American Policy for Human-AI Collaboration:

In Which Our Heroes Discover We’re All Fucked

Battista, Jude

**Executive Summary**

Paper summarized. Executively.

**Problem Description**

The development of Artificial General Intelligence (AGI) 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, 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 (NIST) 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**

NIST predicts three ways in which AI and humans are likely to interact. 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 (NIST, 2019).

Explain each tenet. Boring, but necessary.

Algorithms –

Augmentation –

Visualization –

NLP -

**Problem Identification**

In its current incarnation, the NIST policy more closely resembles wishful thinking than a plan constructed to bring about a desired outcome. It is both short-sighted and narrow in focus and reads 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 (NIST). As the subprime mortgage crisis demonstrates, the profit motive that drives a capitalist economy can easily override actors’ better judgement, 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 better judgement. Instead, a single slipshod AI implementation has the potential 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 NIST 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). Then the creation of true AGI , or even a sufficiently advanced narrow AI along the correct axis will 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. 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.

Human-AI compatibility – not explicitly addressed

* 1. AI goals – should they align with human goals or human welfare?
     1. Goals - How do we codify human goals?
        1. How do we understand human goals?
           1. How do we ascertain human goals?
     2. Welfare - Who defines welfare for a given population.
        1. Which populations should be considered?
           1. Global - Global welfare will certainly conflict with American welfare. What happens when the AI recommends that Americans be subject to a substantial QoL drop?
           2. National – This is likely to spawn competing nationalistic AI development since the world can clearly not trust America to be an impartial leader towards a better future.

Is AI war an inevitable outcome here?

Goals and constraints – not explicitly addressed. Needs to be!

1. Constraints implicit in goals
   1. Define goals
   2. Let the AI learn what constraints are necessary to accomplish those goals
   3. Requires very careful goal construction
   4. Learning process can be… expensive.
2. Explicitly defined constraints
   1. Inevitable conflict with goals – do we really want an AI trying to reason out such a conflict?
   2. Impossible to enumerate all required constraints
   3. Malicious compliance?

**Actor Identification**

As an AI researcher I (appropriate first person?), I am concerned by:

Measurable metrics – how do we track progress and compliance?

Conflict of profit motive with humanity

AGI inevitably decreases human agency. Should this power devolve to corporate hands or be held by actors at least nominally responsible to the public.

Executive authority –

Coordination of effort

Motivation of competition vs. the efficiency of cooperation

Nationalization of goals

Centralized command and control

**Potential Solutions**

Government Oversite – Regulation. Watchdog agency.

Government Contracting – Government as sole employer of AI systems. Industry as contractors

Nationalization – Ultimate centralized command and control

**Recommendation**

Nationalization. AGI inevitably decreases human agency. Should this power devolve to corporate hands or be held by actors at least nominally responsible to the public?

**Evaluation**

Massive pushbacks from industry.

Slower progress

**Policy Breakdown**

**Five principles for AI**

1. Promote Sustained AI R&D investment
2. Unleash Federal AI resources
3. Remove barriers to AI innovation
4. Empower the American worker with AI-focused education and training opportunities
5. Promote an international environment that is supportive of American AI innovation and its responsible uses
6. (bonus) Leverage AI to help Federal government work smarter

Potential problems:

1. Nothing about ensuring responsible development
2. Nothing about ensuring responsible employment.
3. Descriptive vs. prescriptive: Lots of buzzwords, no plans

**Strategy 2: Human-AI collaboration**

“While completely autonomous AI systems will be important in some application domains (e.g., underwater or deep space exploration), many other application areas (e.g., disaster recovery and medical diagnostics) are most effectively addressed by a combination of humans and AI systems working together to achieve application goals”.

This vision is both short-sighted and narrow-minded.

Presumes no transformational changes in AI capability.

1. We already see AI performing medical diagnoses.
   1. Skin Cancer
   2. Eye conditions – Deep Mind at Moorfield’s Eye Hospital and University College London. 94%
   3. Image-based diagnosis – 14 studies combined in Lancet (Denniston, Liu, et al.)
      1. Better disease state detection
      2. Better all-clear accuracy

“Complementary nature of humans and AI systems”

What is the nature of an AI system? Since AGI does not yet exist, how do we characterize it?

Better to *define* the nature of the AI systems we want to build

Enforcement? Hahah.

**AI-Human divisions:**

1. AI performs functions alongside the human – Partnership
   1. Why is the human necessary here? Safety? Comfort?
   2. Or is this just a transitional state as AI grows?
2. AI performs functions when the human encounters high cognitive overload – Supplemental
   1. Which system has precedence in a conflict?
3. AI performs functions in lieu of a human – Replacement
   1. Phrased as handling tasks for which human capacity is limited
   2. Isn’t that ultimately every task as AI capability scales more rapidly than humanity?

Human-centered automation principles:

1. Employ intuitive, user-friendly design of human-AI system interfaces, controls, and displays.
2. Keep the operator informed of critical information, AI system state, changes to state
3. Keep the operator trained.
   1. Recurrent training for general knowledge, skills, abilities (KSA).
   2. Training in algorithms and logic employed by AI system
   3. Expected failure modes of the system
4. Make automation flexible. Operators decide whether to use the AI system.
   1. Adaptive AI systems to support human workers during excessive workload or fatigue
      1. This is essentially replacement with some gift wrap. There is no way this technology is not used to replace the worker.

**Algorithms for human-aware AI.**

Interact intuitively with users.

1. Shallow interactions – User discards option recommended by the system
2. Model-based approaches – use the prior actions of the user
3. Deep models of user intent – based on accurate human cognitive models
   1. Yes, the plan is for the machine to know what you’re thinking. This definitely won’t be abused by every government agency ever.

Interruption models – know when to interrupt the human. This implies super-human understanding of the task. How does this not become replacement?

Develop emotional intelligence – Good goal. Good luck.

Generalization – system of systems

Many AIs interacting with many humans

**Human Augmentation**- Transition from narrow devices to multifunction devices

1. Types
   1. Stationary
   2. Wearable
   3. Implantable
   4. Environmental – this seems redundant with stationary
2. They’re basically chasing Scalzi’s BrainPal
   1. Privacy concerns.
   2. What about explicit and implicit control
      1. Information flow
      2. Physical control
3. Active learning – relies on human SMEs when the AI is uncertain during training
   1. Who chooses the SMEs? This definitely won’t devolve into a huge Charlie Foxtrot.

**Visualization and human-AI interfaces**

1. Who is responsible for designing these? Many current interfaces are designed by programmers and it shows. Which disciplines should be responsible for the interface design, and who should have precedence when back-end and front-end functionalities collide?
   1. If we are to have truly human-friendly AI the interface should drive functionality, not vice versa. This will be a difficult prospect to sell.

**Language processing systems**

1. Currently capable of understanding fluent English in quiet surroundings
   1. How to handle biases inherent to languages? Xenophobia?
   2. What about the multilanguage question? Should English be the official national language for AI interaction? How does this affect the spread of AI? Reinforce classism?
2. Working towards establishing emotional and environmental context for speech
   1. Yeah, that’s probably a good idea.

**Policy Updates**

First and foremost, we need to clarify an overarching set of goals for how our AI should interact with humans. <These need to be specific. Measurable?><Analogous to Asimov’s Laws?> These goals need to be generated by a multidisciplinary task group consisting of philosophers and ethicists as well as the traditional computer scientists and engineers. <How do we reach concensus>. 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.

<Should we suggest goals? Probably beyond the scope of this project. Instead consider the objective of the goals?>

Possible objectives:

1. Preservation of human life
2. Preservation of human life with dignity
3. Preservation of human agency – meaningful decisions within own lives
   1. Priority of this vs. welfare?
4. Quality of life?
   1. How do we measure this?

Is priority absolute or do extreme values override low values for higher priority objectives?

**Possible Objections**

Industry – these requirements will substantially slow down development, cutting into profits and potentially putting them behind the international competition

Rejoinder: 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. We may also need to develop AI countermeasures. We cannot necessarily interfere with AI development in other countries, but we can harden our own targets. While chasing AGI, build a parallel suite of narrow AI applications designed to keep us on par in selected critical areas?

Defense – Slower development puts us at a disadvantage in armed conflict.

Rejoinder: Rely on existing military supremacy while our AI catches up. If necessary throw them the sop of increased funding for conventional defensive measures. A well-constructed AI should be able to iterate more rapidly over itself. Additionally, a poorly-constructed AI will pass its flaws on to its children. It is much easier to build correct software (TDD?) than to debug broken software. This exponentiates as software designs software. Inherited defects rapidly become impossible to correct since nobody understands the system.

**Works Cited**

National Science & Technology Council (NIST). (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>.