Outcome effects, moral luck and the hindsight bias

Markus Kneer (University of Zurich)
Izabela Skoczeń (Jagiellonian University)

John and Sally drive to work. They are well-rested, alert and stick to the speed limit. A child jumps in front of John's car and dies, Sally arrives at work without incident. Who is more to blame? In between-subjects designs, a pronounced outcome effect tends to arise: John is judged morally and legally more culpable than Sally (henceforth the Outcome Effect). This might strike us as unjust, if we hold, with Kant (1978), that agents are morally responsible only for features of their actions over which they have control (the Control Principle).

Philosophers assume that a difference in moral judgment arises even within-subjects, i.e. when people directly compare John's and Sally's case (the Difference Intuition). This would give rise to the Problem of Resultant Moral Luck: We must square the consequentialist Difference Intuition with the Kantian Control Principle, but the two are fundamentally inconsistent. However, Folk Morality disagrees: When presented with John and Sally's cases side by side, the vast majority of participants evaluate the two agents identically. Western Criminal Law, with its deep distaste for strict liability, sides with the Folk in this regard. So there might not be a complex philosophical problem (the within-subjects Difference Intuition is simply an oddity of philosophers hunting for a paradox). However, in everyday life, we are not confronted with two neat cases side-by-side. Usually, we assess situations where a concrete harm has occurred and here outcome is likely to have distorting effect on our judgment, violating the Control Principle to which both the Law and the Folk are committed.

How can we alleviate the outcome effect? There is evidence in favour of a probabilistic account of moral luck-type phenomena (Kamin & Rachlinski, 1995; Kneer & Machery, 2019). On this account, the post-hoc probability of harming a child is perceived higher for John than for Sally. It thus seems more appropriate to judge that John incurred a substantial risk than that Sally did, which, in turn would mean he was more reckless or negligent than Sally. If this account is on the right track, then a perceived difference in probability and risk drives an asymmetry of risk-related inculpating mental states and hence moral (and legal) evaluation. The whole series of inferences from descriptive features to normative evaluation is innocuous, except for the first step, which is affected by the hindsight bias: in John's case, people tend to exaggerate the degree to which a harmful outcome could, or should, have been anticipated (Fischhoff, 1975; 1980). To address the distorting effect of outcome on culpability judgments, this suggests, we must find ways to alleviate the hindsight bias.

We first explore whether the probabilistic account of the effect of outcome on culpability replicates. Our experiments (total N = **2043**) are the first to control explicitly for the distinction between objective probability (probability from the perspective of the universe) and subjective probability (as perceived from the agent's context). Having replicated the outcome effect on probability, mens rea and moral judgment, we show that it must be considered a bias. The effect of outcome is much more pronounced in between-subjects designs than in within-subjects designs. Next, we turn to debiasing strategies: first, probability anchoring. We test whether giving participants the possibility to evaluate the likelihood of a harmful outcome before the consequences are revealed has an impact on their probability assessments ex post. Next, counterfactual priming: we investigate whether entertaining alternative outcomes reduces the outcome effect on probability, mens rea and moral judgments. Finally, probability stabilizing, in which an expert provides the actual ex ante probability of a harmful outcome from the point of view of a scientifically informed perspective. Probability anchoring and counterfactual priming attempt to prevent inappropriate inferences from outcome information to probability ex post in indirect fashion. By contrast, probability stabilizing makes short shrift of the problem by directly stipulating the probability ex post so as to prevent inadequate downstream consequences on mens rea and culpability assessment. Consistent with previous research, the effects of outcome on probability post hoc and downstream variables such as mens rea and culpability are persistent and robust across experiments with different scenarios. These effects are the results of a cognitive bias (though not for punishment judgments). Neither strategy fully eradicates inappropriate inferences from outcome to probability and distorted downstream effects on mens rea and culpability judgments thus remain. What works best is probability stabilizing, which is indeed a means courts all too frequently do not resort to.

Selected References:

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Kamin, K. A., & Rachlinski, J. J. (1995). Ex post ≠ ex ante: Determining liability in hindsight. *Law and Human Behavior*, 19(1), 89–104. https://doi.org/10.1007/BF01499075

Kant, I. (2009). *Critique of pure reason* (15th reprint). Cambridge University Press.

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, 182, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003