Battista, Jude

CO-350

Dr. Measor

Policy Paper Outline

**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-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?

**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.