In addition to the causal and the epistemic conditions, it is standard to demand that responsibility also requires the fulfilment of a **control condition** (sometimes also called freedom condition), which expresses the fact that the agent had the right sort of control whilst performing their action [25]. Due to its close connection to issues of free will and determinism, this condition is heavily debated within philosophy. Within the context of (current) AI systems, however, the control condition can take on a more mundane form: any action that was a result of the correct operation of its program can be viewed as being under the AI's control. Therefore I simply take there to be a specific action variable that ranges over a set of possible actions, and assume that whenever the AI system is running successfully it has control over the value that this variable takes.

My approach proceeds along the same lines as that of Braham and van Hees (BvH) [5]. They offer the most influential formalization of moral responsibility that incorporates both the causal and the epistemic conditions, and therefore their work forms an appropriate point of comparison. Although I agree with the spirit of their approach, I disagree with its formulation. First, their causal condition defines causation as being a *Necessary Element of a Sufficient Set* (NESS). However, their use of game-theory instead of causal models results in an overly simplistic view of NESS-causation that cannot handle indirect causation. Therefore I first formulate their definition using causal models, and then show how to modify it so that it can overcome this limitation. Second, I disagree with the particulars of both their causal and their epistemic conditions. I argue for replacing the NESS definition of causation with my recently developed *Counterfactual NESS* (CNESS) definition [2]. Their epistemic condition states that the agent should minimize the probability of causation. I argue for giving that condition a secondary role: minimizing the probability of causing the outcome is subservient to minimizing the probability of the outcome simpliciter. I analyze several examples to illustrate the superiority of my conditions.

More recently, Halpern & Kleiman-Weiner (HK) [17] used causal models to propose definitions of several concepts that are closely related to moral responsibility. Although they do not explicitly define moral responsibility, they do suggest using the modified Halpern & Pearl (HP) definition of causation for the causal condition [16]. The HP definition correctly handles most of the counterexamples to the NESS definition here presented, but I discuss two types of example for which it fails (whereas the CNESS definition does not). HK also offer a definition of "degree of blameworthiness" that for all intents and purposes is very similar to an epistemic condition: it measures the extent to which the agent minimized the probability of the outcome. I present a case in which the epistemic conditions of BvH and HK conflict in order to argue that a more elaborate epistemic condition is required. My epistemic condition combines that of HK with that of BvH by demanding that an agent minimizes the probability of the outcome, but if possible also minimizes the probability of causation. [1]

Here is the general schema that encompasses all definitions of responsibility that I aim to consider.

Responsibility Schema. *An agent who performs $A = a$ is responsible for outcome $O = o$ if:*

- **(Control Condition)** *The agent had control over $A = a$.*
- **(Causal Condition)** $A = a$ *causes* $O = o$.
- **(Epistemic Condition)** *The agent believes that they could have avoided being responsible for $O = o$ by performing some alternative action $A = a'$.*

As mentioned, I simply assume that the **Control Condition** is always met (as do BvH, who call it the Agency Condition). For sake of brevity, I leave it implicit from now on.

Formalizing the **Causal Condition** comes down to settling the discussion on how to formalize actual causation, which has received considerable attention over the past two decades [22, 31, 15, 30, 16, 1]. A full discussion of causation would be too ambitious for the present purposes. Instead, I evaluate the suitability of several definitions of causation within the context of responsibility by presenting examples that bring across how they differ. On the basis of this evaluation I suggest adopting the CNESS definition and refer the reader to [2] for a more general motivation.

[1] Note that the epistemic conditions of HK and BvH are not necessarily inconsistent: if one simply defines causation as an increase in the probability of the outcome occurring, they become equivalent. Except for the fact that he uses objective probabilities rather than those of the agent, this is roughly the proposal defended in [29]. As exemplified by the examples to be discussed (and as exemplified by browsing the recent literature on causation) such a naive probabilistic approach to causation is unable to deliver sensible verdicts.