to $A = a'$: $CS_e(o, a, a') = \Pr(O = o|[A \leftarrow a]) - \Pr(O = o|[A \leftarrow a'])$ [13, 28]. Sprenger [28] argues for accepting the Eells measure as a general measure of causal strength, which is in line with the priority that my **Epistemic Condition** attributes to it. Moreover, when restricted to positive values, this is in fact HK's definition of the degree of blameworthiness. Likewise, the obvious counterpart of BvH's condition is to look at the increase of probability in causing the outcome. Thus I also define the *actual causation measure of causal strength* as[6]

$$CS_{ac}(o, a, a') = \Pr(A = a \text{ CNESS-causes } O = o|[A \leftarrow a]) - \Pr(A = a' \text{ CNESS-causes } O = o|[A \leftarrow a']).$$

Taking into account that our **Epistemic Condition** is a mixture of those of BvH and HK, I suggest the following definition, where the value of α expresses the relative importance of both measures.

Definition 11 (Degree of Responsibility). *The* degree of responsibility d *for* $O = o$ *of an agent who performs* $A = a$ *w.r.t. a responsibility setting* $(M, \vec{u}, \mathcal{E})$ *is* 0 *in case the agent is not responsible, otherwise let* $S = \operatorname{argmin}_{a^* \in \mathcal{R}(A)} \Pr(O = o|[A \leftarrow a^*])$*, and let* $a'' = \operatorname{argmin}_{a' \in S} \Pr(A = a'$ *CNESS-causes* $O = o|[A \leftarrow a'])$*, then* $d = CS_e(o, a, a'') + \alpha \cdot max(0, CS_{ac}(o, a, a''))$.

Informally, this measure works as follows. Among all actions that minimize the probability of the outcome, we take one that minimizes the probability of causing the outcome, and then take a weighted sum of both causal strength measures for that action (where the second measure is ignored if it is negative). This captures the idea that in order to avoid responsibility, the agent should choose an action that makes the outcome as unlikely as possible, and then further select their action so that it makes causing the outcome as unlikely as possible. The following example illustrates this definition.

Example 6. *Imagine again our scenario from Example 1, but with the following change:* $Assassin_1$ *is known to be a reliable assassin, whereas* $Assassin_2$ *is known to have second doubts and almost never shoots. In other words, it is reasonable for* $Assassin_2$ *to expect that* $Assassin_1$ *will shoot, and it is reasonable for* $Assassin_1$ *to expect that* $Assassin_2$ *will not shoot. On this particular occasion, both assassins shoot and kill victim.*

Although both assassins are responsible according to my definition, it is easy to see that $Assassin_1$ is responsible to a higher degree: the measures of actual causation are identical for both and so are their respective probabilities of the outcome occurring given that they shoot (namely 1), but $Assassin_1$'s probability of the outcome occurring given that they do not shoot is far lower, and thus[7]

$$CS_e^{Ass_1}(V = 1, A_1 = 1, A_1 = 0) > CS_e^{Ass_2}(V = 1, A_2 = 1, A_2 = 0).$$

Interestingly, recent studies offer empirical confirmation that the agent's epistemic state does indeed impact people's judgments in precisely this way: in a disjunctive scenario (like ours), an agent who performs an action that is *typical* (for them) is considered to be more responsible than an agent who acts *atypically* [18]. The authors contrast this disjunctive scenario, which they have trouble explaining, with a conjunctive one in which both agents' actions are necessary for the outcome to occur, which their account explains quite well. In a conjunctive scenario (in other words, if the equation were $V = A_1 \wedge A_2$), an agent who performs an action that is *atypical* is considered to be more responsible than an agent who acts *typically*, flipping the judgments compared to the disjunctive scenario. That is also the verdict of my degree of blameworthiness: in this case, the atypical agent can reasonably expect the outcome to depend on them performing the action whereas the typical agent can reasonably expect that their action has little impact, which translates into a larger measure of causal strength (both CS_e and CS_{ac}) for the former. So in contrast to the account of Kirfel and Lagnado [18], my proposal applies equally to both scenarios and can thus be seen as a formal extension of their work.

7 Conclusion and Future Work

Based on a comparison with the work of BvH and HK, I have offered a novel formal definition of moral responsibility that is particularly suited for AI systems by filling in the causal and the

[6] Surprisingly, to my knowledge this rather obvious measure of causal strength has been overlooked so far in the literature. (For *any* definition of causation of course, not just CNESS.)

[7] The superscripts Ass_i indicate that we are using each agent's subjective probabilities to assess their degree of responsibility.