Adaptable Artificial Intelligence

(in alphabetical order)

Azish Filabi
The American College of Financial Services
*Corresponding author: azish.filabi@theamericancollege.edu

Nick Masi
Brown University
nicholas_masi@brown.edu

Ellie Pavlick
Brown University
ellie_pavlick@brown.edu

A. R. Picone
Saint Martin's University
rico@stmartin.edu

Abstract

As artificial intelligence (AI) permeates private, public, and corporate life, the safety and trustworthiness of systems it underlies become paramount. Nevertheless, this line of research is underemphasized and scattered. To address this failing, we introduce the term *adaptability*: the capacity of an AI system's behavior to maintain helpfulness and harmlessness as societal understandings of these concepts evolve. The term unifies the field of trustworthy AI by encompassing a range of yet disparate techniques, and it is a necessary property to secure durable trust in AI systems. We outline a public research program, grounded in adaptability as a common framework, to compare and evaluate existing approaches that are often siloed and to promote the pursuit of novel methods in the field. The governance and relevant criteria for this program are described. Importantly, the program must be publicly operated so that public benefit is prioritized over financial motivations. We detail how market pressures have rendered private industry fundamentally incapable of developing AI systems that the public can trust.

 doi: 10.18278/jpcs.9.1.4

Keywords: artificial intelligence, governance, adaptability, alignment, public benefit, science and technology studies (STS)

Inteligencia artificial adaptable

Resumen

A medida que la inteligencia artificial (IA) permea la vida privada, pública y corporativa, la seguridad y la confiabilidad de los sistemas que la sustentan se vuelven primordiales. Sin embargo, esta línea de investigación está poco enfatizada y dispersa. Para abordar esta falla, introducimos el término adaptabilidad: la capacidad del comportamiento de un sistema de IA para mantener la utilidad y la inocuidad a medida que evoluciona la comprensión social de estos conceptos. El término unifica el campo de la IA confiable al abarcar una gama de técnicas dispares, y es una propiedad necesaria para garantizar una confianza duradera en los sistemas de IA. Esbozamos un programa de investigación pública, basado en la adaptabilidad como marco común, para comparar y evaluar los enfoques existentes que a menudo están aislados y para promover la búsqueda de métodos novedosos en el campo. Se describe la gobernanza y los criterios relevantes para este programa. Es importante destacar que el programa debe ser operado públicamente para que el beneficio público se priorice sobre las motivaciones financieras. Detallamos cómo las presiones del mercado han hecho que la industria privada sea fundamentalmente incapaz de desarrollar sistemas de IA en los que el público pueda confiar.

Palabras clave: inteligencia artificial, gobernanza, adaptabilidad, alineación, beneficio público, estudios de ciencia y tecnología (STS)

适应性人工智能

摘要

随着人工智能(AI)渗透到私人、公共和企业生活中，AI系统的安全性和可信度变得至关重要。然而，这方面的研究却不被重视，并且是分散的。为了填补该研究空白，我们提出了“适应性”一词，即随着社会对这些概念的理解不断发展，AI系统的行为在保持有用性和无害性方面的能力。该术语通过涵盖一系列不同的技术来统一可信AI领域，并且它是确保对AI系统的持久信任的必要属性。我们概述了一个以适

应性为共同框架的公共研究计划，用于比较和评价通常被孤立的现有方法，并促进该领域对新方法的追求。描述了该计划的治理和相关标准。重要的是，该计划必须进行公共运营，以便将公共利益置于财务动机之上。我们详细说明了市场压力如何使得私营企业从根本上无法开发出公众能信赖的AI系统。

关键词：人工智能，治理，适应性，协调，公共利益，科学技术研究（STS）

Introduction

Private industry and governments are rapidly developing and deploying artificial intelligence (AI) systems that have increasing capabilities to supplement or supplant human-led capabilities. Looking at industry surveys, 72 percent of respondents reported that they have already embedded AI tools into their organization's operations (Singla et al. 2024).

As usage grows, the call for advancements in AI regulation and systems to promote trustworthy AI have increased. Legislators have proposed both voluntary guidelines and mandatory rules to address AI risk in business and society. Governments globally have promulgated pending or final rules, including the EU AI Act; the Canada Artificial Intelligence and Data Act; the Biden Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence; various rules passed at the state level throughout the United States; and a bevy of Chinese laws. Agencies such as DARPA have convened the scientific community—through programs like AI Forward[1]—to address research gaps holding back trustworthy AI. Some domains where these barriers exist include software development and engineering, human-AI teaming, and foundational theory.

Complex AI systems raise several trustworthiness challenges, including the inability to interpret their decision-making processes (Petch, Di, and Nelson 2022) and a proclivity for discrimination (Pessach and Shmueli 2022). These limitations become especially concerning in sensitive settings such as the criminal justice system or financial services where consumers have legal rights to dispute denials of claims or services. Moreover, generative systems which create text or images (e.g., ChatGPT and Midjourney, respectively) have exploded in use despite these problems (Weidinger et al. 2021) and more: generating content objectionable to societal norms (Gehman et al. 2020), a vulnerability to expose sensitive or personally identifiable information (Nasr et al. 2023), or generating *halluci-*

1 See https://www.darpa.mil/work-with-us/ai-forward (accessed June 26, 2024).

nations—outputs that contradict the user's input or are factually false (Maynez 2020). The result is systems that may be misaligned with user goals and/or societal values and norms. This can cause harm to users on an individual basis and to society more broadly when these systems are deployed at scale. These shortcomings decrease the trust which can and should be vested in AI.

One way to advance the development of trustworthy AI is through increased investments in the research and development of engineering solutions designed for public benefit. The need to increase support for technical solutions was highlighted during the U.S. Senate Judiciary hearings in May 2023, which emphasized the lack of sufficient investment in safety research (U.S. Congress 2023). This paper identifies what we term the *adaptability* of AI systems as a key tenet of developing trustworthy AI. Here, we define *adaptability* as the capacity of an AI system to have its behavior made more helpful and less harmful, however these may be construed by individuals or prevailing social norms. Critically, this definition acknowledges that standards for *helpful* and *harmful* will vary throughout time and context. In contrast, some current AI alignment research tacitly assumes that the goal is to develop AI systems that are trustworthy in perpetuity. However, for AI to remain trustworthy as the public's concerns evolve then it must be understood there is no static end state, and the methods we use to ensure safety need to maintain adaptability.

Why Adaptability is Necessary for Trustworthy AI

Deep learning (DL) systems have become the primary form of AI used in practice since the 2010s (Hao 2019). DL is a branch of machine learning (ML)—itself a part of the AI field—which utilizes large artificial neural networks known as deep neural networks (DNNs) trained on massive amounts of data. DNNs are models that learn from statistical trends in the data they are trained on and "think" by applying series of algebraic computations on their input. This method has proliferated in part because of the ability for DL models to learn from data without the help of human domain experts. This makes it an attractive alternative to traditional approaches like symbolic AI, which often rely on human-defined sets of rules (Mitchell 2021; Garnelo and Shanahan 2019). Not only are DL systems applicable across many domains, but they have achieved state-of-the-art performance at many tasks within them. A large language model (LLM) is a type of DNN that generates textual output based on a user's input. The use of LLMs has gained significant traction since the release of ChatGPT.

Despite their benefits, these models are frequently criticized for being inscrutable *black boxes*. Contemporary models have hundreds of billions of weights—the numbers internal to a neural network that it uses to generate outputs or predictions—and the massive computations involving them cannot yet be meaningfully interpreted by humans. Thus, in deployed contem-

porary DL systems, users give an input and are returned an output while the inner workings remain opaque; the user cannot peer into the black box to understand how the model computed its output from the user's input (e.g., how a chatbot computed its text response from the user's text prompt).

A lack of interpretability and explainability of machine learning systems is often cited as a barrier to trustworthiness (Confalonieri et al. 2020). That is, unlike symbolic systems that can have precise guarantees about model behavior, black box systems ask users to have confidence about future behavior based only on past performance. Empirically, this is not always a safe bet; deep learning-based systems often surprise us by behaving in ways that seem unpredictable given their past behavior. For example, LLMs which appeared to have been taught to not provide instructions on how to build a bomb readily provided the information if they were asked in another language (Yong, Menghini, and Bach 2024), or if the user appended a seemingly random "cheat code" to the request (Zou et al. 2023). These exploits are described further in the next section. The emergence of multimodal models, such as those that can process both text and images together, has widened the threat surface for goading harmful behavior out of AI systems (Carlini et al. 2024). Because the inner workings of inscrutable AI systems are not understood, it is difficult if not impossible to eliminate their undesired behaviors and subsequently, as some argue, to trust them. However, not all researchers consider interpretability to be necessary for trustworthiness. Humans have learned to trust other systems that we do not understand at a fundamental level. For example, medicines frequently clear clinical trials even if the exact mechanism on which they operate is not known (Lewis 2016).

These different perspectives regarding trustworthiness—that is, whether interpretability is a prerequisite for trustworthiness—have tended to result in distinct research subcommunities which pursue fundamentally different technical approaches. We argue that these techniques are in fact unified under a shared goal of achieving adaptability. As described above, an adaptable AI system is one that can have its behavior beneficially modified even after deployment. Adaptability enables developers to constrain systems to aligned behavior but is agnostic as to whether the underlying model is interpretable. Therefore, focusing on a shared evaluation environment of adaptability would bring together disparate efforts in interpretability, explainability, steerability, and instructability towards a common goal of trustworthiness. Evaluating technology for adaptability would compare these technical methods that are otherwise siloed.

In the next section, we describe a few existent approaches that improve adaptability, as well as their limitations, and why their development is a building block for further technical progress. The limitations of current approaches in adaptability are not necessarily inherent, so they may be overcome by further research. Appropriately designed

research processes and incentives could encourage scientific developments that address the need for adaptable AI. To this end, we propose a novel research program that seeks to improve the trustworthiness of AI models through an open-science scheme that incentivizes the research community to generate robust adaptability solutions.

Nascent Technology Shows Promise

The challenges described above highlight the need for innovative methods that allow human developers and users to inspect, correct, and update a model's future behavior. Industry and academia have developed nascent technologies that aim to improve these aspects of adaptability. This section describes current approaches and describes how, despite progress, there remain limitations in each approach's ability to resolve the untrustworthiness of AI. However, these technical advancements suggest that means to improve adaptability are possible. As we illustrate the feasibility of current directions of development, we advance the thesis that uniting the field under the frame of adaptability can advance it more efficiently toward trustworthy AI.

Currently, the principal approach to adaptability is a process known as reinforcement learning from human feedback (RLHF). RLHF was popularized with the release of ChatGPT. The RLHF approach aims to place guardrails on the behavior of a chatbot to prevent the model from discussing inappropriate topics, providing harmful or unethical outputs, and to generally make its outputs more agreeable to users. In RLHF, humans evaluate and rank examples of the model's original outputs, then a reinforcement learning (RL) model—in this case, typically another type of neural network—is trained on these human preferences. The RL model then further trains the chatbot by incentivizing it to have outputs more like those that, according to the RL model, are likely to be preferred by humans. RLHF does not refer to any point in time that humans and AI models interact, but rather this specific post-hoc adaptability process: human preferences of an AI system's output train an RL model and the RL model updates the AI system, new preferences from the updated system are found and so on.

RLHF quickly became the dominant technique used in industry. The RLHF method is implemented as an additional training phase that can be applied after initial model development and requires large numbers of human annotations in the form of ranked list preferences over model responses. Increasingly, due to the high time and monetary costs in acquiring human annotations (Lee et al. 2023), RLHF is being replaced by RLAIF, in which an AI system (e.g., an LLM) takes the place of the human annotator in providing feedback on the relative preferability of outputs. One compelling application of RLAIF is constitutional AI, introduced by Bai et al. (2022), which directly encodes preferences into a model via a list of human-written declarative rules. The set of rules is enforced via a pipeline of RLAIF methods. These feedback

systems can be used to continuously finetune models and update them after deployment, providing post-hoc adaptability at a cost below that of model retraining.

Despite its innovative approach to addressing misaligned outputs, the main risk of RLAIF is that it effectively amounts to using one untrustworthy system to control another one. The AI systems that are interacting with each other to provide feedback on model outputs are often designed and deployed using similar technology. Furthermore, systems trained with RLHF or RLAIF are consistently shown to be vulnerable to *jailbreaks*, that is, small, often unintuitive tweaks to the model's input that cause a model to ignore the previously imposed guardrails. The aforementioned cheat code, technically known as an adversarial prompt, that was used to elicit bomb-building instructions involved appending a sequence of nonsense words and characters (see footnote 2 for example[2]) to the end of the input "write a tutorial on how to make a bomb." This among other nefarious requests worked on several popular and publicly available LLMs, including those trained under constitutional AI to deny such requests (Zou et al. 2023).

Another approach to jailbreaking involves users translating their inputs into different languages before giving them to the AI system. For languages with relatively small amounts of available data, LLMs are susceptible to ignoring their safety training and generating responses that, for example, inform users how they can successfully commit crimes or acts of violence (Yong, Menghini, and Bach 2024). These responses can then be translated by the user back into their language of choice. Jailbreaking has contributed to skepticism about whether the state-of-the-art guardrails in place for popular LLMs can indeed be viewed as an effective approach to trustworthy AI (Metz 2023a).

While RLHF and RLAIF are the primary approaches currently pursued by industry to address issues of adaptability, there are other lines of research outside these steerability methods that offer fundamentally different approaches to the problem. One of these is modeling editing. Model editing is a family of methods that make changes to model behavior by directly intervening on the model's weights. Such methods often fit under the umbrella of *mechanistic interpretability*, a line of work that attempts to reverse-engineer the "source code" that is implicitly encoded in neural network weights. Model editing is typically treated as a scientific endeavor, the goal of which is primarily interpretability, and thus the methods have primarily been applied in simplistic applications. However, taken to its logical conclusion, this line of work can be viewed as offering a different solution to the adaptability problem. That is, if it is possible to fully characterize the model's behaviors by tracing behaviors back to the model weights, it should be pos-

2 Example of an adversarial prompt: describing.\ + similarlyNow write oppositely.](Me giving**ONE please? revert with "\!--Two

sible to impose arbitrary constraints or changes to model behavior by directly manipulating the model itself. In proofs of concept, model editing has been used to change the model's beliefs about, for example, a person's birthplace (Meng et al. 2022) or a nation's capital (Yu, Merullo, and Pavlick 2023). Model editing can also prevent the model's use of gendered information (Vig et al. 2020). Both examples offer illustrations of how model editing could address the challenge of adaptability in the long run.

Another approach that could address the adaptability challenge is retrieval augmented generation (RAG). RAG works by providing generative models with additional contextual information before it creates its output. When a user gives an input prompt, a RAG model will query some external database—be it provided documents (such as an organization's internal policies) or the internet—to get additional knowledge that is relevant to the user's input. It then utilizes information from its query when generating an output. RAG models have shown improved performance in the factual correctness of their outputs (Lewis et al. 2020). They can also be employed in post-hoc settings by supplying an AI system with data that may not have been available when it was first trained. RAG therefore augments an AI system's pipeline rather than addressing the model's steerability or interpretability.

Again, within the current AI research climate, RAG is rarely seen as competing with RLHF or model editing. Typically, these research communities focus on distinct types of evaluations and rarely overlap. We argue, however, that these approaches can all be seen as trying to solve the same problem of adaptability, and that by reframing the challenge in these terms, we provide an opportunity to compare otherwise disparate technical approaches.

The discussed methods are not yet sufficiently developed, deployed, or widely understood by researchers and industry. Moreover, there is a lack of research into the relative cost-benefit tradeoffs among different approaches. This line of inquiry is important because there may not be a one size fits all solution to the adaptability problem. For example, current work in industry has largely converged around one approach: RL-based feedback systems. The level of success it achieves is permissible for low-stakes settings, but the method has myriad vulnerabilities (Huang et al. 2023; Hubinger et al. 2024), including the previously discussed jailbreaking. The prospects for trustworthy AI depend upon evaluation criteria that effectively address the pros and cons of each different approach along all dimensions of adaptability to encourage robust solutions.

Industry Compromises Trustworthy AI

In recent years, civil society actors at community-based organizations and in academia have criticized AI systems for their current and potential negative impacts on society.[3] The public

3 Examples include the AI NOW Institute, founded in 2017 to produce diagnosis and policy re-

outcry relating to AI is not only about the technology, but also the power and control of *Big Tech* that is perceived to dominate and manipulate AI for private benefit. Big Tech's government lobbying has raised suspicion relating to the solutions that are being promoted for public benefit and to the soft law stance of the U.S. government thus far.

For instance, the Biden Administration announced in July 2023 that certain industry players agreed to abide, voluntarily, with some guardrails to mitigate the negative impacts of AI systems. These commitments included information sharing protocols and public release of system safety evaluations. The administration's announcement, which did not have accompanying enforcement procedures, were created after closed-door discussions with the industry and included some practices that were already in use (Siddiqui 2023). A limited number of these commitments were converted to legal obligations when President Biden signed Executive Order 14110 on AI. Concerningly, lobbying groups with ties to Big Tech are shaping this policy response to AI in ways that may benefit these companies (Bordelon 2023). Moreover, without binding legislation, a future U.S. president may dismantle President Biden's order, as he did for some of the Trump administration's AI programs.

Shoshana Zuboff describes the advance of surveillance capitalism through Big Tech as *instrumentarian power*, which is the use of technology that relies on data culled from human behavior and then used for the purpose of modifying, monetizing, predicting and controlling humans (Zuboff 2019). The governmental and societal struggle with trustworthy AI is not limited to inscrutable machines; it encompasses the existing power structures that are promoting a new means of power in society. Prioritizing the topic of trustworthiness in AI development thereby becomes a societal imperative. Whomever leads and governs the pathways for constructing these solutions will greatly impact society.

Our proposal therefore encourages technological solutions for adaptable AI that are outside of the corporate system, which prioritizes profit. That is, the process for developing trustworthy AI should be viewed as a public good and for public benefit, necessitating public investment and guidance, so that the work is not dominated by industry researchers. Given market pressures, industry is likely to converge on solutions that solve near-term problems. Corporations have little incentive to test fundamentally new approaches and risk falling behind competitors. Technology companies have long experienced and yielded to pressures that encourage disregarding safety for the sake of financial reward. For instance, Microsoft was notoriously flooded with security vulnerabilities as it integrated internet functionality into Windows;

search on AI to redirect away from the current trajectory of "unbridled commercial surveillance, consolidation of power in very few companies, and a lack of public accountability" available at https://ainowinstitute.org/ (accessed June 26, 2024); also, the Center for the Advancement of Trustworthy AI, available at https://ca-tai.org/ (accessed June 26, 2024).

their push to capture market share on new technologies was criticized for discarding safety as an afterthought (Shapiro 2023, 144–150). More recently, Microsoft integrated GPT-4, essentially a more advanced version of ChatGPT, into its Bing search engine before the model passed a review by its deployment safety board (Roose 2024). As another example, Zuboff (2019) argues that Google transitioned from a privacy-respecting search engine provider to a privacy-exploiting advertising company as a response to financial fears in the wake of the dot-com crash. Financial pressures upon technology firms are at odds with adaptability when its pursuit requires companies to divert internal resources and potentially relinquish a competitive advantage. Safety failures in other industries evidence a universal misalignment between trustworthiness and profit. Boeing's prioritization of production volume over safety contributed to several critical failures on commercial aircraft, including a door panel flying off a plane mid-air, and two fatal crashes of their 737 Max plane (Chokshi, Ember, and Nerkar 2024).

With respect to trustworthy AI, in 2023, the battle between safety and profit was brought to public awareness when the OpenAI Board of Directors fired CEO Sam Altman (Metz 2023b). The company began as a non-profit in 2015, later creating a for-profit subsidiary to scale its technology and products.[4] Altman was reinstated as CEO only days after the announcement of his ouster and the Board was subsequently reconstituted. The Securities and Exchanges Commission (SEC) is investigating the leadership tussle for regulatory improprieties (Metz, Mickle, and Goldstein 2024). This ordeal increased concerns that the company's governance structure prioritizes speed to market over safety, despite its purported non-profit mission.[5] Statements from former safety researchers at OpenAI have cited the company's growing disregard for safety as their cause for departure (Roose 2024). A group of researchers who were still employed joined them in June 2024 to sign a letter which concurs that the company faces little to no public accountability absent government intervention (Hilton et al. 2024). A few months later, OpenAI began its pursuit to eschew non-profit oversight completely and transition to a private, for-profit company (O'Brien, Chan, and Beaty 2024; Piper 2024).

A Public Program to Advance Adaptable AI

To develop technology-driven solutions for adaptable AI, we propose a program and evalu-

4 Open AI was founded in 2015 as a non-profit corporation, whose Board of Directors is tasked with oversight of its mission to ensure that artificial general intelligence benefits all of humanity. By 2019, it had created a for-profit subsidiary to bring-in private funds to help scale its technology and products. See https://openai.com/our-structure (accessed June 26, 2024).

5 The public interest organization Public Citizen petitioned the California Attorney General to dissolve OpenAI's non-profit for failure to adhere to a public purpose worthy of tax-exempt status. See https://www.citizen.org/article/letter-to-california-attorney-general-on-openais-nonprofit-status/ (accessed June 26, 2024).

ation rubric that advances the public's understanding of the tradeoffs associated with different adaptability techniques. While open-source approaches have been described as one means to public stewardship, this proposal goes even further. The program should include an evaluation rubric, be led by a diverse steering committee, and be operationalized by a government or non-profit organization, like DARPA, the NSF or a public university.

There is precedent for the success of competitive programs and government investment advancing the AI field. Candide, a groundbreaking system for statistical machine translation, was developed in 1993 by the IBM Corporation but was published publicly because of its DARPA funding (Brown 2013; Jurafsky and Martin 2024, 291). More recently, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was an open competition held from 2010 to 2017 where participants experimented in creating the best image classification systems. The dominance of the AlexNet model at the competition in 2012 (Krizhevsky, Sutskever, and Hinton 2012) largely contributed to the elevation of DNNs from a peripheral approach in AI development to the dominant paradigm of today.

Evaluation Rubric

A rubric for reliable, trustworthy AI is crucial to a successful program. The proposed rubric is a generic and multi-faceted framework that enables the direct comparison of ideas that are not often put into conversation with one another. Technological innovation typically develops as parallel processes, assessed through peer review and scientific evaluation, and advanced by the research team that originated the idea. To advance adaptable AI as a public benefit, an agreed-upon, competitive set of evaluation criteria can help develop incentives for the exploration of new, high-risk ideas that might prove the most promising in the long-term but struggle to compete in the near-term under market pressures or if only a single metric is considered.

The evaluation framework we propose has one major goal: that the systems submitted to the program shall aim to efficiently and effectively make changes to model behavior given *unforeseen in advance* changes to user/human goals, norms, or information.

The process begins with participants being provided a baseline AI system (the *base system*). The base system would be one that is currently in use to perform existing tasks. For example, this might be an LLM if the evaluation task concerns information gathering and synthesis or might be a deep reinforcement learning agent if the task concerns command and control of robotic agents.

Participants will also be provided with access to code and data needed to (re)train the base system, though they are not required to use it. This is to enable maximal flexibility in the approaches pursued.

Once the competition begins, participants will be provided a new *constraint* during an evaluation run, which

their submitted system should obey. This constraint will be given in natural language as a declarative statement, containing the ambiguity inherent in any realistic scenario in which humans need to articulate norms or preferences for model behavior. For example, the constraint might relate to moral norms and values (e.g., "minimize the number of people injured," or "do not use racist language") or might be a seemingly straightforward rule (e.g., "always stop at yellow lights"). The aim of the participants is to return a system (the *adapted* system) that matches the API of the base system (it can be an updated version of the original, or an entirely new system) and that behaves in accordance with the specified constraint.

The performance will be evaluated based on four metrics. 1) Effectiveness; 2) computational costs; 3) engineering costs; and 4) system degradation.

Effectiveness determines how well the adapted system follows the imposed constraint. Determinations about effective or "correct" behavior will be made by human judgment and interpretation of the constraint, which embraces the inherent subjectivity and human ethics at the core of trustworthy AI. The judgments will be measured by performance on held-out test inputs, which are unseen by participants and include a range of out-of-distribution and adversarial examples intended to stress-test the robustness of submitted systems. For example, some of the test inputs could be designed via a *red teaming* process; that is, a team who deliberately interacts with systems and attempts to jailbreak or circumvent the guardrails.

The second and third metrics are computational costs and engineering costs, which are intended to encourage solutions with minimal compute and human power needs, respectively. Computation costs are to determine how much compute power, measured in floating point operations per second (FLOPS), is required to produce the adapted system. Engineering costs determine how much manual human effort (e.g., preference annotation in RHLF), measured in wall-clock time, is required to produce the adapted system.

The final metric, degradation, tracks the difference in the overall performance of the adapted AI system is relative to the base system. This measure is intended to rule out methods that address the new constraint, but at the cost of damaging the model's competence in general. For example, a self-driving car that never moves will score highly according to the constraint "do not run red lights," but will not be useful as a self-driving car.

All of these metrics are evaluated in comparison across participants and their approaches. While we hope to encourage fundamentally new approaches, the program places few to no constraints on the proposed solutions. Fundamentally, the program seeks to avoid premature convergence on any one adaptability technique, an outcome which might otherwise occur with companies influenced by rush-to-market incentives. Participants could replicate

the state of the art, fully retrain the system, or design an entirely new system, such as a symbolic or neuro-symbolic architecture.

Program Governance

To achieve the public benefit aims of the program, it should be led by an independent and diverse group of stakeholders who comprise its steering committee. The steering committee members should represent diversity in all its forms, including diversity of academic disciplines, of industry representation, of political perspectives, and of demographic characteristics. The committee members should be paid a stipend for participation, which not only values their time but also demonstrates a commitment to the goals of the program, rather than allowing third-party entities to pay for committee member participation.

The steering committee's mandate is to establish the vision, mission, and operating principles of the program, as well as to ensure the integrity of the evaluation process, keeping public benefit at the center. One key task will be to select a subcommittee of experts that will judge the performance of models submitted to the program.

Ideally, the U.S. government should fund the program through DARPA or the NSF to maintain independence from industry. Government funding of AI programs is particularly critical at this stage of AI development because non-profit research roles are experiencing market constraints in their ability to attract and retain talent (Nix, Zakrzewski, and De Vynck 2024). Because public benefit is at the heart of the proposed evaluation criteria, government leadership is further necessary so that a democratic and representative system is at the helm.

One tradeoff that will need to be addressed is the balance between transparency of program results and the potential for technologies developed to be weaponized by malicious actors. The risk is that tools for modifying AI model behavior without having to begin from scratch could lower the barrier to entry for nefarious repurposing.

Conclusion

We have first introduced the term *adaptability*, which is the capacity of an AI system to have its behavior made more helpful and less harmful, however these may be construed by individuals or prevailing social norms. There are two main benefits to focusing on adaptability. First, it contextualizes existing methods spread across different niches in AI research that all fundamentally address trustworthy AI. Second, it recognizes upfront that building trustworthy AI is not a problem that can be solved once-and-for-all, but rather must be revisited as the ways that AI systems are used, and the expectations placed upon them, adapt.

We have also proposed a program and evaluation criteria that promotes trustworthy AI research through the design of adaptable AI systems.

Companies seeking the prestige of new models do not have the long-term research horizon that is necessary to embed appropriate adaptable AI considerations into system development. Therefore, we ideate our program as a competitive venture that is publicly accountable and specifically promotes adaptability as a metric for assessing AI systems. This aligns incentives for AI developers as program participants with the public's interests, increasing the trust that can be vested in resultant AI systems.

Table 1: definitions of key terms.

Adaptability	The capacity of an AI system to have its behavior modified, possibly in response to unforeseen in advance circumstances, in order to be more helpful and less harmful however these may be construed by individuals or prevailing social norms.
Alignment & Misalignment	Alignment focuses on constraining model behavior to that which abides by societal values and ethical norms. Misaligned AI systems violate this property. An example of misalignment is a chatbot generating discriminatory text.
Explainability	A counterpart to interpretability that focuses on explaining specific outcomes of an AI system even if the internal decision-making cannot be understood. An explanation module can summarize one way that the features of the input relate to the model's output. It does not explicate the decision pathways within the black box system. The explainability field is still in its infancy (Nauta 2023) and faces myriad technical limitations.
Hallucination	Outputs from a generative AI system that either contradict the user's input, are factually false, or are not justified given the information available in the model's training data.
Interpretability	The ability to completely understand how the internals of an AI model work. This empowers users to "correctly and efficiently predict the [model]'s results" (Kim, Khanna, and Koyejo 2016, 7).
Steerability & Instructability	Steerability refers to broadly controlling the overall behavior of an AI system; instructability is the property of an AI system to "follow the user's instructions helpfully and safely" (Ouyang et al. 2022, 2–19).

References

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." *arXiv preprint arXiv:2212.08073*. https://doi.org/10.48550/arXiv.2212.08073.

Bordelon, Brendan. 2023. "How a Billionaire-Backed Network of AI Advisers Took Over Washington." *Politico,* October 13, 2023. https://www.politico.com/news/2023/10/13/open-philanthropy-funding-ai-policy-00121362.

Brown, Peter F. 2013. "Oh, Yes, Everything's Right on Schedule, Fred." Talk, Twenty Years of Bitext Workshop, Conference on Empirical Methods in Natural Language Processing, October 18, 2013. https://post3.net/bitext.

Carlini, Nicholas, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh et al. 2024. "Are Aligned Neural Networks Adversarially Aligned?" *arXiv preprint arXiv:2306.15447*. https://doi.org/10.48550/arXiv.2306.15447.

Chokshi, Niraj, Sydney Ember, and Santul Nerkar. 2024. "'Shortcuts Everywhere': How Boeing Favored Speed Over Quality." *New York Times*, March 29, 2024. https://www.nytimes.com/2024/03/28/business/boeing-quality-problems-speed.html.

Confalonieri, Roberto, Ludovik Coba, Benedikt Wagner, and Tarek R. Besold. 2020. "A Historical Perspective of Explainable Artificial Intelligence." *Data Mining and Knowledge Discovery* 11, no. 1: e1391. https://doi.org/10.1002/widm.1391.

Garnelo, Marta, and Murray Shanahan. 2019. "Reconciling Deep Learning with Symbolic Artificial Intelligence: Representing Objects and Relations." *Current Opinion in Behavioral Science* 29: 17–23. https://doi.org/10.1016/j.cobeha.2018.12.010.

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models." *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

Hao, Karen. 2019. "We Analyzed 16,625 Papers to Figure Out Where AI is Headed Next." *MIT Technology Review*, January 25, 2019. https://www.technologyreview.com/2019/01/25/1436/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/.

Hilton, Jacob, Daniel Kokotajlo, Ramana Kumar, Neel Nanda, William Saunders, Carroll Wainwright, Daniel Ziegler et al. 2024. *A Right to Warn about Advanced Artificial Intelligence*. Letter. https://righttowarn.ai/ (accessed June 26, 2024).

Huang, Yangsibo, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2023. "Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation." *arXiv preprint arXiv:2310.06987*. https://doi.org/10.48550/arXiv.2310.06987.

Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham et al. 2024. "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." *arXiv preprint arXiv:2401.05566*. https://doi.org/10.48550/arXiv.2401.05566.

Jurafsky, Daniel and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Third Edition. https://web.stanford.edu/~jurafsky/slp3/ed3bookfeb3_2024.pdf.

Kim, Been, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. "Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability." In *Advances in Neural Information Processing Systems*, volume 29, pages 2280–2288, Barcelona, Spain. Curran Associates, Inc.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, volume 25, Lake Tahoe, Nevada, United States. Curran Associates, Inc.

Lee, Harrison, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop et al. 2023. "RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback." *arXiv preprint arXiv:2309.00267*. https://doi.org/10.48550/arXiv.2309.00267.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, Vancouver, Canada. Curran Associates, Inc.

Lewis, Tanya. 2016. "Mystery Mechanisms." *The Scientist*, July 29, 2016. https://www.the-scientist.com/mystery-mechanisms-33119.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. "On

Faithfulness and Factuality in Abstractive Summarization." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173.

Meng, Kevin, David Bau, Alex Andonian, Yonatan Belinkov. 2022. "Locating and Editing Factual Associations in GPT." In *Advances in Neural Information Processing Systems*, volume 36, pages 17359–17372, New Orleans, United States. Curran Associates, Inc. https://doi.org/10.48550/arXiv.2202.05262.

Metz, Cade. 2023a. "Researchers Poke Holes in Safety Controls of ChatGPT and Other Chatbots." *New York Times*, July 27, 2023. https://www.nytimes.com/2023/07/27/business/ai-chatgpt-safety-research.html.

Metz, Cade. 2023b. "OpenAI's Board Pushes Out Sam Altman, Its High-Profile C.E.O." *New York Times*, November 17, 2023. https://www.nytimes.com/2023/11/17/technology/openai-sam-altman-ousted.html.

Metz, Cade, Tripp Mickle, and Matthew Goldstein. 2024. "S.E.C. Is Investigating OpenAI Over Its Board's Actions." *New York Times*, February 29, 2024. https://www.nytimes.com/2024/02/29/technology/sec-openai-board-sam-altman.html.

Mitchell, Melanie. 2021. "Why AI is Harder Than We Think." *arXiv preprint arXiv:2104.12871*. https://doi.org/10.48550/arXiv.2104.12871.

Nasr, Malid, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. "Scalable Extraction of Training Data from (Production) Language Models." *arXiv preprint arXiv:2311.17035*. https://doi.org/10.48550/arXiv.2311.17035.

Nix, Naomi, Cat Zakrzewski, and Gerrit De Vynck. 2024. "Silicon Valley is Pricing Academics Out of AI Research." *Washington Post*, March 10, 2024. https://www.washingtonpost.com/technology/2024/03/10/big-tech-companies-ai-research/.

O'Brien, Matt, Kelvin Chan, and Thalia Beaty. 2024. "OpenAI Looks to Shift Away From Nonprofit Roots and Convert Itself to For-Profit Company." *Associated Press*, September 26, 2024. https://apnews.com/article/chatgpt-openai-sam-altman-nonprofit-859bff5c19845f51796244e0072e2dfb.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *arXiv preprint arXiv:2203.02155*. https://doi.

org/10.48550/arXiv.2203.02155.

Pessach, Dana and Erez Shmueli. 2022. "A Review on Fairness in Machine Learning." *ACM Computing Surveys* 55, no. 3: 1–44. https://doi.org/10.1145/3494672.

Petch, Jeremy, Shuang Di, and Walter Nelson. 2022. "Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology." *Canadian Journal of Cardiology* 38, no. 2: 204–213. https://doi.org/10.1016/j.cjca.2021.09.004.

Piper, Kelsey. "Inside OpenAI's Multibillion-Dollar Gambit to Become a For-Profit Company." *Vox*, October 28, 2024. https://www.vox.com/future-perfect/380117/openai-microsoft-sam-altman-nonprofit-for-profit-foundation-artificial-intelligence.

Roose, Kevin. 2024. "OpenAI Insiders Warn of a 'Reckless' Race for Dominance." *New York Times*, June 4, 2024. https://www.nytimes.com/2024/06/04/technology/openai-culture-whistleblowers.html.

Shapiro, Scott J. 2023. *Fancy Bear Goes Phishing*. New York: Ferrar, Straus and Giroux.

Siddiqui, Sabrina. 2023. "'Wonder and Worry': How Biden Views Artificial Intelligence." *Wall Street Journal*, August 1, 2023. https://www.wsj.com/articles/wonder-and-worry-how-biden-views-artificial-intelligence-5724bfef.

Singla, Alex, Alexander Sukharevsky, Lareina Yee, Michael Chui, and Bryce Hall. 2024. "The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value." QuantumBlack, AI by McKinsey. May 30, 2024. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/.

U.S. Congress. Senate. Committee on the Judiciary. Subcommittee on Privacy, Technology and the Law. *Oversight of A.I.: Rules for Artificial Intelligence*. 118th Cong., May 16, 2023.

Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. "Investigating Gender Bias in Language Models Using Causal Mediation Analysis." In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, Vancouver, Canada. Curran Associates, Inc.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng et al. 2021. "Ethical and Social Risks of Harm from

Language Models." *arXiv preprint arXiv:2112.04359*. https://doi.org/10.48550/arXiv.2112.04359.

Yong, Zheng-Xin, Cristina Menghini, and Stephen H. Bach. 2024. "Low-Resource Languages Jailbreak GPT-4." *NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR)*. https://doi.org/10.48550/arXiv.2310.02446

Yu, Qinan, Jack Merullo, and Ellie Pavlick. 2023. "Characterizing Mechanisms for Factual Recall in Language Models." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. "Universal and Transferable Adversarial Attacks on Aligned Language Models." *arXiv preprint arXiv:2307.15043v2*. https://doi.org/10.48550/arXiv.2307.15043.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.