American Policy for Human-AI Collaboration:

Battista, Jude

**Executive Summary**

As members of the Future of Life Institute, we believe that the future of humanity is worth preserving. While humanity faces a daunting array of threats as we move towards the second half of the twenty-first century, none pose the existential danger of the development of Artificial General Intelligence (AGI). Once an AGI surpasses human intelligence, it becomes capable of designing even better AI systems, leading to an AI cascade that rapidly outpaces torpid biological evolution. The resultant super-intelligent system will inevitably supersede human control. If humanity wishes to maintain or improve upon its current condition, it must ensure that the AGI’s goals are aligned with those of humanity.

While the United States National Science and Technology Council (NSTC) has implemented a nascent policy on Developing Effective Methods for Human-AI Collaboration, the current iteration prioritizes innovation over safety. NSTC recommends four principles: seeking new algorithms for human-aware AI, developing AI techniques for human augmentation, developing techniques for visualization and human-AI interfaces, and developing more effective language processing systems. Notably absent is any mention of ensuring that the AI shares human values or goals in these interactions. As AI becomes more advanced, any such divergence may well jeopardize its human partner.

By relying on private industry to drive innovation, this policy decentralizes control over AI development, putting humanity’s fate in the hands of the least safe actor in that space. Because humanity cannot afford the apocalyptic consequences of failure, it is incumbent upon the United States government to ensure that American AI research is conducted responsibly, to guarantee that AI goals are consonant with human goals.

In light of these factors, the Future of Life Institute recommends that the Senate Subcommittee on Space, Aeronautics, and Related Sciences author legislation immediately nationalizing all United States Artificial Intelligence research programs under the umbrella of the Advanced Research Projects Agency. While this will sacrifice the benefits of competition in a free market it permits strict regulation of the research and coordination of resources.

**Background**

The development of Artificial General Intelligence (AGI), or a machine intelligence capable of operating at a human level on arbitrary tasks, may well be the single most disruptive event in human history. In 2015, Stephen Hawking described human-level artificial intelligence (AI) as “likely to be either the best or worst thing to ever happen to humanity.” (Tegmark, 2015). He speculates that a truly superhuman intelligence is both unpredictable and uncontrollable by merely human minds. If such an intellect has goals incongruous with those of biological entities, humanity runs the risk assumed by ants in the face of human urbanization. Even in the absence of malice, the ants may be swept away unnoticed, simply as a side of effect of progress. Other flavors of AI apocalypse are less dramatic but pose an existential threat to humanity as we know it. Perhaps humans become explicitly subordinate to AGI, or perhaps the AGI usurps the human agency we associate with free will. Seán Ó hÉigeartaigh, the Executive Director of the Centre for the Study of Existential Risk at the University of Cambridge, pointed out that the risks inherent to such a transformative technology only become more difficult to mitigate the longer we postpone addressing them (Ó hÉigeartaigh, 2015). In June of 2019, the Select Committee on Artificial Intelligence (SCAI) of the National Science and Technology Council (NSTC) released the latest version of the United States’ National Artificial Intelligence Research and Development Plan. This plan focuses on seven strategies for Artificial Intelligence (AI) research and development. Given the potential impact on the future of human development, the second strategy, Developing Effective Methods for Human-AI Collaboration, assumes particular import. Unfortunately, as it stands, this strategy is woefully underdeveloped.

**Policy Description**

NSTC predicts three ways in which AI and humans are likely to interact (NTSC, 2019, p. 16). The first envisions human and AI entities working in partnership. This may entail the AI performing supplemental tasks to support the human, or providing additional capacity for the human agent to access at need. An AI my also act as a failsafe, intervening or providing augmentation when task complexity spikes, temporarily overwhelming human capability. Finally, an AI may operate autonomously in place of human agents in environments or on missions which exceed human limitations, or to which humans are poorly suited.

The current policy revolves around four tenets: seek new algorithms for human-aware AI, develop AI techniques for human augmentation, develop techniques for visualization and human-AI interfaces, and develop more effective language processing systems (NSTC, 2019, p. 17-18). NTSC prescribes the development of algorithms designed to make AI a better partner for humans, by driving intuitive interactions between the pair. This entails broadening the goals and courses of action available to the AI, permitting it the scope of action necessary to adjust to human actions. This may run the gamut from shallow interaction, where the system suggests options to the user, all the way to deep learning algorithms that model the user’s intent.

Human augmentation covers a wide range of capabilities. An AI system may be form-specific such as a workstation, a wearable device, or a smart room designed to assist the user within its field of influence. Alternatively, the system may be domain-specific with an expert system lending capability to a human partner within a set of assigned tasks. The roles may also be reversed with AI systems handling the bulk of the task, calling on a human Subject Matter Expert (SME) to clarify any situations for which the AI is insufficiently trained.

As AI outstrips human capabilities, it also becomes more complex, necessarily rendering it more difficult to predict and understand, even for its developers. For a layperson to successfully interact with such an AI, the system must have a transparent, easily accessible interface, capable of translating large and complex datasets to the human. This requires a level of data abstraction likely beyond human capacity, making an AI responsible for learning to communicate with humans. The human-AI interface must be functional in a wide variety of environments, including those which constrain communication, requiring a substantial degree of autonomy on behalf of the AI system.

Improved language processing systems may ease the interface burden on an AI system. When an AI can understand fluent English and respond in kind, it allows even an untutored user to query the system iteratively to establish a common understanding of the task at hand and the capabilities the AI can offer. Speech also conveys emotional and contextual information absent in a rigid control interface, meshing with human-aware algorithms to provide a flexible interaction model.

**Problem Identification**

In its current incarnation, the NSTC policy more closely resembles wishful thinking than a plan carefully constructed to bring about a desired outcome. It is both short-sighted and narrow in focus, reading as though it was purchased by industry members if not directly dictated by the same. The policy repeatedly cites industry and academia as leaders in AI development, and seems content to maintain this status quo (NSTC, 2019, p. 14). As the subprime mortgage crisis demonstrates, the profit motive that drives a capitalist economy can easily override market actors’ better judgment, leading them to choose short-term gains over long-term health. The sheer power offered by successful AI implementation creates an even more precarious scenario. Not only is the temptation orders of magnitude greater than that of mere lucre, but tragedy no longer requires an entire industry to abandon its social conscience. Instead, a single slipshod AI implementation has the potential to irrevocably alter the course of human development. Further, given the unparalleled advantage of being the first to field a fully-armed and operational AI agent, market competition is likely to devolve into a race to the minimum acceptable outcome as competitors cut corners to expedite delivery. This dramatically increases the chances of an unacceptably flawed AI going live.

While the NSTC plan does propose the implementation of standards and benchmarks for AI development, as outlined in the policy these are largely focused on improving capabilities of and expediting the delivery of AI systems. Notably missing is any explicit requirement for the development of ethical AI. Also absent are any requirements for the goals of and restrictions on AI systems. Instead the emphasis is on “fostering” the development of AI in the commercial realm in keeping with President Trump’s directives to “to enable the creation of new AI-related industries and the adoption of AI by today's industries” and “reduce barriers to the use of AI technologies” (Trump, 2019). According to futurist Max Tegmark, control is correlated with intelligence (Tegmark, 2018). The creation of true AGI, or even a sufficiently advanced narrow AI along the correct axis will then likely be uncontrollable by human agents. This necessitates the installation of human-aligned goals and restrictions in the infancy of any AI system. Patching core modifications into any mature software system is always a fraught business. Doing so in a system as complex as any AGI must be is a Sisyphean task. Even the creation of human-compatible goals and restrictions looks to be a Herculean labor, as it first necessitates understanding and codifying goals for humanity, a complex multidisciplinary effort likely to require years of effort. The humanitarian goals need to have absolute primacy over performance or effectiveness goals. Even a single failure in this regard can create a cascade resulting in an AI whose goals dramatically diverge from those of humankind. Given the complexity and importance of the task, defining the goals of and strictures on AI agents would seem to be at least as urgent as performance standards, yet is thoroughly ignored in the current policy.

While the distribution of AI development between competing corporations, agencies, and universities undoubtedly drives innovation, it also creates a regulatory and enforcement nightmare and incentivizes subversion of any restrictions in place. In addition to the safety concerns previously discussed, this fragmented marketplace creates philosophical issues. The marketplace of ideas works well for easily replicatable systems. These enforce a degree of equity as competitors can easily overtake the market leader simply by adopting the current best technology and iterating on it. With opaque systems such as strong AI, there is no such guarantee. The first entity to create an AGI may obtain an insurmountable advantage as, even with precise replication steps, duplicating the result may be impossible. This possibility escalates asymptotically towards certainty as AI begins to iterate on itself, diverging rapidly from its initial design. This paucity of competition creates an aggregation rather than democratization of power, allowing those who own the means of production to increase and solidify the social and economic chasm between themselves and their supplicants. Should the United States wish to maintain even the pretense of a democratic society, it must ensure egalitarian access to this transformational technology.

**Potential Solutions**

Any solution to the weaknesses in the NTSC policy necessarily revolves around the establishment of goal-alignment standards for current and future AI development. Only by building AI systems around a core set of values that align the AI’s goals with those of humanity will developers be able to influence the course of AI progression once AGI becomes a reality. The question is what course of action best permits the establishment and enforcement of such protocols. The Future of Life Institute’s research to date indicates three possible solutions to these questions. The first is to use government oversight to strengthen the existing proposal. By implementing goal-alignment requirements for any domestic AI R&D program, NTSC could partner with industry and agency actors to promote the survival of the human race. While program control would remain with the original agents, the NTSC can act as a watchdog. Given the limited enforcement powers of an external group in this situation, compliance would need to be voluntary, preserving the agency of each actor. This approach also minimizes the interference of the government in the free market economy, which should be reassuring for privately owned businesses.

An intermediate solution involves emulating the military industrial complex. Congress could enact laws restricting the employment of broad AI to government actors, while leaving the research and development of such AI in private hands. In this case, the academic or industry actors function like defense contractors, with each one able to submit bids for the fulfillment of task-specific AI platforms as intermediate steps in pursuit of true AGI. This maintains elements of free market competition, while reducing the risk that AI goals become dissociated from national priorities.

The most extreme of the three solutions is for the federal government to nationalize all domestic AI research and development. While this is a violation of both property rights and the free market, it offers substantial advantages in terms of centralizing control over the manifold diverse AI programs currently being pursued in the United States. It also permits the NTSC to directly mandate compliance with goal-alignment regulations for all domestic AI programs. The most difficult option to implement, it also offers the greatest likelihood of effectiveness.

**Recommendation**

With good reason the United States government has historically claimed a monopoly on domestic research into transformational technology. The U.S. Army Corps of Engineers did not solicit bids on the Manhattan Project. Unlike traditional technology development which permits iteration over a prototype to define a path of constant improvement, some technologies demand to be treated with more respect. Moving fast and breaking things is a laudable ideology for disrupting the status quo. It is somewhat less appropriate for thermonuclear weapons, AGI, self-replicating nano-tech or any other system with a reasonable chance of kickstarting an apocalyptic positive feedback loop.

Because control is correlated with intelligence, superhuman AGI inevitably exerts some degree of control over humans, eroding their agency. The only question is to what degree human free will is affected. Should this determination devolve to corporate hands or be held by actors at least nominally responsible to the public? Citizens have already tacitly agreed to some degree of tyranny on the behalf of the government in order to promote public good. The government does have some degree of public accountability. Corporations not only have no such accountability but have repeatedly denied that they have any responsibility along that axis. In the landmark case of *Dodge vs. Ford Motor Co.*, the Michigan State Supreme Court affirmed the corporation’s legal responsibility to reward its shareholders monetarily rather than provide a social benefit. The consequences of AI development are simply too severe to be left to the vicissitudes of the market.

The need for accountability coupled with the non-existent margin for error mandate rigorous oversight of and close cooperation between American AI development programs. To ensure this, the Senate Subcommittee on Space, Aeronautics, and Related Sciences (SSARC) should immediately introduce legislation to nationalize all American AI research and development programs. With jurisdiction over both the National Institute of Standards and Technology and the Office of Science and Technology Policy, the SSARC is uniquely positioned to initiate action on the credible grounds of a scientific emergency. Further, Senators Moran and Schatz have dual membership on the Senate Subcommittee for Defense, easing the coordination with that subcommittee to place the newly consolidated research programs under the Defense Advanced Research Projects Agency (DARPA).

While the nationalization of resources is a drastic measure, it has substantial precedent in times of national emergency. During World War I, railroads and munitions plants were nationalized, and in the wake of the 11 September attacks, the federal government assumed control of airport safety, making it the responsibility of the newly formed Transportation Security Administration (TSA). While the risk from AGI development is both less immediate and less obvious than physical violence, the potential consequences vastly outstrip any conventional military threat, warranting an extreme response.

Since this solution consolidates power with the federal government, it should appeal to lawmakers. The appeal to public safety should prove adequate to sway the general public as it has with past measures such as the Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism, or Patriot Act. If necessary, further appeal can be made on behalf of national defense. In light of confirmed national programs from international actors such as China and Russia, a domestic rival program makes a reasonable response. Further, the program consolidation acts to reduce resource waste and streamline development efforts. Since the federal government has already budgeted nearly a billion dollars for AI research in FY2020, this measure should be budget-neutral (CSTE, 2019, p. 7). Ultimately, the argument for this solution boils down to a humanitarian one. It is simply the only option which offers sufficient control over the risks inherent to pursuing AGI to give humanity the best chance of surviving the transition into a world where the dominant intelligence is silicon-based.

**Evaluation**

The primary objection to the proposed nationalization of American AI research programs will undoubtedly stem from industry actors who have invested substantial resources into their own AI programs. In the absence of free market competition, they will argue, research will stagnate and the United States will fall behind its international competitors. However, AI is too important to be a race to the bottom as everyone tries to cut corners to be the first. In this case, it is better to be right than first. Further, given the global markets in which companies like Google and Amazon operate, there is no guarantee that their goals align with those of the United States. The nationalization of their assets may be seen as preemptive self-defense. Even should American AI research fall behind the global pace, the newly centralized coordination allows general AI development in parallel with narrow AI development. This permits the employment of the more easily developed narrow AI to harden American targets as an interim measure until the long term AGI solution can be brought online.

Despite being granted control over all domestic AI development, the United States Department of Defense (DoD) may also find this solution unpalatable. In addition to suddenly assuming responsibility for perhaps the most important research project in human history, the military, understanding the strategic value of being the first nation to field AGI may object to any process that slows down development. If the Joint Chiefs of Staff believe free market competition is the most efficient way to create an AGI, they are likely to argue in favor of a more laissez-faire approach. Due to her present conventional and nuclear military superiority, America is in a uniquely fortunate position to trade AI delivery speed for greater quality control.

Furthermore, since ingrained design defects are likely to be self-replicating, a well-constructed AGI should be able to iterate more rapidly over itself than a more shoddily, if rapidly, constructed version, eventually outpacing less robust constructs. If necessary, the DoD can leverage the aforementioned parallel development strategy for defensive narrow AI systems to buy time. In the long run, nationalization is the strategy with the highest probability of successfully integrating AI and human goals. With the future of humanity at state, the United States should settle for nothing less.

**Works Cited**

Committee on Science and Technology Enterprise (CSTE). (2019). Supplement to the President’s FY2020 Budget. Retrieved from <https://www.whitehouse.gov/wp-content/uploads/2019/09/FY2020-NITRD-AI-RD-Budget-September-2019.pdf>.

Michigan State Supreme Court. (1919) *Dodge vs. Ford Motor Co.* Retrieved from https://h2o.law.harvard.edu/cases/3965.

National Science & Technology Council (NSTC). (2019). The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update. Retrieved from <https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf>.

Ó hÉigeartaigh, Seán. (2015). The Future of Artificial Intelligence. Retrieved from <https://futureoflife.org/2015/01/31/the-future-of-artificial-intelligence/>.

Tegmark, Max. (2015). Hawking Reddit AMA on AI. Retrieved from <https://futureoflife.org/2015/10/11/hawking-reddit-ama-on-ai/>.

Trump, Donald. (2019). Executive Order 13859. Retrieved from <https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence>.