A Causal Analysis of Harm

Sander Beckers[1], Hana Chockler[2] and Joseph Y. Halpern[3]

[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]Department of Informatics, King's College London
[3]Computer Science Department, Cornell University

Abstract

As autonomous systems rapidly become ubiquitous, there is a growing need for a legal and regulatory framework that addresses when and how such a system harms someone. There have been several attempts within the philosophy literature to define harm, but none of them has proven capable of dealing with the many examples that have been presented, leading some to suggest that the notion of harm should be abandoned and "replaced by more well-behaved notions". As harm is generally something that is caused, most of these definitions have involved causality at some level. Yet surprisingly, none of them makes use of causal models and the definitions of actual causality that they can express. In our full paper [1] we formally define a qualitative notion of harm that uses causal models and is based on a well-known definition of actual causality [2]. The key features of our definition are that it is based on *contrastive* causation and uses a default utility to which the utility of actual outcomes is compared. We show that our definition is able to handle the examples from the literature, and illustrate its importance for reasoning about situations involving autonomous systems.

Keywords

harm, actual causation, default utility

1. Introduction

The notion that one should not cause harm is a central tenet in many religions; it is enshrined in the medical profession's Hippocratic Oath, which states explicitly "I will do no harm or injustice to [my patients]" [3] it is also a critical element in the law. Not surprisingly, there have been many attempts in the philosophy literature to define harm. Motivated by the observation that we speak of "causing harm", most of these have involved causality at some level. All these attempts have encountered difficulties. Indeed, Bradley [4] says:

Unfortunately, when we look at attempts to explain the nature of harm, we find a mess. The most widely discussed account, the comparative account, faces counterexamples that seem fatal. But no alternative account has gained any currency. My diagnosis is that the notion of harm is a Frankensteinian jumble . . . It should be replaced by other more well-behaved notions.

The situation has not improved much since Bradley's paper (see, e.g., recent accounts like [5, 6]). Yet the legal and regulatory aspects of harm are becoming particularly important now,

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
srekcebrednas@gmail.com (S. Beckers); hana.chockler@kcl.ac.uk (H. Chockler); halpern@cs.cornell.edu (J. Y. Halpern)
https://www.sanderbeckers.com (S. Beckers); https://www.hanachockler.com/ (H. Chockler); https://www.cs.cornell.edu/home/halpern/ (J. Y. Halpern)

as autonomous systems become increasingly more prevalent. In fact, the new proposal for Europe's AI act [7] contains over 25 references to "harm" or "harmful", saying such things as "...it is appropriate to classify [AI systems] as high-risk if, in the light of their intended purpose, they pose a high risk of harm to the health and safety or the fundamental rights of persons ..." [7, Proposal preamble, clause (32)]. Moreover, the European Commission recognized that if harm is to play such a crucial role, it must be defined carefully, saying "Stakeholders also highlighted that ...it is important to define ... 'harm' [7, Part 2, Section 3.1].

Imagine a UAV used by the military has to decide whether or not it should bomb a suspected enemy encampment. The problem is that the target is not clearly identified, because there are two camps close to each other: one consisting of civilian refugees, another consisting of a rebel group that is about to launch a deadly attack on the refugee camp, killing all of its inhabitants. The UAV's decision is based only on the expected utility of the refugees, and therefore it bombs the camp. Tragically, as it turns out, the camp was that of the refugees. Here we have the intuition that the UAV harmed these refugees, despite the fact that both actions would have led to all the refugees being killed. Examples in which one event (the bombing) preempts another event (the attack) from causing an outcome are known as *Late Preemption* examples in the causality literature; we discuss them in the full paper.

In the healthcare domain, autonomous systems are used for, among other things, classifying MRI brain images suspected of containing a tumor. If an image is classified as having a tumor, the system decides whether to recommend a surgery. While the overall accuracy of the system is superior to that of humans, in some instances the system overlooks an operable tumor. Imagine a patient who has such a tumor and dies from brain cancer as the result of not undergoing surgery, leading to a dispute between the patient's family and the hospital regarding whether the patient was harmed. Even if both parties agree that the patient would probably have been alive if the diagnosis had been performed by a human, the hospital might claim that using the system is the optimal policy, and therefore one should compare the actual outcome only to those that could have occurred under the policy.

Fortunately, the formal tools at our disposal to develop a formal notion of harm have also improved over the past few years; we take full advantage of these developments in our full paper [1]. Concretely, we provide a formal definition of harm that we believe deals with all the concerns that have been raised, seems to match our intuitions well, and connects closely to work on decision theory and utility. Here we briefly give a high-level overview of the key features of our approach and how they deal with the problems raised in the earlier papers.

There is one set of problems that arise from using counterfactuals that also arise with causality, and can be dealt with using the by-now standard approaches in defining causality. For example, Carlson, Johansson, and Risberg [5] raise a number of problems with defining harm causally that are solved by simply applying the definition of actual causality given by Halpern [8, 2]. The issue of whether failing to take an action can be viewed as causing harm (e.g., can failing to water a neighbor's plants after promising to do so be viewed as causing harm) can also be dealt with by using the standard definition of causality (which allows lack of an action to be a cause).

Just applying the definition of causality does not deal with all problems. The other key step that we take is to assume that there exists a *default* utility. Roughly speaking, we define an event to cause harm whenever it causes the utility of the outcome to be lower than the default utility. The default may be context-dependent, and there may be disagreement about what the

default should be. We view that as a feature of our definition. For example, we can capture the fact that people disagree about whether a doctor euthanizing a patient in great pain causes harm by taking it to be a disagreement about what the appropriate default should be. Likewise, the dispute between the family and the hospital described above can be modeled as a disagreement about the right default. Moreover, by explicitly bringing utility into the picture, we can connect issues that that have been discussed at length regarding utility (e.g., aggregating utilities of different individuals, weighting losses vs gains, etc.) to issues of harm. We develop a quantitative notion of harm in a follow-up paper that focuses on harm in the context of uncertainty and when involving multiple individuals [9].

Acknowledgements

Sander Beckers was supported by the German Research Foundation (DFG) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645, and by the Alexander von Humboldt Foundation. Hana Chockler was supported in part by the UKRI Trust-worthy Autonomous Systems Hub (EP/V00784X/1) and the UKRI Strategic Priorities Fund to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1). Joe Halpern was supported in part by NSF grant IIS-1703846, ARO grant W911NF-22-1-0061, and MURI grant W911NF-19-1-0217.

References

[1] S. Beckers, H. Chockler, J. Y. Halpern, A causal analysis of harm, Proc. Advances in Neural Information Processing Systems 35 (NeurIPS '22) (2022).
[2] J. Y. Halpern, Actual Causality, MIT Press, Cambridge, MA, 2016.
[3] N. L. of Medicine, Hippocratic oath, https://www.nlm.nih.gov/hmd/greek/greek_oath.html, 2002.
[4] B. Bradley, Doing away with harm, Philosophy and Phenomenological Research 85 (2012) 390–412.
[5] E. Carlson, J. Johansson, O. Risberg, Causal accounts of harming, Pacific Philosophical Quarterly (2021).
[6] N. Feit, Harming by failing to benefit, Ethical Theory and Moral Practice 22 (2019) 809–823.
[7] European Commission, Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 2021. https://artificialintelligenceact.eu/the-act/; accessed Aug. 8, 2021.
[8] J. Y. Halpern, A modification of the Halpern-Pearl definition of causality, in: Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), 2015, pp. 3022–3033.
[9] S. Beckers, H. Chockler, J. Y. Halpern, Quantifying harm, Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023) (2023).