Many Kinds of Minds are Better than One: Value Alignment Through Dialogue

Sanjay Modgil
Department of Informatics, King's College London

Recent successes in artificial intelligence (AI) have in large part been due to advances in machine learning, and have been accompanied by leading researchers warning of the possible dangers of AI [1,17]. It is argued that the forseeable benefits of AI will license the development of, and trust in, machines that are increasingly more powerful (with cognitive powers far outstripping those of humans), autonomous and capable of acting in diverse and open environments. However, such machines may formally achieve their operator's goals in ways that not only diverge from their operators intentions, but may actually be contrary to the interests and values of their operators [1,17,21]. This concern recalls arguments to the effect that adhering to any rule based ethical system may result in unintended, harmful consequences (as exemplified by Asimov's laws of robotics [14,15]). However, this problem has acquired renewed urgency given that it is *a feature* of learning systems that they find unforseen ways of achieving goals, and that achievement of *any* operator's goal will incentivise 'instrumental goals' (such as self-preservation) that thwart corrective measures to prevent harm [1,20,21,19]. The need to ensure AI acts in accordance with human values has prompted considerable intellectual investment into what in a machine learning context has been termed the 'value loading (alignment) problem' [1,21], more broadly understood as the problem of how to design 'ethical agents [1]. Whether the envisaged agents' ethical behaviour is implemented through use of machine learning techniques via the maximising of utility functions encoding human preferences, and/or through the use of 'top down' symbolic logic based reasoning adhering to explicitly encoded ethical theories [12,23], two key research problems need to be addressed [1,16,21]. Firstly, there is the problem of how to specify objective utility functions (deontic axiomatisations) that are perfectly aligned with human values and applicable in changing environments and to novel situations (in particular ethical challenges that lack precedent and thus most saliently expose the is/ought gap, such as those arising from the use of radically new technologies). Secondly, there is the above described problem of unintended behaviours misaligned with human values. Run time learning of values has been proposed to address these problems [19], for example through the use of inverse reinforcement learning [13] in which AI systems are incentivised to observe and query humans [16]; the assumption being that actions reveal preferences and hence values, and that humans are sufficiently informed and have the requisite capacity 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; in particular those arising from the use of

[1]Note the assumption that however advanced the AI (including the 'super-intelligent' machines whose developmental trajectory Bostrom rigorously charts once human level artificial *general* intelligence is achieved), deciding on moral issues in isolation, cannot, even in principle, always align with the moral decision making of humans. Since whether one is a Aristotliean type virtue ethicist, a Kantian deontologist, or a consequentialist, one must neccessarily access (reports of) first person subjective experience in deciding what oughts pertain when faced with ethically challenging issues.This is most explicitly acknowledged by the utilitarian account of consequentialism, which advocates ethical choices that impartially maximise total happiness, where happiness is broadly construed as the subjective experience of well being.

novel technologies, as highlighted above). I argue that we therefore require that AI systems and humans engage in comprehensive, rational exchange of arguments purposed to decide ethical issues (as humans do when faced with difficult ethical decisions). Indeed, I contend that such *'value deliberation' dialogues will be better purposed to decide ethical issues*, by integrating the vastly superior epistemic and causal reasoning capacities of AI systems, with human deliberations over values and preferences that are in turn informed by the qualitative aspects of human *experience* (Recall Footnote 1).

To illustrate, consider a recent variation on the trolley problem thought experiments [4] in which an autonomous vehicle (*AV*) is on course to hit and kill five pedestrians. The *AV* can swerve, but then will hit an obstacle and kill the driver. Should the *AV*'s algorithms be designed to continue on its course or swerve ? We can speculate a dialogue purposed to decide this issue, integrating arguments from a utilitarian AI agent and human agents. The former submits an argument claiming that the car should swerve, given the standard utilitarian calculation that the number of lives lost are minimised by so doing. As reported in [7], while human agents typically support such an argument, a significant majority would not themselves buy an *AV* that implements such a calculation. On eliciting reasons as to why, humans might be expected to appeal to the value of self-preservation as trumping the utilitarian principle, where ceteris paribus, such an appeal is arguably contingent on the *experienced* desire for self-preservation [2]. The AI agent may then seek to resolve the apparent dilemma by firstly estimating the likelihood of such scenarios, which is likely to be extremely low (reflect for a moment on how often you think such scenarios have actually occurred in all the billions of car journeys since the invention of motorised vehicles). Then given the 'argument from self preservation', the AI agent estimates the reduced number of sales of *AV*s that implement the above utilitarian calculation, and subsequently argues that the increased fatalities resulting from reduced sales far outnumber the tiny number of fatalities that would result if the above utilitarian calculation were not implemented. Indeed, this line of argument is itself a utilitarian argument, and shows that the above calculation, properly considered, does in fact *violate* the utilitarian principle.

To meet requirements for realising such 'value deliberation' dialogues (cf. 'value learning') will initially require building on current research in logic-based models of argument and dialogue. In particular works on dialectical characterisations of non-monotonic logics that instantiate Dung's general theory of argumentation [2,11], and that are: 1) extended to incorporate reasoning *about* values and preferences [8,10]; 2) reformulated as dialogical exchanges so as to provide formal frameworks for joint reasoning that accommodate humans and machines [9]; 3) provably rational under bounds on computational/cognitive resources and when employing real world modes of dialectical reasoning [3]. In particular, one will require extending the aforementioned work to accommodate arguments based on probabilistic, causal and hypothetical reasoning, as well as more sophisticated modes of dialectical reasoning other than the use of attacks (e.g., the weighing up - *qua accrual* – of arguments) and the dialectical demonstration of inconsistency modelled in [3]. A non-exhaustive list of other research challenges includes argumentative characterisations of action, dynamic epistemic and deontic logics, argument mining [5], and interdisciplinary collaborations with researchers in informal logic and philosophy (e.g., further integration of logic based argumentation with schemes and critical questions [24], speech act theory [18] and the pragma-dialectic school of argumentation [22]) and integration with machine learning to address the symbol grounding problem [6].

[2]The trolley problem thought experiments explicitly indicate that the only consequential distinction that can be made is that more lives are lost in one scenario than the other; we are explicitly told that, for example, nothing is known about whether the future life of the one has more or less beneficial consequences than the future lives of the ten.

References

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

[2] 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.

[3] M. DAgostino and S. Modgil. Classical logic, argument and dialectic. *Artificial Intelligence*, 262:15–51, 2018.

[4] P. Foot. The problem of abortion and the doctrine of the double effect in virtues and vices. *Oxford Review*, (5):5–15, 1967.

[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] I. Rahwan J. F. Bonnefon, A. Shariff. The social dilemma of autonomous vehicles. *Science*, (6293):1573–1576, 2016.

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

[9] S. Modgil. Towards a general framework for dialogues that accommodate reasoning about preferences. In *Fourth International Workshop on Theory and Applications of Formal Argument (TAFA 2017 co-located with IJCAI17*, pages 175–191, 2017.

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

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

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

[13] 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.

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

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

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

[17] 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.

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

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

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

[21] 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.

[22] 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.

[23] 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.

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