Autonomous Systems Safety & Governance:

Thursday Lab: Short Answer Worksheet

Jeffrey Ding

AIMS CDT November 2019

This worksheet aims to help unpack some of the insights from the lecture on AI governance and AI ethics. Responses to each question should be in short-answer format (2-3 paragraphs).

I. AI Governance Landscape

We've discussed how the technical landscape of AI will affect important issues in AI governance. Let's zoom in on the relative importance of compute as an input to training and deploying large-scale AI system (whether compute becomes a key bottleneck in training and deploying large-scale AI systems, the scarcity/availability of compute). How could changes in the relative importance of compute affect the AI governance problem?

1. *Explore this with respect to the expected speed and scale of AI progress.*

2. *Explore this with respect to the key actors in AI governance.*

3. *Explore this with respect to potential mechanisms of AI governance.*

AI risks can be divided into *misuse risks*, *accident risks,* and *structural risks.*
See: https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure. It's useful to conceptualize how each of these risks could characterize one specific AI application. For example, for AI-enabled cyberoperations, an automated cyberdefense system could malfunction (*accident*), a hacker could use precise, automated spearphishing to compromise a key entity (*misuse*), or AI advances could strengthen offensive cyber capabilities more than defensive ones, exacerbating the cybersecurity dilemma (*structural*).

4. *Give an example of how each one of these risks could apply to a different application of AI, outside of enabling cyberoperations.*

5. *The misuse-accident-structural framework is one way categorize AI risks. Try to construct an alternative way to think about AI risks:*

II. AI Governance in Comparison with other Dual-Use Domains

For this section, the following supplementary text may be useful.
Harris, E. D. (ed.) (2016), Governance of Dual-Use Technologies: Theory and Practice, esp. the Concluding Observations by Elisa D. Harris,
https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/7

6. *What is the governance potential of AI as compared to biotechnology? How would you need to define the scope of "AI" to make this comparison possible?*

7. *What lessons can we learn from progress of government regulation and international cooperation in the cyber and biotechnology domains that can be applicable to the governance of AI?*

III. Existing AI Ethics Principles and Governance Mechanisms

Below is a collection of different AI principles issued by different companies, government bodies, and civil society groups.

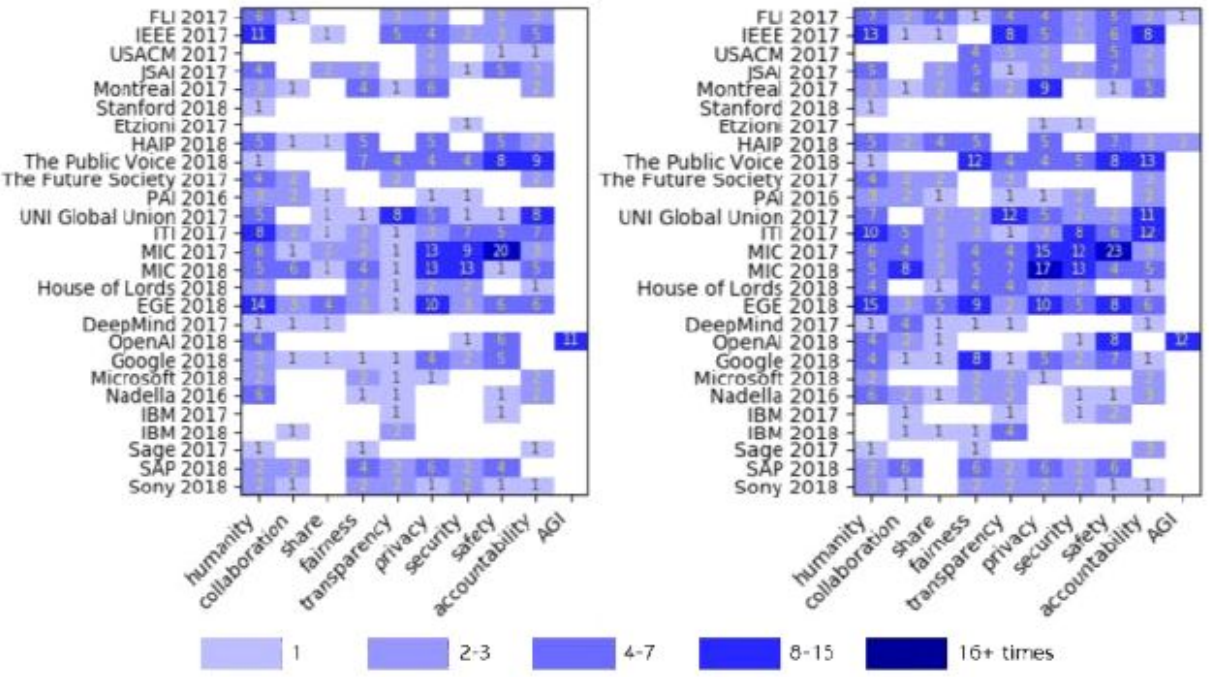

Figure 1. Topic coverage of principle titles and explanatory texts based on manually chosen keywords (A) and extended keyword groups by semantic similarity (B).

Source: https://arxiv.org/ftp/arxiv/papers/1812/1812.04814.pdf

8. *What, if anything, surprises you about the topic coverage analysis of these AI principles? If nothing is surprising, explain why.*

9. *These ten topics were chosen manually by the authors of the paper* (see Table 1 below)*. Are these principles mostly cheap talk? If you could just choose one principle that all firms and governments should adhere to when developing AI systems, what would it be and why?*

Table 1.Topics and Manually Chosen Keywords for AI Principles

Topics	Keywords
Humanity	humanity, beneficial, well-being, human value, human right, dignity, freedom, education, common good, human-centered, human-friendly
Collaboration	collaboration, partnership, cooperation, dialogue
Share	share, equal, equity, inequity, inequality
Fairness	fairness, justice, bias, discrimination, prejudice
Transparency	transparency, explainable, predictable, intelligible, audit, trace, opaque
Privacy	privacy, personal information, data protection, informed, explicit confirmation, control the data, notice and consent
Security	security, cybersecurity, cyberattack, hacks, confidential
Safety	safety, validation, verification, test, controllability, under control, control the risks, human control
Accountability	accountability, responsibility
AGI/ASI	AGI, superintelligence, super intelligence

Below is the U.S.Department of Defense's ethics principles for its use of AI systems.

1. **Responsible.** Human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of DoD AI systems.

2. **Equitable.** DoD should take deliberate steps to avoid unintended bias in the development and deployment of combat or non-combat AI systems that would inadvertently cause harm to persons.

3. **Traceable.** DoD's AI engineering discipline should be sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including transparent and auditable methodologies, data sources, and design procedure and documentation.

4. **Reliable.** DoD AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use.

5. **Governable.** DoD AI systems should be designed and engineered to fulfill their intended function while possessing the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behavior.

Source: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

10. *Choose two of these principles, and identify possible AI applications where the DoD may face a tradeoff between adhering to the principle and fulfilling its mission of upholding national security.*