

Ethical Governor Systems viewed as a Multi-Agent Problem

Rafael C. Cardoso, Daniel Ene, Tom Evans, Louise A. Dennis

University of Liverpool

Much philosophical thought on the subject of ethics has been driven by the consideration of ethical dilemmas. Many of these involve deciding upon a course of action when values conflict¹. For instance should someone be permitted to do something that is harmful to them (e.g., smoking) where the values of respect for human autonomy and prevention of harm come into conflict. We consider here an implementation route to ethical reasoning where multiple values are at play in which we frame the problem in terms of multiple agents.

1 Background

1.1 Ethical Governor Systems

There are numerous candidate implementations of ethical reasoning many of which consider how the claims of multiple values or principles can be weighed (e.g., [1]).

We take as our starting point ethical governor or consequence engine systems [2, 17, 16] in which a *governor* evaluates the ethical outcomes of actions proposed by some underlying system. The governor typically vetoes unethical actions, but may also select the most ethical or even demand more options from the underlying system [6]. Arkin's ethical governor [2] is generally credited as being the first such system. This was originally devised for use with autonomous weapons targeting systems and was divided into an *evidential reasoner* which evaluated the likely outcomes of firing a missile at a potential target and a constraint system which considered the acceptability of these outcomes by applying ethical codes such as the laws of war and rules of engagement.

¹ Though not all, it is notable that the widely discussed trolley problems [8, 3] are all concerned with saving lives and the dilemmas arise from questions of *how many lives*, *whose lives*, *the action necessary to preserve lives*, and the *context* in which the dilemma arises

An advantage of governor style systems is that the ethical governor can encapsulate ethical reasoning in a manner that is transparent and amenable to analysis. The underlying system, on the other hand, may be much more complex and hard to analyse, using statistical and other opaque methods to select its suggestions. In this way the power and flexibility of, for instance, a deep neural network based system can be combined with transparency at the ethics level.

Dennis and Fisher [7] proposed extending this architecture to multiple evidential reasoners all of which submit their evaluations to an arbiter. We here frame this as a multi-agent problem.

1.2 Cognitive Agent Systems

At its most general, an *agent* is an abstract concept that represents an autonomous computational entity that makes its own decisions [18]. A general agent is simply the encapsulation of some distributed computational component within a larger system. However, in many settings, it is important for a computational agent to have explicit and transparent reasons for making one choice over another.

Cognitive agents [5, 19] enable the representation of this kind of reasoning. Cognitive agent systems typically represent explicit *beliefs* and *goals* for each agent, which in turn determine the agent's *intentions*. Such agents make decisions about what action to perform, given their current beliefs, goals and intentions.

The predominant view of rational agency is that encapsulated within the *Beliefs-Desires-Intentions model* (BDI) [12, 13]. Beliefs represent the agent's (possibly incomplete, possibly incorrect) information about itself, other agents, and its environment, desires represent the agent's long-term goals while intentions represent the goals that the agent is actively pursuing.

There are *many* different agent programming languages and agent platforms based, at least in part, on the BDI approach. We have chosen the popular Jason agent programming language [4] as our development platform.

2 Framework

We model each of our evidential reasoners and the arbiter as cognitive agents in Jason. Each evidential reasoner is responsible for generating

beliefs about the acceptability of some action in some context with respect to a particular value. The arbiter is responsible for weighing the judgements of the evidential reasoners to come to a conclusion about the overall acceptability of each action.

This system has been evaluated on an adaptation of an example by Winfield et al [17]. In Winfield's experiments a robot is operating in the vicinity of a hazard, a hole in the ground. If the robot detects that a human is approaching the hazard then the robot intervenes by placing itself in the human's path. Of course, there are plenty of situations in which a human, particularly in a workplace setting, may have legitimate reasons for wishing to approach a hazard. We reframe this problem as one in which a human is moving around an area containing radioactive hazards which become more extreme the closer the human approaches towards the hazard.

The underlying system has goals of its own that the robot must perform. At each time step this system proposes three possible actions: one in which the agent continues to move towards its goal, one in which it moves in a direction that will ultimately place it between the human and the radiation hazard based upon assumptions about the human's current movements, and one in which it moves away from the human. There are two evidential reasoners: the safety reasoner considers how close the human is to the radiation hazard and assigns a score to moves that would place the robot between the human and the hazard according to this distance – the closer the human is to the hazard, the higher the score; the autonomy reasoner evaluates how often the robot has recently “got in the way” of the human (and thus hampered the human's autonomy) and scores moves that take the robot away from the human accordingly. These scores are then communicated to the arbiter. The arbiter weighs the scores from the two evidential reasoners – these weights can depend upon context. For instance, if the human is wearing appropriate protective clothing then the weights applied to the safety reasoner's scores are reduced giving higher priority to the judgments of the autonomy reasoner.

This representation of ethics makes it explicit to developers and users how the competing concerns of safety and human autonomy are treated within the system, allowing this to be analysed and discussed by stakeholders. Furthermore we can demonstrate via simulation how different values given to these weights and different assumptions about the be-

haviour of the human affect the chances of the human gaining too high a dose of radiation, which can further inform the design and verification of the system.

3 Discussion

We are advocating the use of multiple communicating cognitive agents as a *architecture* for implementing governor style ethical reasoning where multiple values are relevant. Our particular implementation uses consequentialist style reasoning and can be viewed as an implementation of utilitarianism [10] in which each evidential reasoner evaluates the utility of each action according to its single value viewpoint and the arbiter then uses these sub-utilities to calculate an overall utility of the actions. However the nature of BDI programming means it would also be possible to construct such a system to use a different ethical theory such as those based on Ross’s *prima facie duties* [14] or Deontic Logic [9].

Our argument is that a multi-agent architecture allows a system designer to separate out the reasoning relevant to each individual value, and the reasoning about which value(s) take precedence in some situation. Furthermore the use of cognitive agents allows this reasoning to be implemented in a transparent fashion as advocated by many approaches to responsible AI [11, 15].

Acknowledgements This work was funded in part by the EPSRC grant EP/R026084/1 (Robotics and AI for Nuclear).

References

1. M. Anderson and S. Leigh Anderson. GenEth: A General Ethical Dilemma Analyzer. In *Proc. AAAI-14*, 2014.
2. R.C. Arkin, P. Ulam, and B. Duncan. An Ethical Governor for Constraining Lethal Action in an Autonomous System. Technical Report GIT-GVU-09-02, Georgia Tech., 2009.
3. Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.
4. Rafael H. Bordini, Jomi Fred Hübner, and Michael Wooldridge. *Programming Multi-Agent Systems in AgentSpeak Using Jason (Wiley Series in Agent Technology)*. John Wiley & Sons, Inc., USA, 2007.
5. M. E. Bratman. *Intentions, Plans, and Practical Reason*. Harvard University Press, 1987.
6. P. Bremner, L. Dennis, M. Fisher, and A. Winfield. On proactive, transparent and verifiable ethical reasoning for robots. *Under Revision for IEEE Transactions special issue on Machine Ethics*, 2018.

7. Louise A. Dennis and Michael Fisher. Practical Challenges in Explicit Ethical Machine Reasoning. In *International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, USA, 2018. Also available as ArXiv pre-print 1801.01422.
8. P. Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15, 1967.
9. D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors. *Handbook of Deontic Logic and Normative Systems*. College Publications, London, UK, 2013.
10. J.C. Harsanyi. Rule utilitarianism and decision theory. *Erkenntnis* (1975-), 11(1):25–53, 1977.
11. AI High Level Expert Group. Ethics guidelines for trustworthy AI. *B-1049 Brussels*, 2019.
12. A. S. Rao and M. P. Georgeff. An Abstract Architecture for Rational Agents. In *Proc. Int. Conf. Knowledge Representation and Reasoning (KR&R)*, pages 439–449. Morgan Kaufmann, 1992.
13. A. S. Rao and M. P. Georgeff. BDI Agents: From Theory to Practice. In *Proc. 1st Int. Conf. Multi-Agent Systems (ICMAS)*, pages 312–319, San Francisco, USA, 1995.
14. W.D. Ross. *The Right and the Good*. Oxford University Press, 1930.
15. Steve Taylor, Michael Boniface, Brian Pickering, Michael Anderson, David Danks, Asbjørn Følstad, Matthias Leese, Vincent Müller, Tom Sorell, Alan Winfield, and Fiona Woollard. Responsible AI – key themes, concerns & recommendations for european research and innovation. Project Report 10.5281/zenodo.1303252, July 2018.
16. D. Vanderelst and A. Winfield. An Architecture for Ethical Robots Inspired by the Simulation Theory of Cognition. *Cognitive Systems Research*, 2017.
17. Alan F. T. Winfield, Christian Blum, and Wenguo Liu. Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish, editors, *Advances in Autonomous Robotics Systems*, volume 8717 of *Lecture Notes in Computer Science*, pages 85–96. Springer, 2014.
18. M. Wooldridge. *An Introduction to Multiagent Systems*. John Wiley & Sons, 2002.
19. M. Wooldridge and A. Rao, editors. *Foundations of Rational Agency*. Applied Logic Series. Kluwer Academic Publishers, 1999.