Penultimate Draft. Published Version Available at
AJOB Neuroscience:
https://www.tandfonline.com/doi/full/10.1080/21507740.2019.1705428

BCI-Mediated Behavior, Moral Luck, and Punishment

Daniel J. Miller

Abstract: An ongoing debate in the philosophy of action concerns the prevalence of *moral luck*: instances in which an agent's moral responsibility is due, at least in part, to factors beyond his control. I point to a unique problem of moral luck for agents who depend upon Brain Computer Interfaces (BCIs) for bodily movement. BCIs may misrecognize a voluntarily formed distal intention (e.g., a plan to commit some illicit act in the *future*) as a control command to perform some overt behavior *now*. If so, then BCI-agents may be deserving of punishment for the unlucky but foreseeable outcomes of their voluntarily formed plans, whereas standard counterparts who abandon their plans are not. However, it seems that the only relevant difference between BCI-agents and their standard counterparts is just a matter of luck. I briefly sketch different solutions that attempt to avoid this type of moral luck, while remaining agnostic on whether any succeeds. If none of these solutions succeeds, then there may be a unique type of moral luck that is unavoidable with respect to deserving punishment for certain BCI-mediated behaviors.

Keywords: blameworthiness; moral luck; punishment; brain computer interfaces

An ongoing debate in the philosophy of action concerns the prevalence of *moral luck*: instances in which an agent's moral responsibility is due, at least in part, to factors beyond his control. Consider a purported example of *outcome luck*: one assassin succeeds in shooting and killing his target, but a qualitatively identical assassin misses simply because a bird flies in the path of his bullet (Nagel 1979). The only relevant difference between the two lies in factors beyond their control. Though some have the intuition that the successful assassin is more blameworthy than the unsuccessful assassin (since he

killed someone), it strikes others as unfair to judge him as more blameworthy for a difference in outcomes that is completely due to luck.

Rainey, Maslen, and Savulescu seem intent on avoiding a kind of outcome luck for agents who depend upon Brain Computer Interfaces (BCIs) for bodily movement. One concern raised is that BCI devices may misrecognize certain *involuntary* mental events (e.g., involuntarily imagining performing some action) as control commands to perform overt behaviors. It seems unfair to judge that BCI-agents are blameworthy for outcomes of involuntary mental events.

Yet, the concerns with moral luck raised by BCI-mediated behavior extend beyond the authors' observation that *involuntary* mental events might result in overt behavior. A distinct problem of moral luck arises from the possibility that BCIs may misrecognize a *voluntarily* formed distal intention (e.g., a plan to commit some illicit act in the *future*) as a control command to perform some overt behavior *now*. Consider:

> *ABANDONED PLAN*: Typical Terry stands with his business rival, Gary, on top of a skyscraper (their usual meeting place for arranging high-stakes business deals). Terry judges that it would be advantageous to get rid of Gary. Inspired by the view, Terry voluntarily forms a plan to shove Gary off of the skyscraper on some future occasion (when fewer people are around). Moments later, however, Terry has a crisis of conscience and consequently abandons his plan.

> *PREMATURE EXECUTION*: BCI-Barry is just like Typical Terry except that he depends upon a BCI to control his prosthetic arms. Through training, Barry has become aware that BCIs can misrecognize distal intentions to Φ *later* as control commands to Φ *now*. When Barry voluntarily forms a plan to shove Gary off of the skyscraper on some future occasion, his BCI

misrecognizes this as a control command, resulting in Barry's prosthetic arms shoving Gary off of the skyscraper to his death.

Although Barry did not intend to kill Gary *then*, Gary's subsequent death was nevertheless a foreseeable result of voluntarily forming his plan. Consequently, Barry is plausibly blameworthy for Gary's death.[1] So, although both agents are (we can assume) blameworthy for their voluntarily formed plans to kill Gary, Barry seems more blameworthy than Terry. But suppose that the following counterfactual is true: were it not for the BCI's misrecognition, Barry also would have had a crisis of conscience moments later and consequently abandoned his plan. If so, then it seems the only difference between Barry's plan and Terry's plan (such that the former resulted in Gary's death and the latter did not) is a matter of luck. Given this, the intuition that Barry is more blameworthy than Terry seems unjustifiable.

In response, one might suggest the following solution: While Barry is blameworthy for more *items* than Terry (i.e., the voluntarily formed plan *and* Gary's death), this does not entail that Barry is any *more* blameworthy than Terry: we can distinguish between the *scope* of an agent's blameworthiness and the *degree* to which the agent is blameworthy (Zimmerman 2002, 560).

While this solution may successfully avoid outcome luck concerning blameworthiness, the scope-degree distinction does not apply as neatly to concerns about how (in such cases) the law should punish criminal behavior mistakenly mediated by BCIs. While the movement of Barry's prosthetic arms mistakenly mediated by the BCI

---

[1] This is consistent with the claim that Barry is not *as* responsible for Gary's death as he would be if the movement of his arms were the result of an intention to do so *then*, or if Barry actually foresaw at the time that forming his distal intention might have the specific result that it did (Miller 2019).

does not constitute an *action* (since it's not appropriately causally related to his distal intention), it is comparable to the operation of a machine that is known to occasionally malfunction. The case in question is therefore analogous to involuntary manslaughter, where someone's death is an unintended result of an intentional action (in Barry's case, the mental action of voluntarily forming his plan). Given these considerations, BCI-agents like Barry seem deserving of punishment for the unlucky but foreseeable outcomes of their voluntarily formed plans, whereas standard counterparts (like Terry) who abandon their plans clearly cannot be deserving of punishment for outcomes that never come about.

But again, the difference between such agents seems just a matter of luck. If Terry is not deserving of punishment, how is it fair to judge that Barry is? Here I sketch some different solutions that attempt to avoid this type of moral luck, while remaining agnostic on whether any succeeds.[2]

Solution 1: One might *deny* that the only difference between Barry and Terry is just a matter of luck. Since Barry depends upon a BCI for bodily movement, he has a special obligation to be careful in forming *any* intentions (even intentions to perform actions at later times), since he knows that his BCI device might misrecognize such intentions as control commands to perform overt behaviors *then and there*. Because of this, Gary's being shoved *then* was a foreseeable (and thus avoidable) outcome of Barry's plan, whereas this was not so for Terry's plan.

One may object that *Solution 1* avoids outcome luck only at the cost of embracing a different kind of moral luck: *constitutive luck* (i.e., luck in how an agent is constituted).

---

[2] Whether outcome luck is acceptable concerning punishment more generally is contested (Hart 1968, Lewis 1989). I take my focus here to be a unique problem of outcome luck for BCI-agents.

We may presume that it is just a matter of bad luck that Barry requires a BCI and consequently has special obligations to be cautious that standard agents do not have. In response, the proponent of *Solution 1* may point out that *this* sort of luck is ubiquitous: even standard agents commonly have special obligations due to constitutive luck (e.g., I may have an obligation to take extra precautions around fragile items because I happen to be particularly clumsy). Even so, there may be something relevantly different about placing special expectations on agents who depend upon BCIs for bodily movement: it seems unreasonable to expect BCI-agents to be as careful as they would need to be in order to avoid the sort of outcomes in question. Indeed, it would require an enormous shift in the control we normally exercise over our mental lives to monitor each plan we form as though it might be translated into overt behavior then and there.

Solution 2: One might *agree* that that the only difference between Barry and Terry is just a matter of luck, and maintain that BCI-agents are no more deserving of punishment than their standard counterparts. This leaves it open whether the two agents deserve *any* punishment.

> Solution 2a: One might maintain that *both* agents deserve punishment. If so, however, we would want to know in virtue of what Terry deserves punishment. Terry might be *blameworthy* for his voluntarily formed plan, but it is implausible that he deserves punishment for a purely mental action. Nor does it seem reasonable for the law to

treat such fleeting intentions (supposing they were discovered) as *attempts* to commit crimes, especially since agents like Terry never take any steps to actualize their plans.[3]

Solution 2b: One might maintain that *neither* agent deserves punishment. This entails that Barry does not deserve punishment for Gary's death. Someone who opts for *Solution 2b* could support this contention either by maintaining that (i) Barry is not even blameworthy for Gary's death (something plausibly required for deserving punishment), or that (ii) Barry is blameworthy for Gary's death, but this fact is insufficient to make him deserving of punishment for it. (i) is controversial, since it is widely held that agents are blameworthy for foreseeable outcomes of earlier items for which they are blameworthy (Rosen 2008, 604; Fischer and Tognazzini 2009, 537-538).[4] And a defense of (ii) would require an explanation of why being blameworthy for someone's death is insufficient for deserving punishment.

If none of these solutions succeeds, then there may be a unique type of moral luck that is unavoidable with respect to deserving punishment for certain BCI-mediated behaviors.

Acknowledgements: The author is grateful to Gabriel De Marco, Randy Clarke, and Kyle Fritz for helpful feedback on an earlier draft of this paper.

References

[3] In contrast, it is reasonable to judge the unsuccessful assassin as deserving punishment for *attempted* murder. This highlights one way that the problem of outcome luck I raise for BCI-agents is unique.

[4] For an argument that (given other plausible commitments) foreseeability is insufficient as an epistemic condition on responsibility for outcomes, see Miller 2017.

Miller, D. J. 2017. Reasonable foreseeability and blameless ignorance. Philosophical Studies 174(6): 1561–1581.
Miller, D. J. 2019. Circumstantial ignorance and mitigated blameworthiness. Philosophical Explorations 22(1): 33–43.
Fischer, J, and N. Tognazzini. 2009. The truth about tracing. *Noûs* 43(3):531-556.
Hart, H. L. A. 1968. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford University Press.
Lewis, D. 1989. The Punishment That Leaves Something to Chance. *Philosophy and Public Affairs* 96:227–242.
Nagel, T. 1979. Moral Luck. In T. Nagel, *Mortal Questions.* Cambridge: Cambridge University Press.
Rainey, S., H. Maslen, and J. Savulescu. (n.d.) When thinking is doing: Responsibility for BCI-mediated action requires special attention in terms of controllability and foreseeability of outcomes. Forthcoming in *AJOB Neuroscience*.
Rosen, G. 2008. Kleinbart the oblivious and other tales of ignorance and responsibility. *Journal of Philosophy*, 105(10):591–610.
Zimmerman, M. 2002. Taking luck seriously. *The Journal of Philosophy* 99(11):553-576.