The **Epistemic Condition** requires settling the question: what does it take for the agent to believe that performing $A = a'$ allows them to avoid responsibility? Since this condition uses the notion of responsibility, our Schema is circular. There exist different ways of filling it in so that it no longer is circular, and it is this flexibility that makes filling in the condition interesting. One possible suggestion is to demand that the agent believed $A = a'$ would not result in the outcome $O = o$, another weaker suggestion is to demand that the agent believed $A = a'$ would not cause $O = o$, etc..

A note of clarification is in order before we proceed. The current work does not aim to offer a complete theory of moral responsibility for AI systems, but rather zooms in on the above conditions whilst ignoring certain others. Concretely, here are some important issues that I set aside in this paper.

1.1 Some Limitations and Related Work

There exist forms of moral responsibility that do not (always) involve causation, such as those that follow from certain societal norms and expectations. For example, a captain is responsible for everything that happens on their ship, a parent is responsible for the behavior of their child, etc.. More generally, assigning responsibility to AI systems should itself be seen as just one part of the wider discussion on accountability that arises from the introduction of such systems into our society [19, 7].

Relatedly, responsibility is often associated with the morally stronger notions of blame and praise. I take responsibility to be a weaker notion that necessarily precedes judgments of blame and praise: one cannot be blameworthy for an outcome unless one is responsible for it, and similarly for praise. To develop definitions of blame and praise requires bringing into view both the *absolute* moral valence of the outcome $O = o$ (was it good or bad?) and its *relative* valence (was it better/worse than an alternative which it prevented?), as well as the costs incurred by the agent when performing an action. As the vast literature on trolley cases and other moral dilemmas illustrates, these issues make matters significantly more complex [10, 20].

One condition in particular that seems highly relevant to assigning blame (resp. praise) is to consider whether the outcome caused by the agent is *harmful* (resp. beneficial) or not. Indeed, one natural way of implementing a formal definition of responsibility within AI systems is to demand that it tries to avoid becoming responsible for harmful outcomes. This is confirmed by the recent European AI Act, which categorizes the risk that an AI system poses based on how likely it is to cause harm [11]. Beckers et. al. recently proposed a causal analysis of harm that is also formalized using causal models, and thus it could easily be integrated into the present proposal [3, 4].[2] In the present paper, however, I choose to focus exclusively on defining responsibility, thereby paving the way for future definitions of blame and praise.

Duijf presents a formalization of moral responsibility for outcomes that is likewise inspired by, but not an endorsement of, BvH [9]. Rather than defending an alternative definition of responsibility as I do, he presents a broad lanscape of completely formal conditions for responsibility that one might consider and analyzes their logical relations. As with BvH, his definition of NESS causation is formulated using game-theory, and thus it is likewise restricted to applications of direct causation.

The next section introduces the formalism of causal models that will be used to express all candidate definitions and related notions. Section 3 presents the BvH and HK definitions of responsibility and their respective definitions of causation. Section 4 discusses the **Causal Condition** by introducing two more definitions of causation and offers some examples to argue in favor of adopting the CNESS definition. (Some further examples are offered in the appendix.) We move on to a discussion of the **Epistemic Condition** in Section 5, which leads the way to my definition of moral responsibility. Since responsibility is often taken to come in degrees, in Section 6 I define the *degree of responsibility* and sketch how it helps interpret recent empirical work in psychology on responsibility judgments.

[2] I should note that they use the HP-definition of causation, which I criticize below. However, they state explicitly that their approach applies just as well to other definitions of causation.