Neurofeedback for moral enhancement: irreversibility, freedom, and advantages over drugs

Akabayashi and colleagues examine the potential therapeutic applications of decoded neurofeedback in treating psychiatric conditions such as depression, and developmental disorders. Decoded neurofeedback, they argue, is particularly promising in this regard, since it can enable individuals to observe a representation of their brain activity in real time. Consequently, individuals can train themselves to intentionally adjust their brain activity, ultimately in the absence of the visual representation. Akabayashi et al. further hypothesize that decoded neurofeedback techniques could be used for moral enhancement, if individuals were able to train themselves to adjust their brain states to those conducive to moral behavior, such as the brain state correlated with compassion. This, they argue, would be a particularly appealing form of moral enhancement, since modulation of brain states could be 'personalized, or tailor made' to the individual's beliefs about how to live morally.

The paper makes an important contribution in drawing the attention of neuroethicists to the prospect that neurofeedback could enable a greater degree of control over mental phenomena such as emotions or strong desires, which sometimes frustrate our ability to act in line with our moral commitments or prudentially. Although Akabayashi et al. are optimistic about the way in which the personalization of neurofeedback preserves moral pluralism, they suggest that there are significant ethical concerns relating to irreversibility, safety and efficacy. Irreversibility in particular, they argue, is potentially problematic in its implications for freedom, since agents are rendered unable to alter themselves if their moral beliefs change. This, they argue, is in contrast to pharmaceuticals, the effects of which are reversed once they are metabolized (except, perhaps, in the case of addictive substances). Whilst we agree that adverse side effects and efficacy are clearly relevant to the ethical assessment of neurofeedback for the full range of potential purposes, we challenge the authors' claim that the irreversibility of the effects of neurofeedback necessarily presents a problem for its use for moral enhancement. Indeed, we argue that irreversibility could present no problem for the agent's freedom. Indeed, it could promote it. Further, through defending this claim, we show that the particular way in which neurofeedback supports the agent's freedom (by increasing cognitive control) gives us a reason to prefer it to some plausible forms of pharmaceutical moral enhancement.

Akabayashi et al. claim that, "in contrast [to pharmaceuticals], if the alterations of neural networks induced by decoded neurofeedback-based moral neuroenhancement were not or hardly reversible, then this technique would threaten the human freedom of moral development." We take their claim not to be an assertion that freedom from technological intervention is necessary in order for moral development to be truly *moral*. If this were the case, then neurofeedback interventions – or more precisely their effects – would never be candidates for instances of moral enhancement irrespective of reversibility. Rather, the concern seems more plausibly to be that the agent might lack the option to undo the effects of neurofeedback in the event that her beliefs about morality changed. Reversibility, the argument goes, gives us the ability to train and detrain ourselves, as we revise our beliefs and develop morally. In their words: "Reversibility can save our freedom in the sense that can allow us to alter ourselves later again to a certain extent despite being equipped with a particular characteristic already." Pharmaceuticals taken in pursuit of moral enhancement, Akabayashi et al. argue, allow for such reversibility since they are metabolized and thus their effects wear off. We take this to imply, for example, that if an agent revises her beliefs and now considers herself to have evidence that that too much compassion is morally bad (or not morally good), then she can stop taking the compassion drug, or take less of it. Neurofeedback, if irreversible, does not have this quality, the argument goes.

We argue that (if in fact present) the irreversibility of the training effects of neurofeedback is, not problematic for the agent's freedom of moral development. Indeed, if the technology works as described, it can promote the agent's freedom through giving her more control. Akabayashi et al. provide the example of moral enhancement via control of positive or negative emotions. Examples might include training oneself to increase feelings of compassion at will or decrease feelings of aggression at will. Having the capacity to cultivate the brain activity correlated with these emotions, it is assumed, might, for example, allow one to care better for one's sick friend, or remain calm and collected in a hostile disagreement. Further possibilities for moral enhancement might be found in the potential for controlling brain states associated with strong desires that move one to act contrary to one's moral convictions. Control over a desire to eat meat might be morally valuable to a vegetarian; control over a desire to cheat on one's partner might be morally valuable to the monogamist.

So, what if the agent's beliefs about morality change? What if the agent perceived her low aggression to result in a regrettable failure to be sufficiently moved to act against injustice? What if the vegetarian ceased to be persuaded by the arguments for vegetarianism or the monogamist ceased to find any moral value in monogamy? Would the irreversibility of the effects of neurofeedback training restrict the agent's freedom in the way Akabayashi suggest? The answer might be in the affirmative if the intervention were to impose on the agent a particular brain state. We might here draw comparisons with environmental interventions that can sometimes have this effect, such as brainwashing or certain forms of religious indoctrination. Indeed, in the present context, such permanence would seem to imply that the agent would have no or little choice but, for example, to feel no aggression, even if this no longer helped her to do what she believed morality required.

However, neurofeedback training appears to be characterized by increased capacity for self-regulation and control. Indeed, such effects seem to be closely tied to the goal in using it for moral enhancement. Far from imposing one brain state on the agent, neurofeedback training allows the agent to control her brain activity. Akabayashi et al.'s therapeutic example of chronic pain patients being asked to make the flame on the screen larger and smaller to control the activity of their anterior cingulate cortex (and hence the intensity of their pain) demonstrates this principle. If agents use neurofeedback to train themselves to be able to dial up and dial down a particular emotion or desire, then irreversibility poses no problem. This is so even if the agent revises her beliefs about morality such that a particular degree of emotion or strength of desire ceases to move her to act as she believes she should. Indeed, the ability to control brain states should in fact enable her to also avoid (now) unhelpful degrees of emotion or strengths of desires. So, if neurofeedback training facilitates greater control over a state, in line with the therapeutic examples discussed by Akabayashi and colleagues, rather than conditioning the permanence of a particular state, irreversibility does not present a problem for the freedom of moral development. Rather, greater ability to regulate one's brain states can only increase the agent's freedom, especially given the way in which strong emotions and desires often frustrate our autonomy when we fail to prevent them influencing our behavior. Just as one might decide that one is too aggressive, and retrain the brain, so too one might decide one is too passive, and retrain the brain.

The contrast between neurofeedback and pharmaceuticals in terms of their *potential irreversibility*, then, does not illuminate a problem with the former regarding the agent's freedom to revise her moral beliefs and behavior. However, we suggest that the contrast between neurofeedback and some pharmaceuticals in terms of the agent's *cognitive control* reveals that neurofeedback in some cases presents a superior route to moral enhancement. In the above discussion, we argued that the ability to regulate one's brain states at will, in line with one's beliefs about what one should do promotes the agent's autonomy. This ability to intentionally align one's emotions and desires with what one believes one has reason to do constitutes an ideal moral enhancement. Candidate moral enhancers often face the objection that any positive effects on behavior are not 'rationally mediated' or 'bypass rational deliberation', and thus are not true moral enhancements. A drug that imposed on the agent inescapable urges to act compassionately might constitute such an example, especially if these feelings were unaffected by whether the particular situation called for compassion and resulting compassionate action. On this view, even compassionate acts performed in the presence of justifying moral reasons would, *ex hypothesi*, result from the drug rather than the agent, in a way that devalued their moral praiseworthiness.

For example, John Harris (2012: 172) has argued that 'moral enhancement, properly so called, must not only make the doing of good or right actions more probable and the doing of bad ones less likely, but must also include the understanding of what constitutes right and wrong action…if, "once the enhancement has been initiated, there is no further need for cognition", then the morally enhanced action is effectively automatic, unconscious and therefore unintended, entirely outside the realm of moral responsibility'. Whilst we do not wish to argue that pharmaceuticals could never be moral enhancers (an agent can respond to reasons to take a drug that she believes will force her behave more morally and continue to feel the weight of those reasons under its direct effects; or, a drug might in fact render an agent more responsive to moral reasons, despite the primary effect of the drug involving no rational mediation), an intervention that trains the agent to integrate and precisely celebrate its influence internally would make the moral enhancement effects dependent on rational deliberation, in the way that some moral enhancement skeptics claim is not possible in the case of some pharmaceuticals. Indeed, exercise of the effects of neurofeedback training will partly constitute rational deliberation.

Thus, holding adverse side effects and effectiveness constant, neurofeedback seems to give us prospects for moral enhancement more promising than some drugs; due to the degree of cognitive integration facilitated, and the potential for finer-grained control and calibration by the agent.

A possible objection might be that, in addition to or instead of neurofeedback engendering greater cognitive control over emotions and desires (with implications for moral behavior), learning to control brain states could have irreversible effects on the agent's moral beliefs, making it difficult or impossible to decide ex post to implement changes. For example, inculcating more compassionate brain states might make one more receptive to facts about the suffering of others such that one cannot now ignore the moral weight of this suffering: one's beliefs about how one should treat people, and about which brain states one should continue to cultivate are in this sense irreversible. However, this does not seem to be a 'permanent' change distinct in kind from more traditional moral education: witnessing, in a particularly vivid and engaging way, others' suffering might have the same effect on empathy or compassion, and it might be equally irreversible. Further, the plausible idea that individuals can make moral progress relies on there being reasons to behave in certain ways to which one can become more responsive. Becoming more aware of these reasons, as perhaps one does through increasing compassion or reducing spontaneous aggression, should not necessarily be viewed as problematic for an agent's freedom, even if the force of these reasons then endures. As long as an intervention does not distort or hijack rational evaluation, persisting downstream changes in moral beliefs do not constitute a problem per se; indeed, the very notion of moral development depends on reassessing and revising beliefs, some of which the agent may end up holding firmly.

Irreversibility of increased cognitive control, we have argued, does not necessarily pose an ethical challenge; permanence of a particular brain state, or unmediated changes to beliefs may be more problematic from the perspective of 'freedom of moral development', but the evidence Akabayashi and colleagues present does not demonstrate neurofeeback having such effects. Moreover, indirect changes to beliefs, resulting from the agent's greater exposure to reasons, constitutes archetypal moral development. Thus, we conclude the mechanism of action of neurofeedback is either more advantageous or at least no more disadvantageous than many other proposed mechanisms of moral enhancement.

References

Harris, J. (2012), 'Ethics is for Bad Guys! Putting the "Moral" into Moral Enhancement', *Bioethics*, 27 (3): 169–173.