The Role of Logic based Argument and Dialogue in Addressing the AI Value Loading Problem (Extended Abstract)

Recent successes in artificial intelligence (AI) are largely attributable to advances in machine learning, and have been accompanied by leading researchers warning of the possible dangers of AI [16]. It is argued that machines may formally achieve their operators' goals in ways that may conflict with their operators' values [1, 16, 20]. This concern recalls the argument that adhering to any rule based ethical system may result in unintended, harmful consequences [13, 14], and has acquired renewed urgency given that *a feature* of learning systems is the discovery of unforseen ways of achieving goals, and that achievement of *any* operator's goal will incentivise machines to thwart corrective measures to prevent harm [1, 18, 19, 20]. To ensure AI actions are aligned with human values has in a machine learning context been termed the 'value loading (alignment) problem' [1, 20]. Whether ethical behaviour is implemented using machine learning techniques, via the maximising of utility functions encoding human preferences, and/or through use of symbolic logic based reasoning adhering to explicitly encoded ethical theories [10, 22], two key research problems need to be addressed [1, 15, 20]. Firstly, the problem of specifying objective utility functions (deontic axiomatisations) that are perfectly aligned with human values and applicable in changing environments and to novel situations (in particular ethical dilemmas with no precedent). Secondly, the above described problem of unintended behaviours misaligned with human values. Run time learning of values has been proposed to address these problems [18], for example through use of inverse reinforcement learning [11] in which AI systems are incentivised to observe and query humans [15]; the assumption being that actions reveal preferences and hence values, and that humans are sufficiently informed and have the requisite capacities to definitively arbitrate on matters of ethical importance. However, humans clearly do not always behave ethically, and moreover are often uncertain about how to resolve ethical issues; particularly those arising from deployment of novel technologies (that hence lack precedent).

I argue that we therefore require that AI systems and humans engage in comprehensive, rational exchange of arguments purposed to decide ethical issues; indeed, in dialogues that will *be better purposed to do so* by virtue of incentivising and harnessing AI's vastly superior access to information and capacity to look further into the future. To meet requirements for such 'value deliberation' dialogues (cf. 'value learning') will require building on current research in logic-based models of argument and dialogue. In particular works on dialectical characterisations of non-monotonic logics [9] that instantiate Dung's general theory of argumentation [3], and that are: 1) extended to incorporate reasoning *about* values and preferences [7, 8]; 2) reformulated as formal dialogical frameworks for joint reasoning that accommodate humans and machines [4, 12]; 3) provably rational under bounds on computational/cognitive resources and when employing real-world modes of dialectical reasoning [2]. A non-exhaustive list of further research challenges includes argumentative characterisations of action, epistemic and deontic

logics, argument mining [5], interdisciplinary collaborations with informal logic and philosophy (e.g., further integration of logic-based argumentation with schemes and critical questions [23], speech act theory [17] and pragma-dialectics [21]), and integration with machine learning to address the symbol grounding problem [6].

References

[1] M. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[2] M. D'Agostino and S. Modgil. A rational account of classical logic argumentation for real-world agents. In *European Conference on Artificial Intelligence*, pages 141 – 149, 2016.

[3] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[4] X. Fan and F. Toni. A general framework for sound assumption-based argumentation dialogues. *Artificial Intelligence*, 216:20 – 54, 2014.

[5] G.Ami. Computers that can argue will be satnav for the moral maze. *New Scientist*, September, 2016.

[6] M. Garnelo, K. Arulkumaran, and M. Shanahan. Towards deep symbolic reinforcement learning. *CoRR*, abs/1609.05518, 2016.

[7] S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9-10):901–934, 2009.

[8] S. Modgil and T.J.M Bench-Capon. Metalevel argumentation. *Journal of Logic and Computation*, 21(6):959–1003, 2011.

[9] S. Modgil and H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.

[10] J. H Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.

[11] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 663–670, 2000.

[12] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15:1009–1040, 2005.

[13] R.Clarke. Asimov's laws of robotics: Implications for information technology - part 1. *Computer*, 26(12):53–61, 1993.

[14] R.Clarke. Asimov's laws of robotics: Implications for information technology - part 2. *Computer*, 27(1):57 – 66, 1994.

[15] S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *CoRR*, abs/1602.03506, 2016.

[16] S. Russell, T. Dietterich, E. Horvitz, B. Selman, F. Rossi, D. Hassabis, S. Legg, M. Suleyman, D. George, and S. Phoenix. Research priorities for robust and beneficial artificial intelligence: An open letter. *AI Magazine*, 36(4):3–4, 2016.

[17] J. R. Searle. Speech acts. In *Cambridge University Press, Cambridge UK.*, 1962.

[18] N. Soares. The value learning problem. In *Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence (IJCAI-2016)*, 2016.

[19] N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky. Corrigibility. In *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop*, 2015.

[20] J. Taylor, E.Yudkowsky, P. LaVictoire, and A. Critch. Alignment for advanced machine learning systems. `https://intelligence.org/2016/07/27/alignment-machine-learning/`, 2016.

[21] F. H. van Eemeren, B. Garssen, E. C.W. Krabbe, A. Henkemans, S. Francisca, , B. Verhei, and J. H. M. Wagemans. *The Pragma-Dialectical Theory of Argumentation*, pages 517–613. Springer, Dordrecht, 2014.

[22] W. Wallach, C. Allen, and I. Smit. Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI & Society - Special Issue: Ethics and artificial agents*, 22(4):565–582, 2008.

[23] D. N. Walton. *Argument Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, 1996.