Moral Responsibility for AI Systems

Sander Beckers
University of Amsterdam
sanderbeckers.com
srekcebrednas@gmail.com

Abstract

As more and more decisions that have a significant ethical dimension are being outsourced to AI systems, it is important to have a definition of *moral responsibility* that can be applied to AI systems. Moral responsibility for an outcome of an agent who performs some action is commonly taken to involve both a *causal condition* and an *epistemic condition*: the action should cause the outcome, and the agent should have been aware – in some form or other – of the possible moral consequences of their action. This paper presents a formal definition of both conditions within the framework of causal models. I compare my approach to the existing approaches of Braham and van Hees (BvH) and of Halpern and Kleiman-Weiner (HK). I then generalize my definition into a *degree of responsibility*.

1 Introduction

As more and more decisions that have a significant ethical dimension are being outsourced to AI systems, it is important to have a definition of responsibility that can be *applied* to the decisions of AI systems, and that can be *used by* AI systems in the process of its decision-making [8]. To meet the first condition, such a definition should require only a minimal notion of agency and instead focus on those aspects of responsibility that are readily applicable to (current) AI systems. To meet the second condition, such a definition should be formulated in a language that can be implemented into an AI system, so that it can integrate judgments of responsibility into its decision-making. This paper sets out to propose such a definition using the well-established framework of causal models [22, 23].

There exist different notions of moral responsibility that one might be interested in, and here we restrict attention to just one of them, namely *responsibility for consequences*, meaning the responsibility one has for a particular outcome that is the result of performing a particular action. This can be expressed more clearly by saying that the action *caused* the outcome, and therefore the first condition of concern here is the **causal condition** on responsibility [12, 26, 21, 5]. The past two decades have seen immense progress on offering formal definitions of actual causation by way of using causal models, and the definition here developed takes maximal advantage of this progress by comparing some recent proposals and choosing the one that correctly handles several complicated cases to be considered [33, 16, 2].

Our actions can cause all kinds of outcomes for which we are clearly not morally responsible: if a train crashes into a car that illegally crosses the railroad then the train conductor is not responsible for the car driver's death, if you turn on a light switch in a hotel room then you are not responsible if a short-circuit follows, etc.. The standard intuition that we have in such cases is that the agent "could not have known" that their action would cause the outcome. This is why definitions of responsibility also invoke an **epistemic condition**, stating roughly that the agent should have been able to foresee that they are performing an action which could result in them being responsible for the outcome [5, 27, 25].

Preprint. Under review.