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

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

AI goals – should they align with human goals or human welfare?

How do we codify human goals?

How do understand human goals?

How do we learn human goals?

Who defines welfare?

**Goals and constraints** – not explicitly addressed

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**

**Visualization and human-AI interfaces**

**Language processing systems**

**Policy Evaluation**

The National Artificial Intelligence Research and Development Strategic Plan spends five pages outlining its policy for ensuring that AI development is human-compatible.