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