The Philosophy of X in XAI

Neil McDonnell [1]

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

Abstract

Explanation is a topic in its own right in philosophy, and a topic of newfound interest in AI research given the need in some domains for explainable AI (XAI). This paper traces some of the progress in the philosophical discourse and applies to a realistic application of AI where explanation is required. The aim is to show that philosophy may be of use in the search for the X in XAI.

Keywords

Explanation, AI, XAI, Philosophy

1. Introduction

Explanation is a central topic in the philosophy of science and it retains its status in part because there is, as yet, no consensus on what the necessary and sufficient conditions are for counting as an explanation. This makes the challenge of producing explainable AI (XAI) a philosophically interesting one. In this short paper I introduce some toy cases from the philosophy literature that illustrate the challenge of accounting for explanation, and draw on recent philosophical progress to sketch a potential path forward.

2. Causal Dependence

In classes around the world the phrase "correlation is not causation" is drummed into students, but it remains contentious what the missing ingredient is that you need to add to correlation to get genuine causation. David Lewis [1] offered an answer: dependence. Whereas two common effects of a cause (say, the bang and the muzzle flare of a fired gun) correlate with one another, we know they are not mutual causes. As Lewis pointed out the flare does not depend on the bang or vice versa – they both depend on the firing of the gun. This is what causation has that correlation does not: dependence.

When we ask "why was there a bang?" one obvious answer is "because a gun was fired" and thus we cite a cause in giving an explanation, and we find the cause by examining the dependency in the situation. There is a problem though:

Case 1 – Late Pre-emption: Billy and Suzy are throwing rocks at a window. Both are accurate, but Suzy throws harder and her rock reaches the window first. The window breaks, then Billy's rock passes through the empty space.

This case is a famous counterexample to Lewis' dependence thesis. In this case it is obvious that Suzy is the cause of the window breaking, but because Billy is there as backup the window breaking does not depend on Suzy. So, Suzy is the cause (by common sense) but Suzy is not a cause (by Lewis' theory). So much the worse for Lewis' theory.

But there is an obvious comeback that shows a problem for explanation. The window breaking *rather than not breaking at all* did not depend on Suzy (hence the problem) but the window breaking *exactly like that* rather than a fraction of a second later did depend on Suzy's throw. This shows that we can talk about the same event (window breaking) two different ways and come to two different conclusions about what it depended on and hence what causes or explains it.

Neil McDonnell. 2023. The Philosophy of X in XAI. In Joint Proceedings of the ACM IUI Workshops 2023, March 2023, Sydney, Australia EMAIL: neil.mcdonnell@glasgow.ac.uk (A. 1)
ORCID: 0000-0001-7279-5277

So explanation is sensitive to the description or categorization of the consequent event.

A parallel issue afflicts the antecedent (cause) event. Here is another famous case [2]:

Case 2 – Sophie: Sophie the pigeon is trained to peck all and only red patches. A scarlet patch is placed in front of Sophie and she pecks.

There are three candidate explanations we can offer for why Sophie pecked:

1. Because a scarlet patch was placed.
2. Because a red patch was placed.
3. Because a coloured patch was placed.

Stephen Yablo [2] introduced this case to show that we have a strong preference for explanation 2. Explanation 3 is not specific enough because it might lead someone to think a blue patch could have worked instead. Explanation 1 is too specific because it might make you think your crimson patch would not have triggered a peck. These misleading implications make 1 and 3 less good than 2 as an explanation. If you agree with Yablo on this then it shows that explanation is also sensitive to how you describe or characterize the antecedent event, and that it is easy to give a misleading explanation.

Complicating the case a little more, we can specify that the lab in which Sophie is being experimented on only has two types of patches: scarlet and transparent (colourless). If we know this additional information it seems that explanations 1 and 3 are no longer misleading about crimson or blue patches since they are already ruled out as viable alternatives. This external fact about the context seems to change the quality of an explanation without changing anything about the specific interaction between Sophie and the scarlet patch that we are seeking the explanation about. This shows that the quality of an explanation can vary with contextual information about viable alternatives.

3. A Realistic Problem

There are a host of other problem cases from the causation literature that are relevant to the wider issue of XAI, but these examples illustrate one strand where recent progress has offered a potential solution. I will illustrate with the realistic case of loan viability as assessed by AI.

Loans cannot legally be denied on the basis of a protected characteristic in (at least) Germany and the United States [3] and to protect against abuse of this the candidate loanee is entitled to an explanation if they are rejected. If an AI is used to reach that determination in a more efficient way, then it must be XAI so that the legal requirement for an explanation is met.

Our Billy and Suzy case shows us one type of structure that could be a problem. Suppose the system rejects candidate P and an explanation is sought. The explanation offered is that P's employment contract is too short, but what is not made clear is that the system would have rejected P in any case based on P's ethnicity (due to a biased historical dataset, let us suppose). The system is clearly flawed but this explanation disguises the fact.

This Sophie scenario also showed us a problem that emerges in this scenario. The explanation offered (that P's contract is too short) implies that extending the contract will change the verdict. It won't in the case as described, and so whilst it does seem to qualify as an explanation, it is an incomplete or misleading one that obscures the problematic reasoning that is waiting in the wings. It is analogous in a way to the first explanation in the Sophie case (that a scarlet patch was placed) since the explanation masks the presence of an alternative cause, crimson in the Sophie case, ethnicity in the loan case.

4. Lessons from Philosophy

The lesson from these examples is that we need a better form of explanation. This is where some recent work in philosophy can help. The three main insights are that good explanations often have a contrastive structure, that we care about what we can intervene upon, and that robust/stable dependence relations make for better explanations. I will unpack each briefly.

Contrastive explanations do not seek to explain just the outcome in isolation (e.g. the broken window) but to explain the difference between two states: the window breaking at that moment *rather than* slightly later [4], [5]. Suzy made that difference but did not make the difference between it breaking and not breaking at all. In our other example, 'placing a scarlet patch *rather than* no patch at all' does explain Sophie's peck, but 'placing a scarlet patch *rather than* a crimson one' does not. Thus, making our explanation query contrastive in the form "Why X rather than Y?" is likely to yield a better explanation. Applied to the loan case, if we ask "Why was P rejected *rather than* accepted for the

loan?" the answer cannot just be that the contract was too short, since a longer contract would not have brought about the second contrast (acceptance). This may then flag up the problematic ethnic profiling that was previously disguised. It also helps avoid the ambiguity about what relevant alternatives are viable in the context (blue?, transparent?) since the alternative is made explicit in the contrastive target.

A related view of explanation from Woodward [6] holds that what we care about is what we need to intervene upon to get the outcome that we want. To stop the window breaking we need to intervene on *both* Suzy and Billy. To get Sophie to peck we need to ensure some kind of red patch is placed, but we need not intervene to make it some specific shade. And in the case of P's loan they need to change both their contract and (absurdly) their ethnicity to get the loan. Thus making explicit the interventions required to change the outcome from one outcome to an explicitly stated contrasting outcome gives a richer explanation.

Finally, it is worth noticing that both the output of our target process can be graded, and so can the inputs that yield that output. For example, it might be the case that an applicant could be offered a larger or smaller loan, based on better or worse rates, depending on how risky a prospect the system takes them to be. Suppose Q applies for a loan and is accepted at a lower amount and poorer rates than hoped. A good explanation of this outcome – why the loan offered was low and expensive *rather than* higher and cheaper – will show Q what variables to intervene on for a better outcome (reduce outgoings, clear existing debt, extend contract etc.). It makes a difference whether Q needs to remove just one of these barriers, two of them, or all three before getting the desired outcome and so an even better explanation will additionally express how *robust* the connection between these explainers and the outcome is. *Robustness*, or *sensitivity* as it is sometimes known [7], is a measure of the range of counterfactual scenarios where the putative cause is present and the effect still occurs. A small range indicates that the relationship is sensitive, a broader range indicates that it is robust. In general, citing more robust causes provides a better explanation as it extends into more scenarios.

5. Lessons from Philosophy

I have here briefly shown the benefits of Contrastive [4], [5], and Interventionist [6] approaches to explanation, and introduced the recent insights about causal robustness [7] from the philosophical literature. The aim has been to show the potential for philosophical reasoning around causation and explanation to inform the desiderata for what counts as explainable in XAI. A highly influential figure in these recent debates, both in philosophy and computer science, is Judea Pearl [8], [9], and it is to his formalism for capturing the sorts of counterfactual dependence relations at the heart of this discussion that I direct interested practitioners.

6. References

[1] D. Lewis, Causation, Journal of Philosophy 70(17) (1973) 556-567. doi: 10.2307/2025310

[2] S. Yablo, Mental Causation, Philosophical Review 101 (2), (1992) 245-280. doi: 10.2307/2185535

[3] N. Savage, Breaking into the black box of artificial intelligence, 2022. URL: https://www.nature.com/articles/d41586-022-00858-1

[4] B. van Frassen, The Scientific Image, Oxford University Press, Oxford, 1980.

[5] J. Schaffer, Contrastive Causation. Philosophical Review 114 (3) (2005) 327–358. doi: 10.1215/00318108-114-3-327

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

[7] L. Fearnley, Moral worth, right reasons and counterfactual motives, Philosophical Studies 179 (9), (2022) 2869-2890. doi: 10.1007/s11098-022-01805-6

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

[9] J. Pearl, D. Mackenzie, The Book of Why: The New Science of Cause and Effect, Allen Lane, 2018.