Joint Proceedings of the ACM IUI Workshops
2023, March 2023, Sydney, Australia

Norms and Causation in Artificial Morality

Laura Fearnley [1]

[1] *University of Glasgow, University Avenue, Glasgow, UK*

Abstract

There's been increasing interest into how to build Artificial Moral Agents (AMAs) that make moral decisions on the basis of causation rather than mere correction. One promising avenue for achieving this is to use a causal modelling approach. This paper explores an open and important problem with such an approach; namely, the problem of what makes a causal model an appropriate model. I explore why we need to establish criteria for what makes a model appropriate, and offer-up such criteria which appeals to normative considerations.

Keywords

Artificial Moral Agents, Causation, Normativity, Philosophy

1. Introduction

Artificial Morality is an emerging interdisciplinary field that centres around the creation of artificial moral agents, or AMAs, by implementing moral competence in artificial systems. The demand for moral machines comes from the changes in our everyday practices; artificial systems are rapidly being used in a variety of situations from home help and elderly care purposes to banking and court algorithms. It is therefore crucial to create reliable and responsible machines that make sound moral judgements. In this paper I introduce some cases from the philosophy of causation literature that generate problems for developing efficient and accurate AMAs. I also investigate how an appeal to normative considerations can provide a potential solution to these problems.

2. Causal Models

Plausibly morality deals in causation rather than mere correlation. As such there's recently been a growing interest in how to build AMAs that make moral decisions based upon cause and effect. One popular methodology for achieving this goal has been to use a causally modelling approach. Advocated of this approach often take as their point of departure the idea that causal relationships are relationships that are potentially exploitable for the purposes of manipulation and control. Roughly, if X is a cause of Y, then I should be able to manipulate X in the right way that would bring about a change in Y. In this way, causal relationships are thought to be relationships of dependency potentially exploitable for manipulation and control — X's causal status in regards to Y depends upon how Y reacts under changes to X. Typically the causal modelling approach takes the dependency relation to be one that holds between variables and their values (James Woodward 2003). Variables can be taken to represent one's preferred choice of causal relata — events, facts, properties, instantiations etc. Whether one variable is a cause of another is determined by whether some manipulation on the first variable changes the second variable; that is, whether a change in one variable makes a difference to another.

Following Judea Pearl (2000), the causal model is formalized using causal Bayes nets. These comprise of systems of structured equations and directed graph, which taken together, represent the causal relationships within the model. Directed graphs consist of an ordered pair ***{V, E}***, where ***V*** is a set of variables representing the causal relata**,** and ***E*** is a set of directed edges (arrows) representing the causal

Proceedings Joint Proceedings of ACM IUI, March, 2023 Sydney, Australia

EMAIL: laura,fearnley@glasgow.ac.uk (A. 1)
ORCID: XXXX-XXXX-XXXX-XXXX

structure by way of connecting the causal relata. Structural equations, on the other hand, define the causal structure between the variables in the model.

As opposed to other models, which use statistical predications to track mere correlation, the structural causal model approach relies upon counterfactuals and structural equations to determine bone fide causal relations. Given that morality relies upon causation, rather than correlation, the interventionist's causal modelling approach promises to provide an excellent starting point for informing artificial moral decisions.

Despite its initial appeal however, there is still much work to be done before the structural causal model approach can be fully implemented. One pressing difficulty is to identify what exactly makes a structural causal model an *appropriate* model. That is, what kind of things ought to be represented in the model in order for it to accurately and sufficiently express the essential causal structure of the actual situation. To illustrate, consider the following cases:

Case 1 – Forest Fire: Suppose I wanted to launch an inquiry to determine the causes of a forest fire. What variables ought to be included in the model? It seems reasonable to include a variable that represents the occurrence of the lightning strike, but it's less clear whether one should include a variable representing the presence of oxygen in the atmosphere, or whether oxygen should be relegated to a mere background condition. Whether we do include oxygen in the model will have a decisive effect on what kind of causal information is produced by the model. This is because manipulations to the presence of oxygen in the model, will make a different as to whether the forest fire occurs. For instance, changing the value of oxygen in the model from 1 to 0 will create a change in the occurrence of the forest fire – turning it from 1 to 0. As a result, oxygen would be a cause (rather than mere background condition) to the fire.

Case 2 – Plant Watering: Suppose I wanted to launch an inquiry to determine the causes of the death of my house plant. It seems reasonable to include in the model my failure to water my plant. It seems less reasonable to include, say Bono's failure to water my plant. Again, whether we include Bono will make a difference to the causal information produced by the model. Suppose we change Bono's not watering the plant, to Bono's watering the plant, then a manipulation on the variable representing Bono's failure to water will make a difference as to whether the plant dies. Thus, the model would determine Bono as a cause of the plant's death. This is surely the wrong result. We need some way to screen-off these irrelevant variables and values, lest we are left with erroneous causal verdicts.

Settling the question of what makes a model appropriate is an open and important problem in the philosophical and scientific literature. According to Paul and Hall (2013), it is also a problem that has been inequality addressed.

3. Causal Models and AMAs

Supplying criteria for what makes a model appropriate is crucial in the creation of AMAs (Kušić and Nurkic 2019). For if AMAs are to make moral decisions based upon faulty causal information generated by these models, then plausibly the moral decisions themselves will be flawed. Consider again *Case 2 – Plant Watering.* Suppose that we do include Bono's failure to water the plant as a variable in the model, and that therefore the model does recognise him as a cause of the plant's death. Thus the model establishes a causal connection between Bono and the dead plants. This causal connection can then partly justify and inform allocations of moral culpability. Yet, surely it is absurd to think that Bono is in anyway morally culpable for my dead house plants.

This is a simple toy example to illustrate the pitfalls of the causal modelling approach. But we can well imagine the implications of such errors in high-stakes moral domains, such as prison sentencing and medical treatment.

4. Norms and Causal Models

The lesson from these examples is that we need clearer criteria for establishing the aptness of causal models. Otherwise, AMAs which use such models to inform their moral decision-making will generate surprising, and perhaps unsettling moral decisions. In this final section, I'll explore criteria for establishing the aptness of a model.

One promising avenue for specifying the aptness of a model draws heavily on normative considerations. In particular, considerations about what's normal or abnormal. The idea that causal relations are sensitive to what's normal and abnormal is often credited to Hart and Honoré (1985). They contend that that a cause should be understood as an intervention, analogous to a human action, that makes a difference to the way things normally develop. For instance, “[w]hen we assert that A's blow made B's nose bleed or A's exposure of the wax to the flame caused it to melt, the general knowledge used here is knowledge of the familiar way to produce, by manipulating things, certain types of change which do not normally occur without our intervention.” (1985, p.31). Since Hart and Honoré, several philosophers, including McGrath (2005), Menzies (2009), and Hall (2007) have begun to invoke normality into their theories of causation. Some have even done so in the context of the causal modelling approach to overcome the problems of what makes a model apt Hitchcock (2007), Halpern (2016).

The strategy begins by using considerations about what’s normal and abnormal to constrain the kinds of values and variables to be represented in the model. Specifically, the idea is that the variables and values which go into the model ought to include abnormal occurrences. Whilst, the variables and value that should be omitted from the model should include abnormal occurrences.

To illustrate the strategy consider *Case 1 – Forest Fire.* Here we were wondering whether to include the presence of oxygen into the causal model; if it were included it would likely come out as a cause of the forest fire since a manipulation on the presence of oxygen would cause a change in the occurrence of the forest fire. A strategy which appeals to normative considerations would say that the presence of oxygen ought to be omitted from the model, because the occurrence of oxygen in earth’s atmosphere is normal. Thus, oxygen would not be a cause of the fire. This strategy gives us the right result. Plausibly, we want to say that oxygen is a mere background condition to the fire (not a cause).

Next consider *Case 2 – Watering Plants*. Here we wanted some way to exclude Bono’s failure to water the plant from entering into the model. For if his failure was represented in the model, it would come out as a cause of the plant’s death. Again, an appeal to normality allows us to do this. Bono’s failure to water my plants is a normal occurrence. It is both statistically and prescriptively normal for Bono *not* to walk into my house, watering can in hand, to water my plants. Hence the variable representing his failure should not be represented in the model. Again, this gets the right result – Bono is not a cause of the plant’s death.

As these two examples illustrate, an appeal to normative considerations to govern what kind of variables and values are represented in a model yields highly intuitive results. In particular, it yields causal information that seems to be *correct*. Importantly, correct causal information is the kind of information that AMAs ought to be basing their morally charged decisions on. In this way an appeal to normative considerations in the causal modeling mythology provides a promising pathway to overcoming some problems in the development of AMAs.

References

[1] Hall, N. Structural Equations and Causation. Philosophical Studies, (2007). 132(1), 109–136.

[2] Hall, N., & Paul, L. A. Metaphysically Reductive Causation. Erkenntnis, (2013). 78(S1), 9–41.

[3] Halpern, J. Y. Actual Causality. 2016 The MIT Press.

[4] Hart, H. L. A., & Honoré, T. Causation in the Law. 1985 Second Edition. Oxford University Press.

[5] Hitchcock, C. Prevention, Preemption, and the Principle of Sufficient Reason. The Philosophical Review, (2007). 116(4), 495–532.

[6] Kušić, Marija & Nurkić, Petar. Artificial morality: Making of the artificial moral agents. 2019. Belgrade Philosophical Annual 1 (32):27-49.

[7] McGrath, S. Causation By Omission: A Dilemma. Philosophical Studies, 2005. 123(1–2),

[8] Menzies, P. Platitudes and Counterexamples. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), 2009. The Oxford Handbook of Causation (Vol. 1). Oxford University Press.

[9] Pearl, J. Causality. 2000. Cambridge University Press.

[10] Woodward, J. Making Things Happen: A Theory of Causal Explanation. 2003 Oxford University Press.