Reasoning about ethical and legal responsibility using formal logic

Sjur K Dyrkolbotn and Jan M Broersen

Department of Philosophy and Religious Studies, Utrecht University

Developing a principled and effective approach to *responsibility tracking* is a key challenge in robot ethics, asking us to develop models of how responsibility relations form alongside complex causal chains involving the agency of robots and other AI systems.[1] The complexity of providing such models, combined with the need for clarity and precision, means that formal logic is a good candidate for choice of methodology. Below is a brief summary of three key challenges, accompanied by some possible applications of formal logic.

Causality

The first step towards effective responsibility tracking is a set of conditions to determine when a given combination of actions counts as a (contributing) *cause* of an event. Philosophers and legal scholars have done important unifying work on legal and moral causality, work that has resulted in abstract formulations of various causal tests that have also been (partly) formalised.[2]

More work needs to be done on combining these refined approaches to causality with existing logics of time and agency.[3] Moreover, open conceptual questions concern the best way to approach *proximate cause* doctrines[4], the notion of *moral luck*[5], and the controversial idea (both in philosophy and law) that causal attributions (should) depend on normative reasoning.[6] A motivating case that illustrates these three questions is the following: who caused the so-called flash crash in 2010, when the Dow Jones index plummeted by 5 per cent in three minutes?[7]

According to the US regulator, a single individual trading out of his suburban home outside of London was a "contributing factor", apparently guilty because he tricked robot traders into behaving irrationally.[8] This man was arrested in April, indicted in early September, and is currently awaiting a decision on whether he will be extradited to stand trial in the US. Was his actions truly the cause of the flash crash, or should attention rather be directed at the robot traders themselves and their inherent inadequacies as (moral?) agents on financial markets?

Expectation

Judgements about responsibility for intelligent systems are strongly shaped by expectations regarding the capabilities of such systems. How can we formalise such expectations, to ensure that the systems themselves, those who program them, and those who use them, share a common understanding of what can reasonably be expected? Moreover, how can we track the normative implications of robot behaviour that fails to live up to reasonable expectations?

---

[1] The authors are working on this as part of the ERC-funded REINS project at Utrecht University, see Jan Broersen, "Responsible Intelligent Systems" (2014) 28(3) KI – Künstliche Intelligenz 209.

[2] See generally Richard W Wright, "Causation, responsibility, risk, probability, naked statistics, and proof: pruning the bramble bush by clarifying the concepts" (1988) 73(5) Iowa Law Review 1001; Matthew Braham and Martin van Hees, "An Anatomy of Moral Responsibility" (2012) 121(483) Mind 601; Joseph Y Halpern, "Cause, responsibility and blame: a structural-model approach" [2015] Law, Probability and Risk ([to appear in print, available online first]).

[3] For instance *stit*-logic or ATL, see Nuel Belnap, Michael Perloff, and Ming Xu, *Facing the future: agents and choices in our indeterminist world* (Oxford University Press 2001); Rajeev Alur, Thomas A Henzinger, and Orna Kupferman, "Alternating-time temporal logic" (2002) 49(5) J. ACM 672.

[4] Used to determine when a cause is "close" enough to the outcome to be morally and/or legally relevant (e.g., selling a gun to a killer is typically too far away to be relevant to a murder).

[5] The idea that one's (degree of) moral and/or legal responsibility can depend on events over which one has no control.

[6] See, generally David Rose and David Danks, "Causation: Empirical Trends and Future Directions" (2012) 7(9) Philosophy Compass 643; Christian von Bar, "Causation or attribution" in *The Common European Law of Torts vol. 2* (Oxford University Press 2000).

[7] See generally Andrew J Keller, "Robocops: regulating high frequency trading after the flash crash of 2010" (2012) 73(6) Ohio State Law Journal 1457.

[8] There is no indication that he anticipated the algorithmic chain reaction that eventually led to the flash crash, see Andrew Verity, Could one man cause a stockmarket crash?, "BBC" (May 6, 2015) ⟨http://www.bbc.com/news/business-32598084⟩ visited on September 20, 2015.

One approach is to combine existing frameworks for deontic logic with base logics developed to formalise agentive causation.[9] Deontic standards can then be formalised relative to a model of agency and time that allows moral inferences to be drawn across causal chains.

A concrete challenge is the following: map out differences and similarities between two candidate AI systems, in terms of the expectations it is reasonable to form about them, e.g., the *da Vinci* surgical system on the one hand and the (potential) AI diagnostician *Watson* on the other.[10] Intuitively, it can be tempting to view the former as a tool, while the other (also) as an autonomous agent. If we can make this intuition more precise and explicate its moral content using formal logic, it would be an important contribution to ethical and legal reasoning about both these systems.

Guilt

Can a machine cause harms on purpose, or do damage as a result of subjective negligence? The question has a potential philosophical aspect to it, but it is also a practical question. Specifically, even if robots cannot think, feel or intend, postulating notions of *culpa* and *mens rea* that apply to them might well be expedient for both ethical and legal reasoning purposes.

Doing so using formal tools is well within reach; important progress on defining and classifying different kinds of *mens rea* using logic has already been made.[11] Further work should be devoted to coming up with appropriate notions of *culpa* and *mens rea* for more refined models of agentive causation and deontic expectation.

A concrete challenge to motivate this work is to adequately model the ethical implications of intelligent systems that actively *deceive* their users. Such systems are becoming increasingly prevalent, and honesty as a design principle appears to be under pressure in the computer science industry.[12] Indeed, it has been argued recently that "benevolent" forms of deception should be embraced as a way of shaping user perception.[13] But when is deception benevolent, and with respect to whom? This is not clear, particularly not in high-risk situations. Indeed, the deceptive turn in computer science raises the prospect of a future where we have to deal with intelligent systems that not only cause harms, but then also *lie* about it afterwards. Here it seems that a notion of *culpa*, or even *mens rea*, applied directly to the behaviour of intelligent systems, might well be both warranted and useful.

Conclusion

This has been a brief sketch of three building blocks for (different kinds of) responsibility attribution for which an analysis based on formal methods has a lot to offer. Importantly, approaching these notions from the angle of responsibility tracking leads us to investigate issues not typically raised in machine ethics. Indeed, none of the points discussed above are primarily about how to teach ethics to robots.[14] Rather, they concern how we can develop our own moral reasoning about AI technology, to tackle new ethical challenges that await us in the robot age.[15] Logic might well have a crucial role to play also in this regard, above and beyond what it can offer as an instrument for ethical AI design.

---

[9] There is much related work going on in computer science and philosophy at the moment. The best known example so far is probably John Horty, *Agency and Deontic Logic* (Oxford University Press 2001).

[10] The da Vinci system is already in wide use and liability disputes have already arisen, see Keith Kirkpatrick, "Surgical robots deliver care more precisely" (2014) 57(8) Communications of the ACM 14; Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (Springer 2013) 91-95. For a legal assessment of the liability questions that arise with respect to Dr. Watson (still a system in development), see Jessica A Allain, "From Jeopardy! to Jaundice: The Medical Liability Implications of Dr. Watson and Other Artificial Intelligence Systems" (2013) 73 Louisiana Law Review 1049.

[11] See Jan Broersen, "Deontic epistemic stit logic distinguishing modes of mens rea" (2011) 9(2) Journal of Applied Logic 137.

[12] See Kate Green, "How Should We Program Computers to Deceive?" *Pacific Standard* (Santa Barbara, California, September 3, 2014) ⟨http://www.psmag.com/nature-and-technology/technology-deception-elevator-crosswalk-programming-robots-lie-89669⟩ visited on May 23, 2015.

[13] Eytan Adar, Desney S Tan, and Jaime Teevan, "Benevolent Deception in Human Computer Interaction" (CHI '13, 2013).

[14] This challenge has received much attention recently, with formal logic apparently gaining ground as the formalism of choice for those who address it, see Boer Deng, Machine Ethics: The Robot's Dilemma, "Nature" (7558, 2015).

[15] An analogy can be made with how logic has been successfully used in the context of hardware verification, to tackle the growing complexity of hardware design, see generally Orna Grumberg and Helmut Veith (eds.), *25 Years of Model Checking: History, Achievements, Perspectives* (Springer 2008).