AI Models for Sociotechnical Virtue

Mark Graves (mgraves@nd.edu)

Preliminary Material for Presentation to Center for Digital Culture, Pontifical Council for Culture on March 23, 2022

In my presentation, I will describe how one could use computational AI models, such as from machine learning (ML), to construct AI systems to interpret human morality and show how these models could also serve as a source for AI morality. To situate these interpretive models within broader topics of study, I have written some preliminary remarks and describe three aspects of my work, which I believe roughly correspond to the interests of the three working groups. I also suggest one article for each thread. Please let me know if you have trouble accessing any of the articles or would like additional material.

Three aspects of my research are:

- Construction of Moral AI – *How can AI acquire moral skills similar to what humans have or aspire?*
- Data Science Ethics – *How can ethics be incorporated into the development of complex AI and ML sociotechnical systems?*
- Human-AI Interpretive Relationship – *How can people and AI mutually interpret each other?*

Preliminary Remarks

As preliminary framing of my research, rather than consider AI and people as distinct things that interact, I tend to think in terms of sociotechnical systems that include people, AI, and other technologies and institutions. I do so for three reasons.

First, more philosophically, as AI and related advanced digital technologies become progressively more incorporated into social structures and relationships, it becomes hard to consider a person as entirely isolated from technology, e.g., close relationships often use technology for communication; one's "extended mind" includes calendars, notes, and tools; longevity and health often depend upon medical technologies.

Second, more technically, even as someone who works closely with complex AI technologies, it can be hard to tell where the boundary is between what is done computationally and what is done manually. IBM Watson was (in)famously dependent upon lots of human knowledge workers, yet DeepMind's AlphaZero and subsequent self-supervised learning tasks automate highly technical machine learning processes that only a few thousand people at most would have the skills to do manually. Even among fairly mundane aspects of AI software development, there are many tasks that might be manually done at one company and automated at another.

Third, sociotechnical systems capture the mutual causality of people defining technology that significantly affects people's lives. This focuses ethical examination of those systems on the interdependent relationships between people and AI as formative to both.

Considering people and AI as components of a sociotechnical system impacts how I approach different aspects of my research and connect them together. For constructing moral AI, I extend the sociotechnical system spiritually toward community using the pragmatic philosopher of religion Josiah Royce's foundation of spirituality in a "community of interpretation." Because of the importance of communication to interpretive communities, I focus my Human-AI relational research on natural language processing (NLP). For data science ethics, I consider the microethical skills (i.e., virtues) and decisions needed to construct and do analysis with machine learning and statistical models, and for constructing moral AI, I consider how more expressive models could capture morality.

Construction of Moral AI

In setting a foundation for long-range, speculative AI development, this project draws upon theological anthropology and moral theology to define models for morality inspired by human moral psychology but that could be implemented using current or near-future AI methods. I adapt for AI a scientifically plausible theological anthropology that uses systems theory to organize AI models of its physical,

biological, psychological, sociotechnical, and moral worlds as well as its reckoning of itself in those worlds. Distinguishing levels of models clarifies the nexus of AI decision making/proto-self as sociotechnically embedded (like most social scientists would argue for humans in society) rather than reductively built from physical or idealistic foundations. In the following article, I close the gap between overly reductionist approaches to AI and implicit or explicit assumptions of human dualism by arguing that an emergent monism can better characterize both human and AI morality.

M Graves. Emergent Models for Moral AI Spirituality. *International Journal of Interactive Multimedia and Artificial Intelligence – IJIMAI.* 7:1. Special Issue on AI, Spirituality, and Analogue Thinking. 2021. (open access)

Some of my presentation also describes work to appear in the forthcoming article:

M Graves. Theological Foundations for Moral Artificial Intelligence. *Journal of Moral Theology*. 11(Special Issue 1). Special Issue on Artificial Intelligence and Machine Learning. 2022. (in press)

Data Science Ethics

Software developers used to design and build monolithic software systems, but current practice for cloud software development is to develop highly-decentralized software components with continuous deployment (into the overall product). As a personal consumer, one typically updates software like a word processor or web browser by substituting a newer incremental version for a previous one, but large AI software—like Google Search, Facebook rankings, or Netflix recommender—consists of hundreds or thousands of interacting components, each of which can be replaced independently or even partially replaced for just some users. At the core of the ML and other analytical components are models trained on disparate data sources, which are built as part of a data science process.[1] Rather than one source of data being processed centrally, the numerous components process thousands of data streams individually, and thus practitioners have limited visibility into the consequences of many of their decisions. Although professional ethics can guide overall behavior, it is difficult for AI developers to parse broad ethical principles, like beneficence, when the effects of their immediate work is deeply embedded in complex sociotechnical systems. In the following paper, my collaborator and I argue for a microethical approach where the development of ethical skills is tightly coupled with development of technical skills.

E Ratti and M Graves. Cultivating Moral Attention: A Virtue-oriented Approach to Responsible Data Science in Healthcare. *Philosophy & Technology*. 34(4): 1819-1846. 2021. (open access)

Human-AI Interpretive Relationship

As AI systems and data practices become more sophisticated and as human-AI interactions become more complex and with more opaque consequences, then communication and trust become more essential. At the heart of that interaction, and through the lens of Royce's semiotic philosophy, is interpretation of AI by humans, interpretation of humans by AI, and each interpreting one to others. Although the interpretive relationship is highly limited currently by AI's restricted abilities, there are ways to investigate how AI systems can interpret human communication through natural language processing (NLP) and how people can interpret models created by AI. Although either could be pursued broadly, at their intersection and relevant for moral investigation is using NLP text analysis tools to build computational models of human moral psychology, or in the case of the following article, moral theology from the perspective of Thomas Aquinas. Although some NLP models can generate human-readable text (e.g., GPT-3), in this article I describe close examination of the model itself in order to interpret Thomistic moral theology and its interpretation of papal encyclicals.

M Graves. Computational Topic Modeling for Theological Investigations. *Theology and Science.* 20(1): 64-84. 2022. (online)

[1] Data science is defined in many ways, but here it refers to the data processes around development of machine learning models for analysis, prediction, or experimentation.