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 has been an 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 centers around the creation of artificial moral agents (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 which use causal modelling frameworks. I also investigate how an appeal to normative considerations can provide a potential solution to these problems.

2. A Problem: Causal Models and Faulty Causal Information

Machine learning algorithms that make decisions based upon statistical correlations have produced unjust and discriminatory outcomes. In the judicial system, for example, AMAs have informed prison sentencing decisions by calculating how likely it is that a defendant will reoffend. The AMA makes these recidivism risk assessments on the basis of statistical patterns found data sets. This has meant that defendants with certain characteristics, such as low incomes, were being assigned particularly high recidivism risk scores because low incomes are correlated with reoffending (Villasenor and Foggo 2020). To safeguard against discriminatory outcomes such as these, there has been a substantial shift towards developing AMAs that make morally charged decisions on the basis of cause and effect rather than mere correlation.

There's a plethora of causal solutions to developing AMAs, but it is counterfactual analyses that are gaining serious traction in the field (Machamer et al., 2000). Especially promising are counterfactual approaches that are formalised through structural causal models (Woodward 2003). Advocates of this approach take as their point of departure the philosophical idea that causal relationships are relationships that are potentially exploitable for the purposes of manipulation and control. According to this view, if X is a cause of Y, then I should be able to manipulate X in a way that would bring about an associated 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

Proceedings Joint Proceedings of ACM IUI, March, 2023 Sydney, Australia
EMAIL: laura,fearnley@glasgow.ac.uk (A. 1)
ORCID: XXXX-XXXX-XXXX-XXXX

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. 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 models are represented 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 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 AMA 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 accurate and just moral decisions must be made on the basis of causation, rather than correlation, the 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 variables 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 lightning hitting the tree. But what's less clear is whether one should include a variable representing the presence of oxygen in the earth's atmosphere or whether the presence of oxygen should be considered as a mere background condition and therefore excluded from being represented in the model. Importantly, 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 will bring about an associated change in the occurrence of the forest fire. For instance, changing the value of oxygen from its actual value of 1 (is present in the atmosphere) to 0 (not present in the atmosphere) will create a change in the occurrence of the forest fire – turning it from 1 (the fire occurs) to 0 (the fire does not occur). As a result, the model would determine the presence of oxygen as a cause of the fire. This is potentially problematic insofar as oxygen is typically considered a mere background condition for fires (not a cause of them).

Case 2 –Flowers: Suppose I wanted to launch an inquiry to determine what caused the death of my flowers. It seems reasonable to include in the model a variable which represents the gardener's failure to water my flowers. It seems considerably less reasonable to include, say, U2 singer Bono and his failure to water my flowers. But again notice that if Bono's omission is represented in the model, it will make a difference to the causal information produced by that model. Suppose we do express his omission in the model, and then intervene on it changing its actual value from 0 (not watering the flowers) to 1 (watering the flowers). A manipulation of this kind will bring about an associated change in the state of the flowers, tuning them from 0 – dead – to 1 – alive. Thus, the model would determine Bono as a cause of the flower's death. This is surely the wrong result. Despite the fact that Bono and the gardener failed to do exactly the same thing, only one of their omissions has causal salience. We therefore need some way to screen-off these irrelevant variables and values, lest we are left with erroneous causal verdicts.

3. Causal Models and AMAs

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 inadequately addressed. We must make progress in this area if causal analyses are to underpin machine learning approaches to AMA (Kušić and Nurkic 2019). For if AMAs are to make moral decisions based upon

faulty causal information generated by these models, then the moral decisions themselves will be flawed. Consider again *Case 2 – Flowers.* Suppose that we do include Bono's failure to water the flowers as a variable in the model, and that therefore the model recognises him as a cause of the flower's death. The established causal connection can partly justify and inform allocations of moral culpability; causal responsibility being a necessary condition for moral culpability (Driver 2007). Yet, it is surely absurd to think that Bono is in anyway sense a morally salient factors in the death of my flowers.

This is a simple toy example taken from philosophy 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. In a model mapping out the causal factors of a patient's unexpected death, for example, do we include a variable expressing the doctor's failure to administer medical treatment? Yes, perhaps. Do we also include a variable expressing the hospital porter's failure to administer medical treatment? Surely not. Notice though that including the night porter's omission would make them a cause of the patient's death, since intervening on their failure by seeing them administer the medicine would produce an associated change in whether the patient dies. Yet, it would be detrimental to make any legal or moral judgments on the basis that there is a causal connection between the night porter and the patient.

4. A Solution: Normative Considerations

The lesson from these examples is that we need principled criteria for establishing the aptness of causal models. Otherwise, AMAs which use such models to inform their moral decision-making will potentially generate surprising, and often unsettling moral decisions. In this final section, I'll provide one solution 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 a cause should be understood as an intervention, analogous to a human action, that makes a difference to the way things would 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 invoked normality into their theories of causation. Some have even done so in the context of the structural causal modelling approach with the primary aim of identifying what makes a model appropriate (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 represent abnormal occurrences. Whilst, the variables and values that should be excluded from the model should represent normal occurrences. The notion of normality beyond deployed by these philosophers is of a prescriptive and statistical kind. To say something is statistically normal is to say that it conforms to a statistical mode. For example, it is statistically normal for Glasgow to have more rainfall than Milan during the winter, so if Glasgow were to have less rainfall than Milan one winter, Glasgow's weather would violate a statistical norm. By contrast, to say something is normal in a prescriptive sense is to say that it follows a prescriptive rule governing the way things ought to be or are supposed to be. So it would be normal for me to keep my promise because I morally ought to keep promises. Broadly speaking then, a variable can be categorised as normal to the extent that it abides by statistical and prescriptive norms.

To illustrate how ideas of normality and abnormality can set parameters around what makes a model apt, let's return to the previous cases. Consider *Case 1 – Forest Fire.* Here it seemed obvious to represent the lightning hitting the tree in the model, but it seemed less obvious to include the presence of oxygen. The problem is that oxygen is typically thought to be a background condition for the occurrence of fires, not causes of them. Hence, we need a strategy that excludes oxygen from being represented in the model. Incorporating normative considerations allows us to do exactly this. The occurrence of oxygen in the earth's atmosphere is statistically

normal, and as such should not be represented in the model. Whereas, lightning strikes are statistically unusual and therefore should be represented in the model. This would make the lightning strike, and not oxygen, a cause of the fire, giving us the right result.

Next consider *Case 2 – Flowers*. Here we wanted some way to exclude Bono's failure to water the flowers from entering into the model, for if his failure was represented in the model, it would come out as a cause of the flower's death. Again, an appeal to normality allows us to do this. Bono's failure to water my flowers is a normal occurrence. It is both statistically and prescriptively normal for Bono *not* to walk into my garden, watering can in hand, to water my flowers. 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 flower's death. Conversely, the gardener's failure to water the flowers is abnormal; her failure deviates from a statistical norm and presumably a prescriptive norm – contractually she ought to water my flowers. Hence an apt model will represent the gardener's failure, but not Bono's.

As these two examples illustrate, an appeal to normative considerations to govern what kind of variables and values are represented in a model in a way that 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 methodology provides a promising pathway to overcoming some of the problems in the development of AMAs.

References

[1] Driver, J. Attributions of causation and moral responsibility. In Moral psychology, Vol 2: The cognitive science of morality: Intuition and diversity. (2007). MIT Press.

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

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

[4] Halpern, J. Y. Actual Causality. (2016). The MIT Press.

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

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

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

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

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

[10] Pearl, J. Causality. (2000). Cambridge University Press.

[11] Villasenor, J & Foggo, V. Artificial Intelligence, Due Process, and Criminal Sentencing. Michigan State Legal Review (2020.) 295

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