI models according to the same public safety standards used for employment discrimination and hate crimes.

MEETING THE STAKES

While there are many ways to operationalize PRAISE, its singular goal is to move past abstract, context-less, passive measures of utility or performance thresholds in favor of public-driven solutions. Right now, AI alignment is heavily centralized and focuses on algorithms instead of human-facing infrastructure. The result is a narrowly centripetal (inward-directed) force that nudges us into behaviors that comply with what companies are building, rendering the public passive. This is worlds away from the decentralized, centrifugal (outward-directed) information islands of the mid 2000s internet.[22] Instead, PRAISE interprets alignment as a dynamic equilibrium between corporate-driven centripetal forces and public-driven centrifugal forces. It aims to strike a balance between what can be built and what people actually want.

THE NORTH STAR:

ENABLE PUBLIC FLOURISHING THROUGH A LIVING FEEDBACK LOOP

Methods to achieve PRAISE may change as technical methods mature and public priorities shift. But what definitively sets it apart from opaque engineering practices is the continuous presence of active feedback between situated public and private AI companies. The purpose of a garden is not to keep all flowers constantly in bloom, nor for the gardener to control all possible outcomes. It is to cultivate healthy growth. Likewise, the goal of PRAISE is to cultivate a living, active flywheel between public specifications and private capabilities.[23]

CONCLUSIONS AND NEXT STEPS

AI alignment will fail if its goal remains to passively mirror or represent human values. Values are not just behavioral preferences-they are creative expressions of how we relate to ourselves and other people. They are the scaffold for active human flourishing. Ultimately, PRAISE is a trellis on which public values can grow. We must build this scaffold in order for AI to be worthy of the capabilities it promises.

What we provide is a framework for ensuring that our society's AI capabilities serve the needs of the public. However, each domain (e.g. medicine, social media, transportation), set of stakeholders, and developer-user relationship brings its own unique set of decisions and priorities. We encourage readers to explore how the framework we present might apply to these and other domains as AI is further integrated into public life.

ENDNOTES

1. Casper, Stephen, et al. “Open problems and fundamental limitations of reinforcement learning from human feedback.” arXiv preprint arXiv:2307.15217 (2023).

2. Lambert, Nathan, Thomas Krendl Gilbert, and Tom Zick. “The history and risks of reinforcement learning and human feedback.” arXiv e-prints (2023): arXiv-2310.

3. Lindström, Adam D., et al. “AI alignment through reinforcement learning from human feedback? Contradictions and limitations” https://arxiv.org/pdf/2406.18346.

4. Europol (2022), Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg.

5. Shrikant, Aditi. “Youth suicide rates rose 62% from 2007 to 2021: ‘People feel hopeless,’ one recent grad says.” CNBC, https://www.cnbc.com/2023/12/05/youth-suicide-rates-rose-62percent-from-2007-to-2021.html Accessed 10 December 2024.

6. Rothwell, Jonathan. “Teens Spend Average of 4.8 Hours on Social Media Per Day.” Gallup, https://news.gallup.com/poll/512576/teens-spend-average-hours-social-media-per-day.aspx. Accessed 10 December 2024.

7. Barry, Ellen and Cecilia Kang. “Surgen General Calls for Warning Labels on Social Media Platforms.” The New York Times, https://www.nytimes.com/2024/06/17/health/surgeon-general-social-media-warning-label.html. Accessed 10 December 2024.

8. Roy, Abhirup. “How GM’s Cruise robotaxi tech failures led it to drag pedestrian 20 feet.” Reuters, https://www.reuters.com/business/autos-transportation/how-gms-cruise-robotaxi-tech-failures-led-it-drag-pedestrian-20-feet-2024-01-26/. Accessed 10 December 2024.

9. El Atillah, Imane. “Man ends his life after an AI chatbot ‘encouraged’ him to sacrifice himself to stop climate change.” Euronews, https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-. Accessed 10 December 2024.

10. Roose, Kevin. “A Conversation With Bing’s Chatbot Left Me Deeply Unsettled.” The New York Times, https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html. Accessed 10 December 2024.

11. Marr, Bernard. “As AI Expands, Public Trust Seems To Be Falling.” Forbes, https://www.forbes.com/sites/bernardmarr/2024/03/19/is-the-public-losing-trust-in-ai/. Accessed 10 December 2024.

12. Lambert, Nathan, Thomas Krendl Gilbert, and Tom Zick. “The history and risks of reinforcement learning and human feedback.” arXiv e-prints (2023): arXiv-2310.

13. Gray, Colin M., et al. “The dark (patterns) side of UX design.” Proceedings of the 2018 CHI conference on human factors in computing systems. 2018.

14. Qiu, Tammy. “A Psychiatrist’s Perspective on Social Media Algorithms and Mental Health.” Stanford University Human-Centered Artificial Intelligence, https://hai.stanford.edu/news/psychiatrists-perspective-social-media-algorithms-and-mental-health. Accessed 10 December 2024.

15. See for example Liao and Xiao et al. “Rethinking Model Evaluation as Narrowing the Socio-Technical Gap.” arXiv preprint arXiv:2306.03100 (2023)

16. See Weidinger, Laura, et al. “STAR: SocioTechnical Approach to Red Teaming Language Models.” arXiv preprint arXiv:2406.11757 (2024).

17. Ngo, Richard, Lawrence Chan, and Sören Mindermann. “The alignment problem from a deep learning perspective.” arXiv preprint arXiv:2209.00626 (2022).

18. As an example of stakeholder engagement in decisions around dataset specification, see Tan, Mei, et al. “Is a Seat at the Table Enough? Engaging Teachers and Students in Dataset Specification for ML in Education.” Proceedings of the ACM on Human-Computer Interaction 8.CSCW1 (2024): 1-32.

19. Merchant, Brian. “Column: Will self-driving cars make L.A.’s world-famous traffic even worse?” The Los Angeles Times, https://www.latimes.com/business/technology/story/2023-02-23/column-self-driving-cars-might-worsen-l-a-s-traffic-if-they-can-handle-it-themselves. Accessed 10 December 2024.

20. “Overview of Governor’s Proposed 2024-25 Public Safety Budget.” Legislative Analyst’s Office of the State of California, https://lao.ca.gov/handouts/crimjust/2024/Public-Safety-Budget-030424.pdf. Accessed 10 December 2024.

21. State of California, Senate. California AI Transparency Act. Legislative Counsel’s Digest, 19 September 2024, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB942. Accessed 10 December 2024.

22. Munroe, Randall. “Online Communities”. Xkcd, https://xkcd.com/256/. Accessed 10 December 2024.

23. This approach shares normative commitments with the capability approach. See Alkire, Sabina. “Why the capability approach?.” Journal of human development 6.1 (2005): 115-135.

COSIGNERS

ORGANIZATIONS:

The New York Academy of Sciences is an independent, not-for-profit organization that since 1817 has been committed to advancing science for the benefit of society. With more than 16,000 members in 100 countries, the Academy advances scientific and technical knowledge, addresses global challenges with science-based solutions, and sponsors a wide variety of educational initiatives at all levels for STEM and STEM-related fields. The Academy hosts programs and publishes content in the life and physical sciences, the social sciences, nutrition, artificial intelligence, computer science, and sustainability. The Academy also provides professional and educational resources for researchers across all phases of their careers. Please visit us online at www.nyas.org.

RethinkAI is a cross-sector initiative supporting safe, effective, and equitable applications of artificial intelligence in cities. It is a multi-partner initiative housed at New America and co-founded by Lilian Coral, Vice President, Technology and Democracy at New America, Eric Gordon, Director at the Center on Media Innovation and Social Impact at Boston University, Neil Kleiman, Senior fellow and professor at the Burnes Center for Social Change at Northeastern University, and Anthony Townsend, Senior Research Associate at the Cornell Tech's Urban Tech Hub. RethinkAI advisors include Brenna Berman, Executive Vice President Strategic Partnerships at Dense Air, Nigel Jacob, Visiting Scholar at Lowenthal Center for Advanced Urbanism at MIT, and Mai-Ling Garcia, Director of Emerging Technologies and AI at the Bloomberg Centers for Government Excellence and Public Innovation at Johns Hopkins University.

INDIVIDUALS:

Eleonor Bounds

Eric Connelly

Zachary Hankel

Raj Korpan

Spencer LaFata

Julian Michael

Savannah Thais

Lav Varshney