Moral responsibility for AI systems

Sander Beckers (UvA)

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 our definition into a degree of responsibility.