**From:** Future of Life Institute
**Re:** HELP Committee White Paper on Advantages and Drawbacks of AI in US Healthcare System, in the Classroom, and in the Workplace
**Date:** September 22, 2023

The **Future of Life Institute (FLI)** is an independent non-profit dedicated to realizing the benefits of emerging technologies and minimizing their risks. Since 2014, FLI has worked closely with experts in government, industry, civil society, and academia to envision and encourage positive futures through policy research and advocacy, grant-making, and educational outreach.

In 2017, FLI coordinated development of the Asilomar AI Principles, one of the earliest and most influential frameworks for the governance of artificial intelligence (AI). FLI serves as the United Nations Secretary General's designated civil society organization for recommendations on the governance of AI, and has been a leading voice in identifying principles for the responsible development and use of AI for nearly a decade.

## Background

As advances in the capabilities of AI systems accelerate, AI is increasingly being integrated across the healthcare, education, and labor domains. While AI carries promise of significant benefits to these industries through improved product service, access, and innovation, integration of AI in such sensitive fields must be carried out with caution, mindful of considerable risks.

Three characteristics of AI systems present significant challenges for their integration in critical domains: 1) The functionality and decision-making criteria of AI systems, i.e. the processes through which outputs derive from a given set of inputs, are opaque and unpredictable; 2) AI systems must be trained on large amounts of past, potentially outdated data; and 3) many AI systems are "general-purpose," and may be applied in ways not intended or expected by their developers.

Because the criteria by which AI systems arrive at output is not evident to or interpretable by humans, the risk that these systems make decisions or produce output that is subtly misaligned with human values is particularly high. Coupled with the fact that AI systems are primarily trained through the identification of probabilistic patterns in past data, this also means the training underlying the output of the systems may reflect antiquated norms and practices, and may therefore further entrench existing cultural biases and power structures. Broader adoption of these systems across a greater breadth of critical domains increases these risks considerably.

A cautionary analogy for this type of misalignment can be seen in the rise of social media, which initially promised outstanding benefits for interpersonal connection and social discourse. Unbeknownst to users, however, the recommendation algorithms determining content displayed to those users was optimized for increasing engagement, rather than for providing those potential benefits. The result has been a fracturing of public discourse and a significant detriment to mental and social health due to the primacy of negative

emotion in driving engagement. Notably, the decision to prioritize engagement was intentional and human-derived, but the preference for negative emotion emerged as a misaligned byproduct of that decision-making process. In the case of advanced AI systems, these potentially-harmful preferences driving output are even more difficult to identify, and therefore even more insidious, if adequate oversight and regulation of AI in critical domains is not imposed.

The capacity for general-purpose systems to be adapted to or employed in a variety of contexts presents additional obstacles from both a practical and legal standpoint. Inevitably, general-purpose AI systems will be deployed in contexts for which they were not specifically designed or tested, but may nonetheless provide convincing output that appears valid, whether or not it actually is. Particularly in domains such as education, healthcare, and labor, extensive testing is necessary to ensure that students, patients, and workers do not face undue hardship, yet well-documented human biases for overestimating the validity of machine-derived information mean that adoption of unreliable systems, even only as support for human decision-making, could lead to harm.

Furthermore, when systems can be applied in numerous ways including some that are not intended or expected, it is not clear which existing legal frameworks are applicable, nor which bodies have jurisdiction to carry out enforcement actions. For example, if a general-purpose chatbot is used by a teacher to personalize lesson plans and also by a healthcare worker to evaluate symptoms, is the information collected by the controller of that chatbot bound by medical privacy laws (e.g. HIPAA) or privacy laws pertaining to student records (e.g. FERPA)? Is the developer of the AI system absolved of liability because they didn't explicitly design the system to be used in that way, or must they account for reasonably feasible uses before deploying the product?

The potential benefits of advanced AI systems in the healthcare, education, and labor domains are enticing, but we urge Congress to emphasize frameworks that are cautious and deliberate, lest we lose control and understanding of some of our most essential societal functions. This memo outlines some brief comments to the Ranking Member of the HELP Committee based on the White Paper titled "Exploring Congress' Framework For The Future Of AI: The Oversight And Legislative Role Of Congress Over The Integration Of Artificial Intelligence In Health, Education, And Labor" to aid in developing such frameworks.

**Comments and Recommendations**

1. **Analogies to Software (p. 1).** The White Paper draws early analogies to traditional software, pointing out that "software is regulated based on how it is used, whether in power plants, airplanes, or X-ray machines." However, advanced AI has some fundamental distinctions from traditional software that complicate this analogy, with important policy implications.

    a) **AI Systems are much more opaque than traditional software.** If a computer programmer wants to scrutinize why software is behaving in a certain way, the

programmer can simply review the source code underpinning the software. This is not the case for advanced AI systems - the reasons for their behavior are often completely opaque to their developers. While AI systems are partly reliant on computer code, their reasoning capabilities come from machine-learning functions which are incredibly difficult to inspect and understand. Additionally, while the Federal Trade Commission requires that traditional sponsored digital content be labeled as such, developers of AI systems are under no obligation to disclose conflicts of interest that may materially impact the output of those systems. As an example, either as a byproduct of a training dataset or as an intentional, profit-seeking design choice, a chatbot intended to serve as a virtual medical assistant may be intrinsically biased to prefer one pharmaceutical over another and preferentially recommend that pharmaceutical, even when it is not optimal for a given patient.

b) **Traditional software is much less general-purpose than AI.** Software tends to have a limited set of horizontal capabilities, confined by the code of the computer program that generates it. This is not the case with advanced AI systems, which are becoming increasingly general-purpose over a wide range of domains. This is especially important in light of c), below.

c) **It is currently impossible to predict the behavior and function of AI systems.** Traditional software executes functions in a predictable fashion, as the functions it can execute are hardwired into the program itself. With AI systems, in contrast, there is great unpredictability with regard to what functions they can fulfill - previously latent functions that developers were completely unaware of have been discovered with nearly every major general-purpose AI system.

d) **AI systems continue to learn and change after release.** As AI systems have been engineered to learn and adjust, they continue to change meaningfully even after they have been released to consumers. This is unlike software which does not change after launch unless its code is intentionally changed or its software is updated by a user. In other words, software requires considerable human intention to change, while AI systems continue to learn and change even in the absence of intentional human efforts.

All of these distinctions create the impetus for much greater consideration and caution before integrating AI into critical systems such as healthcare compared to traditional software. This also merits dedicating more resources to developing techniques for making AI systems more explainable and predictable before integration.

2. **A Complementary Approach to Regulation (pp. 1-2).** While it is true that one-size-fits-all regulation for AI could overlook important differences in applications and risks across domains, only having regulation based on individual applications also carries unique risks. Many of the

risks from advanced AI systems are local to the development of these systems themselves and the actions of developers. We suggest a complementary approach that includes both domain-specific and more general oversight. We recommend leveraging existing regulatory frameworks in specific domains, with adjustments where necessary, to oversee narrow AI systems designed for specific functions, such as treatment recommendations or student assessment, along with the establishment of a centralized regulatory body to monitor, evaluate, and regulate advanced general-purpose AI systems that may cut across several jurisdictional domains and are not adequately addressed by existing, domain-specific regulations.

Accordingly, AI subject matter expertise must be distributed across agencies, but the centralized body would additionally serve an advisory function for other agencies as they enact and enforce regulations pertaining to AI, and would coordinate rapid response in the event of an emergency caused by an AI system. Agencies should leverage their existing protocols for evaluating AI technology specific to functions within their jurisdiction. For instance, the FDA can expand on existing procedures for evaluating suitability of medical devices and pharmaceuticals to similarly review AI systems integrated in medical devices or used for drug discovery to ensure their safety before release.[1] Given the breadth of uses and potential risks associated with powerful general-purpose AI systems, the centralized authority reviewing these systems should adopt its own robust, independent auditing and licensure regime to test such systems for bias and fairness, robustness and reliability, suitability in different contexts, and potential for increasing catastrophic risks prior to their deployment. Only if these systems are demonstrably safe and ethical should they receive licensure.

3. **Clarification of Liability (pp. 4-5).** We are pleased to see that the White Paper identifies the need for "a clear understanding of potential liability around the use of AI" to provide the necessary financial incentives for profit-driven AI developers to exercise abundant caution. Due to the complex value chain involved in the development and deployment of AI systems, it can be unclear what theories of liability are applicable in which circumstances. When an AI system provides incorrect information that results in harm, for instance, is the developer of the system responsible for making the harmed party whole, is it the professional deploying that system, or is it neither? The lack of clarity increases the likelihood of a prolonged legal battle with room for plenty of legal maneuvering and exploitation of loopholes, to the benefit of well-resourced corporations and to the detriment of the vulnerable patients, students, and workers who may be harmed.

[1] We note that the FDA "pre-certification pilot for software treated as medical devices that would certify software developers as opposed to the products themselves" does not constitute an adequate review process, as the potential risks and insufficiencies of AI systems can vary considerably, even when manufactured by the same developer for the same purpose. Because the systems are, generally speaking, self-trained, even their developers may not be aware of precisely how they function and what their shortcomings may be.

Due to the opacity of the latest generation of AI systems, those deploying the systems often have little knowledge of how the systems work, what information they rely on for their output, the contexts in which they produce reliable output, and the ways in which they were trained and tested. To the extent any of this information is discernible, it is generally possessed by the developer of the system, and is unlikely to be passed along to the deployer. As such, the responsibility for ensuring a system that is released for public use should reasonably fall on the developer of the system, unless substantial modifications are made to the system by the deployer that materially affect its functionality. This, along with the inherent risks involved with releasing systems that cannot be fully understood, necessitates a strict liability regime that holds the developer accountable for harms resulting from systems designed to be deployed in critical contexts and powerful general-purpose systems. Such a liability regime would encourage meticulous review and red-teaming of systems before release, to the benefit of the general public. In the event an AI system is materially modified before use, joint and several liability based on the totality of the circumstances would be appropriate, depending on the extent to which the design and marketing of the system is implicated in the resulting harm.

That said, those who maintain a fiduciary responsibility to the best interests of their clients, such as in a doctor-patient relationship, also bear responsibility for adopting and relying on such systems only to the extent that the systems can be affirmatively relied on to fulfill that responsibility. The duties of care and loyalty associated with a fiduciary responsibility should not be foregone solely because new technology was involved in the decision-making process.

4. **Dual-use Concerns with AI Applications in Drug Discovery and Development (pp. 3-5).** As stated in the White Paper, AI advancements promise potentially speedier routes toward drugs that have greater levels of efficacy and safety. However, these benefits should be evaluated in light of dual-use concerns. Biomedical advancements in AI raise potential risks of making it easier for malevolent actors to cause harm by reducing the resource and expertise bar to manufacture dangerous pathogens. Establishing FDA guidance building on earlier comments from the CDC and NIH can help establish guardrails that ensure we maximize the benefits and minimize the downsides of biomedical research using AI. Guidelines detailing requirements for withholding the publication of model weights underlying systems that present particularly significant dual-use risk should also be developed to prevent the proliferation of AI systems that could feasibly facilitate the development of highly-destructive toxins and pathogens.

5. **Privacy Concerns with AI Applications (pp. 5-10).** As the White Paper notes, healthcare providers create, receive, store, and transmit large quantities of sensitive patient data which come under HIPAA and corollary protections. AI systems have been shown to have significant cyber vulnerabilities which put this data at risk of exploitation and misuse by malevolent actors.[2] Large

[2] Our team has done considerable investigation into the the nexus between cybersecurity and artificial intelligence. We are happy to provide more details on this front in future correspondence.

language models and other AI innovations in the classroom can also collect a large amount of data about students, including their academic performance, learning habits, and personal information. We need to be sure that the integration of AI into education does not lead to unintentional bias, misuse, or exploitation of the personal information of students by ensuring compliance of these systems with the Family Educational Rights and Privacy Act (FERPA) and by establishing guardrails before deployment.

6. **Allowing Time to Cushion the Disruptions in the Labor Market (pp. 11-13).** As the White Paper notes, there is a serious concern about job displacement within the labor economy as advancements in and adoption of AI systems accelerate. Estimates vary as to the extent to which jobs will be impacted by these developments, and this uncertainty underscores the need for additional reflection and fact-finding before AI systems are integrated wholesale into these critical facets of society. We believe that the generative AI, and especially general-purpose AI systems, of the last year and several years ahead are different from other 'AI' - or automation - developments in the past, threatening more jobs than prior developments. In either case, it is critical to ensure that disruptions to the labor market are minimized by ensuring that workers have the training and skilling opportunities necessary to transition to positions that will likely require much greater levels of digital literacy. AI developments from major corporations are taking place at a breakneck pace, which reduces the window of time for the labor economy and government to adjust to the transition. Indeed, potential disruption effects on the labor economy highlight the broader need to slow down the rapid pace of AI development and adoption to allow our institutions, labor market, and society as a whole to adjust. The mechanisms for such a slow down are not currently in place - development of increasingly powerful AI systems is happening outside of any well-established governance infrastructure that could require safety and reliability assurances before systems are deployed. Developing this infrastructure is essential to provide the levers necessary to calibrate AI integration with the tolerances of the economy and society.

## Conclusion

We are grateful to the Ranking Member for his attention to the important issue of AI integration into the domains of healthcare, education, and labor, and appreciate the White Paper's recognition of the need for balance and consideration when adopting these emerging technologies. While we have undertaken efforts to identify the risks less oft-discussed in the White Paper, we share the goals of the HELP Committee in ensuring that AI integration is maximally beneficial to our healthcare providers, patients, teachers, students, workers, and society at large.

We welcome further correspondence, and encourage the Ranking Member and his staff to reach out to us with any questions or requests for additional information. We can be reached at policy@futureoflife.org.