Towards Artificial *Phronesis*: Some potential first steps along the road to moral agency using case studies

John Murray

San José State University, San José, California, USA
jxm@acm.org

Abstract. We will survey recent AI regulatory and policy activities from several perspectives, to understand their potential for helping the development of artificial moral agency, or *Phronesis*.

Keywords: Ethics, AI policy, moral agency, GDPR.

1 Summary

What sort of ethics framework would be needed as a precursor to pursuing the goal of machine consciousness? Are there insights to be gleaned from current policy developments in the field of ethics in the context of current and near-future AI systems? This talk will survey recent regulatory activities in this arena from several perspectives and examine how these might – or might not – help us to explore the development of artificial moral agency as a potential contribution to the mechanistic underpinnings of future machine consciousness.

There are several salient elements of relevance here. One is the need to ensure that ethical considerations are incorporated in the initial specification, engineering design, and development of AI systems. Some of the key concerns here are the need to recognize inherent biases and problematic strategies, such as those related to the collection and/or selection of training data for machine learning systems, for example Microsoft's ill-considered Tay system.[1]

A second element is the application of ethical principles during the prototyping/testing of AI systems. There is a significant concern that, as autonomous systems get more powerful and adept at human tasks, the corresponding ethical considerations scale as exponentially large as the applications. The problem is that, when major AI development organizations put ethics on the backburner, major risks can spiral up and out of control. Of particular interest here is the purposeful incorporation of human participants – who themselves may or may not be aware of their involvement – in the testing and validation of AI systems.[2]

Separately from these elements, which are applicable to most or all human interaction and decision support systems, we also need to concern ourselves with the challenges involved with designing a concept of ethics awareness into the AI products themselves. The vision of building artificial moral agents, which operationalize moral practical wisdom or Aristotelian *Phronesis,* falls under this category.[3]

One way forward is to develop a corpus of case studies, which can be used to guide system designers and researchers to envision techniques for architecting such agents. For a very simplistic example, we can look at the European Union's introduction of the General Data Protection Regulation (GDPR), which directs how an individual's personal data should be handled and processed.[4] The GDPR deployment has triggered calls for regulators to publish case studies that illustrate how its principles are put into practice, such as when implementing data management systems, investigating complaints, etc.[5] This talk will incorporate additional examples of this case studies approach. It is anticipated that analytical tools from research programs like SIMPLEX [6], such as DASL[7], may prove valuable in advancing this work.

The initial benefit of such a body of case law will enable individuals to challenge the data practices of organizations, and in turn allow organizations to take data protection authorities to task over their enforcement actions. However, in the long term, this corpus also provides a valuable resource for characterizing tradeoffs between individual rights vs. common goods, implicit vs. explicit consent, and other issues in the areas like privacy, anonymity, security, etc. In other words, a body of structured examples of how ethical dilemmas and tradeoffs are actually resolved in practice. In this way, such a corpus could eventually become a key resource for developers of artificial phronesis.

References

1. Neff, G. & Nagy, P: Talking to Bots: Symbiotic Agency and the Case of Tay. International Journal of Communication 10(2016), 4915–4931.
2. Lomas, N.: Duplex Shows Google Failing at Ethical and Creative AI Design: https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design, last accessed 2018/11/01.
3. Sullins, J.: Machine Morality Operationalized. Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014.
4. European Union: General Data Protection Regulation. (EU) 2016/679 of the European Parliament and of the Council.
5. Dixon, H.: Regulate to Liberate: Can Europe Save the Internet? Foreign Affairs 97(5), 28—32 (2018).
6. DARPA: Simplifying Complexity in Scientific Discovery (SIMPLEX): www.darpa.mil/program/simplifying-complexity-in-scientific-discovery, last accessed 2018/12/16.
7. SRI International: Deep Adaptive Semantic Logic (DASL): www.sri.com/work/projects/deep-adaptive-semantic-logic-dasl, last accessed 2018/12/16.