

epistemic conditions. I used contrasting examples to argue in favor of the Counterfactual NESS definition of causation over the NESS and the HP definition, and in favor of a nuanced epistemic condition that combines the two conditions of BvH and HK. I connected this work to measures of causal strength to define a degree of responsibility. This quantified approach can be further enhanced by also taking into account the robustness of causation, which recent research suggests plays a role in responsibility judgments that is somewhat independent of causal strength [14], as well as by considering the collective responsibility of groups of agents [6, 8]. Lastly, as discussed, a formal definition of responsibility is a necessary prerequisite for definitions of blame and praise. To develop definitions of the latter requires incorporating harm and benefit [3, 4], and possibly also intention. Therefore the current definition can be extended in several ways, which I aim to do in future work.

## Acknowledgments and Disclosure of Funding

Many thanks to Hein Duijf for helpful comments on an earlier version of this paper, as well as to the Neurips reviewers for their constructive criticism of the original submission. This research was made possible by funding from the Alexander von Humboldt Foundation.

## References

- [1] Beckers S (2021) Causal sufficiency and actual causation. *Journal of Philosophical Logic* 50:1341–1374
- [2] Beckers S (2021) The counterfactual NESS definition of causation. In: *Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pp 6210–6217
- [3] Beckers S, Chockler H, Halpern JY (2022) A causal analysis of harm. In: *Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS-22)*
- [4] Beckers S, Chockler H, Halpern JY (2023) Quantifying harm. In: *Proceedings of the 32 International Joint Conference on Artificial Intelligence (IJCAI-23)*
- [5] Braham M, van Hees M (2012) An anatomy of moral responsibility. *Mind* 121(483):601–634
- [6] Braham M, van Hees M (2018) Voids or fragmentation: Moral responsibility for collective outcomes. *The Economic Journal* 128(612):95–113
- [7] Cooper AF, Moss E, Laufer B, Nissenbaum H (2022) Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FaccT-22)*, pp 864–876
- [8] Dastani M, Yazdanpanah V (2022) Responsibility of ai systems. *AI and Society* (38):843–852
- [9] Duijf H (2023) A logical study of moral responsibility. *Erkenntnis*
- [10] Edmonds D (2015) *Would You Kill the Fat Man?: The Trolley Problem and What Your Answer Tells Us about Right and Wrong*. Princeton University Press
- [11] European Commission (2021) Artificial intelligence act. URL <https://artificialintelligenceact.eu/the-act/>
- [12] Fischer JM, Ravizza M (1998) *Responsibility and Control*. Cambridge University Press
- [13] Fitelson B, Hitchcock C (2011) Probabilistic measures of causal strength. In: Illari RF P M, Williamson J (eds) *Causality in the Sciences*, Oxford University Press, pp 600–627
- [14] Grinfeld G, Lagnado D, Gerstenberg T, Woodward JF, Usher M (2020) Causal responsibility and robust causation. *Frontiers in Psychology*
- [15] Halpern J, Pearl J (2005) Causes and explanations: A structural-model approach. part I: Causes. *The British Journal for the Philosophy of Science* 56(4):843–87
- [16] Halpern JY (2016) *Actual Causality*. MIT Press

- [17] Halpern JY, Kleiman-Weiner M (2018) Towards formal definitions of blameworthiness, intention, and moral responsibility. In: Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), pp 1853–1860
- [18] Kirfel L, Lagnado D (2021) Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition* 212
- [19] Kroll JA (2020) Accountability in computer systems. In: Markus D, Dubber FP, Das S (eds) *Oxford Handbook of the Ethics of AI*, Oxford University Press
- [20] McIntyre A (2018) Doctrine of double effect. *The Stanford Encyclopedia of Philosophy*
- [21] Moore MS (2009) *Causation and Responsibility*. Oxford University Press, Oxford, U.K.
- [22] Pearl J (2000) *Causality: Models, Reasoning, and Inference*. Cambridge University Press
- [23] Peters J, Janzing D, Schölkopf B (2017) *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press
- [24] Rosenberg I, Glymour C (2018) Review of joseph halpern, actual causality. *BJPS Review of Books*
- [25] Rudy-Hiller F (2022) The epistemic condition for moral responsibility. *The Stanford Encyclopedia of Philosophy*
- [26] Sartorio C (2007) Causation and responsibility. *Philosophy Compass* 2(5):749–765
- [27] Sartorio C (2017) Ignorance, alternative possibilities, and the epistemic conditions for responsibility. In: Peels R (ed) *Perspectives on Ignorance from Moral and Social Philosophy*, New York: Routledge
- [28] Sprenger J (2018) Foundations for a probabilistic theory of causal strength. *The Philosophical Review* pp 371–398
- [29] Vallentyne P (2008) Brute luck and responsibility. *Politics, Philosophy and Economics* 7:57–80
- [30] Weslake B (2015) A partial theory of actual causation. *The British Journal for the Philosophy of Science* forthcoming
- [31] Woodward J (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press
- [32] Wright RW (1985) Causation in tort law. *California Law Review* 73:1735–1828
- [33] Wright RW (2011) The ness account of natural causation: A response to criticisms. In: Goldberg R (ed) *Perspectives on Causation*, Hart Publishing

## Appendix

### Frankfurt-Case

The following is an example of a so-called “Frankfurt-case”, taken from HK. An enormous literature in philosophy is devoted to dealing with these kinds of examples, attempting to reconcile intuitions about responsibility with the counterfactual and causal features that these examples contain. Surprisingly, almost none of it uses causal models, and yet doing so reveals the causal structure to be entirely unproblematic.

**Example 7 (Frankfurt).** *Imagine Jones poisons Smith, who dies. Unbeknownst to Jones, Black was observing his behavior: if Jones had not poisoned Smith, Black would have given Jones a gun and manipulated him in some way or other so that Jones would shoot Smith. Black is both a perfect observer and manipulator of Jones’s behavior, and is thus guaranteed to succeed in his plans. Intuitively it is clear that Jones is responsible for Smith’s death, despite the fact that he could not have prevented it. (Typical Frankfurt cases focus on responsibility for an action, as opposed to responsibility for the consequence of an action, and therefore scenarios are normally formulated such that Black manipulates Jones to perform the same action. Except for the shift from the action to the consequence though, those cases are structurally isomorphic.)*<sup>8</sup>

The **Epistemic Condition** of both BvH and HK is obviously fulfilled, for Jones believes that Smith’s death is completely dependent on his poisoning. We consider the following equations to assess the causal condition:  $SD = JP \vee JS$  to capture the fact that Smith dies ( $SD$ ) if either Jones shoots ( $JS$ ) or poisons ( $JP$ ) him,  $JS = BM$  to capture that Jones shoots only when Black hands him a gun and manipulates him ( $BM$ ), and finally  $BM = \neg JP$  to capture that Black’s action depends on Jones’s failure to poison.

Regardless of whether we apply the NESS definition, the CNESS definition, or the HP definition,  $JP = 1$  comes out as a cause of  $SD = 1$ , and thus the **Causal Condition** is satisfied. (This is easy to see by observing that the structure of this example is a standard case of Early Preemption.)

BvH claim that their account can handle Frankfurt-cases like this, but that is a mistake. Recall that their variables represent the agents’ strategies rather than their actions, and that we are limited to using a single equation. The outcome function they use when discussing a Frankfurt-case is equivalent to the equation  $SD = JP \vee B$ , where  $B$  represents Black adopting his preferred strategy. Therefore on their account both Jones and Black come out as causes of Smith’s death, which is not a sensible result. BvH admit that their NESS definition is unable to handle conditional strategies like that of Black, but contend that since we are here focussing on Jones this is not a problem. Obviously simply stating that one should only focus on the sensible results of one’s theory is not a satisfactory way of defending it... (This example also highlights a more philosophical problem with their approach: it is not at all clear what it means for a strategy to be a cause. The broad consensus is that causal relata are either events/omissions or properties of events, whereas conditional strategies are neither.)

### Counterexample to the HP-definition

We here consider a second counterexample to the HP definition that was suggested in [24]. The example is of particular interest as it was presented precisely within the context of the relation between causation and moral responsibility.

**Example 8.** *We have equations  $Y = X \vee D$  and  $X = D$ , and we consider a context such that  $D = 1$ . This looks very much like a standard case of overdetermination in which  $X = 1$  and  $D = 1$  are both overdetermining causes. Yet  $X = 1$  is not an HP-cause of  $Y = 1$  (and it is a CNESS-cause). The reason for this is that  $Y = 1$  depends counterfactually on  $D = 1$  by itself, whereas it does not depend on  $X = 1$  by itself and nor does it when we take  $D = 1$  as our witness  $\vec{W} = \vec{w}$ . Rosenberg & Glymour [24] argue that this result shows the HP definition cannot offer a basis for moral responsibility, by offering the following scenario to go along with these equations:*

---

<sup>8</sup>This analysis can just as easily be applied to these more typical Frankfurt cases. Still, for those who are sceptical that proponents of Frankfurt cases are equally comfortable as I am with moving from actions to consequences, I point out that Fischer & Ravizza apply this shift in exactly the same manner as I do when discussing responsibility for consequences [12, ch. 4].

*An obedient gang is ordered by its leader to join him in murdering someone, and does so, all of them shooting the victim at the same time, or all of them together pushing the plunger connected to a bomb. The action of any one of the gang would suffice for the victim's death. If responsibility implies causality, whom among them is responsible? ... Halpern's theory says the gang leader and only the gang leader is a cause of the victim's death. This is a morally intolerable result; absent a plausible general principle severing responsibility from causation, any theory that yields such a result should be rejected.*

## Bombing

We now go through the details for the **Bombing** example. (Ex. 6) We need to consider the following four scenarios:

1.  $S_2 = 1$  and  $S_1 = 0$
2.  $S_2 = 1$  and  $S_1 = 1$
3.  $S_2 = 0$  and  $S_1 = 0$
4.  $S_2 = 0$  and  $S_1 = 1$

We first go through the details for CNESS-causation.

In scenario 1 we have that  $S_1 = D_1 = D_3 = 0$  and  $S_2 = D_2 = B = 1$ . Here  $\{S_2 = 1, S_1 = 0\}$  is sufficient for  $D_2$ , whereas  $\{S_1 = 0\}$  is not. Therefore  $S_2 = 1$  directly NESS-causes  $D_2 = 1$ . Clearly also  $D_2 = 1$  directly NESS-causes  $B = 1$ , and thus  $S_2 = 1$  NESS-causes  $B = 1$  along  $\{S_2, D_2, B\}$ . What about the counterfactual setting  $(M_{S_2 \leftarrow 0}, \vec{u})$ ? That corresponds to scenario 3. There, the bomb doesn't even explode (so  $B = 0$ ), and thus there are no causes of  $B = 1$ . We conclude that in scenario 1  $S_2 = 1$  CNESS-causes  $B = 1$ .

In scenario 2 we have that  $S_1 = S_2 = D_3 = B = 1$  and  $D_1 = D_2 = 0$ . In this scenario  $B = 1$  is directly NESS-caused only by  $D_3 = 1$ . Since  $S_2 = 1$  does not directly NESS-cause  $D_3 = 1$ , it is not a NESS-cause of  $B = 1$ .

In scenario 4 we have that  $S_1 = D_1 = D_3 = B = 1$  and  $S_2 = D_2 = 0$ . Here  $\{S_2 = 0, S_1 = 1\}$  is sufficient for  $D_1$ , whereas  $\{S_1 = 1\}$  is not. Therefore  $S_2 = 0$  directly NESS-causes  $D_1 = 1$ . Clearly also  $D_1 = 1$  directly NESS-causes  $B = 1$ , and thus  $S_2 = 0$  NESS-causes  $B = 1$  along  $\{S_2, D_1, B\}$ . What about the counterfactual setting  $(M_{S_2 \leftarrow 1}, \vec{u})$ ? That corresponds to scenario 2, in which  $S_2 = 1$  does not NESS-cause  $B = 1$ . So  $S_2 = 0$  CNESS-causes  $B = 1$  in scenario 4.

As a result, if *Assassin<sub>2</sub>* chooses  $S_2 = 1$ , the probability of CNESS-causing  $B = 1$  is the probability that  $S_1 = 0$ , which is 0.4. By contrast, if *Assassin<sub>2</sub>* chooses  $S_2 = 0$ , the probability of CNESS-causing  $B = 1$  is the probability that  $S_1 = 1$ , which is 0.6.

NESS-causation for each scenario is already discussed in the above, so we move on to consider HP-causation. In scenario 1 we have counterfactual dependence of  $B = 1$  on  $S_2 = 1$ , and it is well-known that this suffices for HP-causation (as well as for CNESS-causation, by the way [2]).

In scenario 2, note that  $D_3$  suffices for  $B = 1$ , and thus satisfying AC2 is possible only when either  $D_3 = 1$  or  $S_1 = 1$  is also part of the candidate cause  $\vec{X} = \vec{x}$ . However,  $B = 1$  counterfactually depends on  $D_3 = 1$ , meaning that  $D_3 = 1$  is a cause all by itself. Thus  $\{S_2 = 1, D_3 = 1\}$  is not minimal, and because of AC3 this means that it is not a cause. That leaves  $\{S_2 = 1, S_1 = 1\}$ . But this is not minimal either, for  $S_1 = 1$  is a cause all by itself: one can take  $\vec{W} = \{D_2\}$  as a witness to get  $B = 0$  when  $S_1$  is set to 0. Therefore  $S_2 = 1$  is not part of any cause of  $B = 1$ .

Since  $B = 0$  in scenario 3,  $S_2 = 0$  does not HP-cause  $B = 1$  there either, leaving scenario 4. As with scenario 2, the candidate cause will have to include  $D_3 = 1$  or  $S_1 = 1$ . Contrary to scenario 2 though,  $D_3 = 1$  is no longer a cause by itself, since  $D_1 = 1$  holds, and will remain to hold also when we set  $D_3$  to 0. Since  $B = 1$  counterfactually depends on  $\{S_2 = 0, D_3 = 1\}$ , we get that each of them HP-causes  $B = 1$ .