Culpability, control, and brain-computer interfaces*

Charles Rathkopf†

May 11, 2023

Abstract

When actions are mediated by means of a brain-computer interface, it seems that we cannot assess whether the user is culpable for the action without determining whether the brain-computer interface correctly decoded the intentions of the user. Here I argue that this requirement is confused. I also argue that, at least for the purposes of assessing moral culpability, BCI-mediated action should be viewed on the model of action mediated by ordinary (albeit complex) tools.[1]

*This is a draft for a chapter that will appear in a forthcoming book entitled *Neuroprosthethics*, edited by Jan-Hendrik Heinrichs and Orsolya Friedrich.

†Institute for Brain and Behavior, Jülich Research Center

[1]Thanks to Niel Conradie and Jared Parmer for helpful comments on an earlier draft, and especially to Jan-Hendrik Heinrichs for multiple long and fruitful discussions of the material in this chapter.

Contents

1 Ruling out malfunction

Imagine that you are sitting at a faculty meeting, listening to one of your colleagues chatter on about administrative duties. Another colleague, who is known for her impatience with chatter, suddenly picks up her glass of water and throws it in the chatterer's face. After passions have cooled, she says: "I don't know what happened. It was as if I had *no control* over my own arm." If we know that the thrower does not suffer from mental illness, and that she has a track record of responsible adult behavior, most of us would say that she is blameworthy for her action. According to a standard analysis of blameworthiness, which is accepted both by folk intuition and by philosophical theory, a person cannot be blameworthy for an action unless the action was subject to voluntary control. If we accept that analysis, and we accept that the thrower is blameworthy, we must also accept that the thrower had been in control of her arm, despite her claim to the contrary.

Now imagine a variant of this scenario in which the throw is executed by a robotic arm, which is itself governed by an intra-cortical brain-computer interface (BCI). In this variant, the claim that "It was as if I had no control over my own arm" is not so easy to dismiss. In order to make a defensible judgement that the thrower had been in control, it *seems* as if we need to rule out the possibility that the BCI had malfunctioned. One kind of malfunction that seems particularly relevant to ethical questions involves the *misinterpretation* or *inaccurate decoding* of neural activity. Consider, for example, a case in which the thrower idly *imagines* throwing the glass of water in the chatterer's face, but has no intention to carry out the throw. Nevertheless, the act of imagination corresponds to a distinctive pattern of neural activity. The BCI registers that pattern, incorrectly interprets the pattern as encoding an intention to throw, and then executes the throw. In this scenario, it looks as if the thrower is blameless.

How can we tell whether a pattern of neural activity was interpreted correctly? That is an epistemic question about BCI-mediated action. There is also a more fundamental metaphysical question: what *is it* for a BCI to interpret patterns of neural activity correctly? I suspect that some of the confusion in this area stems from attempts to answer the epistemological question without addressing the metaphysical one. In this chapter, I will discuss the metaphysical question and argue that it does not admit of a good answer because it rests on a confusion about how BCIs work. If I am right, it follows that the corresponding

ethical requirement for interpretive accuracy is also confused. I will argue, in other words, that we *can* blame people for BCI-mediated action *without* having to rule out the possibility of misinterpreted patterns of neural activity. Nevertheless, BCI-mediated action affords the subject with a rather limited degree of voluntary control. In the final part of the chapter, I argue that the relevant sort of limitation has little to do with the interpretation of neural signals, and a lot to do with the causal complexity of the actions carried out by means of certain kinds of tools.

2 Getting the target into view

Before getting into the central argument, I want to narrow down the target phenomenon in two ways. The first narrowing concerns the epistemic state of the BCI user. Philosophical theorizing about moral responsibility and blameworthiness typically invokes an awareness condition (Rudy-Hiller, 2022). The awareness condition is a necessary condition that says something like the following. You are only blameworthy if (i) you know that the behavior in question might generate the sort of consequences that it does, and (ii) you know that those consequences are morally objectionable.[2] In the example above, the thrower knows that her action will cause a shocking and unpleasant experience in the speaker, and that doing so is wrong. This is true even if she happens not be reflecting on these facts at the time the water was thrown. In what follows, I will just assume that we are talking exclusively about cases like this one in which the awareness condition is satisfied. This assumption allows us to focus our attention on the relation between blameworthiness and voluntary control.

The second narrowing concerns the kinds of BCI that are relevant to the discussion. The concerns I want to discuss arise only with respect to what are often called *active BCIs*. Following Steinert et al. (2019), we can characterize active BCIs as those that require the user to intentionally perform a mental task that corresponds to a certain pattern of brain activity that the BCI system detects for processing. Active BCIs differ both from reactive BCIs and from passive BCIs. Reactive BCIs are those in which the neural activity that drives the external effector is prompted by a particular kind of stimulus. I suspect that the analysis I provide here may extend to reactive BCIs, but showing that

[2]It should also be emphasized that I am talking about moral rather than legal culpability. In the legal case, ignorance of the law does not excuse its violation (ignorantia legis non excusat.)

it does would involve various complications better left for another occasion. The analysis provided here does *not* extend to passive BCIs, which differ from active BCIs in two ways. First, they do not require the user to perform any particular mental task. Second, their output is not used as a control signal for an effector system outside the brain. Instead, it is used to control electrical stimulation within the brain itself. In light of both differences, passive BCIs do not clearly facilitate or mediate discrete actions in the way that active BCIs do. Consequently, they do not raise the puzzle about blameworthiness sketched in the previous section.

I would also like to set aside BCIs designed to synthesize speech. What I say here is not intended to apply to them because, when trying to work out what counts as voluntary speech, a number of complicated questions about the relation between thought and language have to be confronted. Those questions are hard, and discussing them properly would be better done in a dedicated article.

3 The faithful interpretation requirement

Now that we have narrowed down the target phenomenon, we can turn back to the intuition I tried to pump in the first few paragraphs. Let us elevate that intuition to the status of a principle, and call it the *faithful interpretation requirement*. Here is what it says:

> In order to have voluntary control over some BCI-mediated action, it must be the case that the BCI device faithfully interprets the patterns of neural activity in the head of the BCI-user.

Lots of commentators have suggested a principle like this one (Miller, 2022; Steinert et al., 2019; Rainey et al., 2020). One of the goals of this chapter is to cast doubt on the very idea of faithful interpretation, and thereby on all of the principles that draw on it. A useful way to begin the discussion is by noticing an ambiguity in a claim about BCI ethics made by Rainey et al. (2020). They present a simple decision-tree diagram, and claim that a BCI-user can only have voluntary control over BCI-mediated action if the activity of the BCI conforms to the decision tree. The crucial step in that decision tree is the first one, according to which the BCI device must check whether the neural state that triggered the device was "intentionally realized." If it is not, the BCI should

refrain from action initiation. If the BCI violates this rule, and initiates action without having confirmed that the pattern of neural activity was "intentionally realized," then the action is not subject to voluntary control.

The question "was the trigger state intentionally realized?" presents us with two interpretive difficulties. One concerns the realization relation. The realization relation relates a higher-level state to a lower-level state. In its most familiar domain of application, it relates a mental state to a neural one. Judging from the nature of the discussion in which their claim appears, it seems reasonable to assume that Rainey et al. have this familiar sort of mental-neural relation in mind. However, because the term "trigger state" seems to refer to the neural side of the relation, when we ask, in the passive voice, whether that state is realized intentionally, it looks as if we are being asked to evaluate some other state, at an even lower level of description. The trouble is that there is no indication of what this yet lower level state would be. My guess is that what Rainey et al. really want to ask is whether the pattern of neural activity was *generated* intentionally.[3]

The other interpretive difficulty concerns the meaning of "intentionally." In this context, that word might be intended to invoke either of two contrasts. One contrast is between voluntary and non-voluntary neural activity. If we foreground that contrast, then the faithful interpretation requirement is a requirement that the pattern of neural activity that triggered the BCI was generated voluntarily. The other contrast is between cases in which the pattern conforms to a particular intention, and cases in which the pattern fails to conform to that intention. Both contrasts are interesting, so, rather than choose between them, let us explore both, and see where the exploration leads.

4 The voluntary pattern requirement

What does it take to have voluntary control over the activity of a small population of neurons? This question has the ring of paradox to it. On the one hand, we have neither introspective nor observational access to the activity of our own neurons. In this respect, neurons are similar to internal organs whose activities are governed by the autonomic nervous system. How can you have voluntary control over some internal biological structure without ever being able to know

[3] In an email, Rainey acknowledged the ambiguity and explained that the passage was left ambiguous precisely because the authors were not prepared to address the difficult philosophy of mind questions that an attempt at disambiguation would bring to the fore.

what it is doing? On the other hand, paradigmatic voluntary actions - such as raising your hand - are determined by activity in motor cortex. How can a voluntary action be determined by a non-voluntary cause? This line of reasoning makes it seem that voluntary control over a small population of neurons is both impossible and necessary.

Cognitive psychologists typically assume that an experimental subject acts voluntarily if (i) the psychologist requests that the subject perform some action that the subject would not otherwise be likely to perform, and (ii) they perform it (Wegner, 2002). This suggests that the appropriate test for whether a person has voluntary control over patterns of neural activity is to ask them to "steer" the relevant population into a pattern of activity that would not likely be observed in the absence of such a request. Call this the voluntary pattern requirement.

The most obvious problem with the voluntary pattern requirement is that it can be satisfied trivially. Imagine you come to learn that whenever you contract your abdominal muscles intensely, the activity of a small population of neurons in motor cortex increases. You then proceed to contract your abdominal muscles intensely, and claim to have demonstrated voluntary control over that neural population. In this case, your purported control over your own neural own activity depends on your having had antecedent control over your abdominal muscles. This kind of derivative control is inappropriate for evaluating BCI-mediated action because, for the purpose of assessing blameworthiness, we cannot assume that you had antecedent control over your BCI. To make that assumption is to beg the question against anyone who, like the water thrower in the opening vignette, claims not to have been in control.

Perhaps what has gone wrong is that we have neglected to include a constraint on the intentional object of the action. In the case just described, the intentional object of the action is presumably the abdominal muscles, rather than the neural population. We have to ask, therefore, what it takes to ensure that the intentional object of someone's voluntary action is the activity of a neural population, rather than the bodily *effects* of that activity. Above, I suggested that voluntary control of some biological structure requires knowledge of what that structure is doing. Perhaps, in order to ensure that the neural population is the intentional object of the action, what is required is a particular sort of knowledge: knowledge of real-time neurodynamics. If so, voluntary control over a small neural population requires neurofeedback, which we can think of as any technology that provides the subject with a real-time perceptual signal

that accurately represents some measurement of the relevant neural activity.

Imagine someone who has an electrode array implanted in motor cortex. Rather than controlling a BCI, this array simply measures neural activity, and sends those measurements to a computer, which represents them by means of a graphic display on a screen. After having spent a considerable amount of time training with such a neurofeedback setup, this person claims to be able to act by attending to the neurons themselves, rather by attending either to the muscles in the arm controlled by those neurons, or to the worldly target of the action (such as, for example, a glass of water.) In this case, the kind of control the person has over the relevant neural population is more direct than it was in the abdominal case. Plausibly, this is a case in which the intentional object of the action really is the neural population.

Let us also imagine that, after the neurofeedback user has mastered control of the relevant population, the electrode array is decoupled from the neurofeedback program, and hooked up to a robotic arm instead. After some time, the user acquires the capacity to execute complex actions by means of the robotic arm. Can we now say that the user has achieved voluntary control over the relevant neural population? Is this a good refinement of the voluntary pattern requirement? I can think of two reasons for doubt.

The first reason is that neurofeedback seems superfluous. Imagine two expert BCI-users, each of whom is outfitted with a robotic arm. They have acquired the same highly impressive degree of behavioral competence. They can catch fly balls, knit socks, and shake hands. However, one of the two was trained with neurofeedback, and the other was trained simply by experimenting with the BCI itself. If neurofeedback is a criterion for voluntary control of a BCI device, then only one of these two BCI experts has any voluntary control at all. This strikes me as a case of the philosophical tail wagging the psychological dog. As independently interesting as it may be, the capacity to pass a neurofeedback test cannot be a necessary condition on having voluntary control over a BCI device.

The argument I just gave assumes that neurofeedback is genuinely distinct from BCI-use. Another way to challenge the neurofeedback test is to say that the robotic arm is itself an elaborate kind of neurofeedback. You can think of the movement of the arm as a convoluted representation of the activity of the neural population. If we say that BCIs are a kind of neurofeedback, then the neurofeedback test will inherit the same triviality concern that arose in the abdominal case. Every BCI-user who has acquired behavioral competence, and

who has some general knowledge about the mapping between neural activity and arm movements, will satisfy the neurofeedback test and thereby satisfy voluntary pattern requirement. But in that case the voluntary pattern requirement will turn out to be useless, because it will not help to distinguish the blameworthy from the blameless.

5 The matching requirement

Here is a rough statement of the matching requirement: in order to have voluntary control over a BCI device, the pattern of neural activity that triggers the BCI must conform to the relevant occurrent intention.

I will give two reasons to doubt the matching requirement. But first, let us look more carefully at how the requirement is formulated. The phrase "conforms to" is unusual in this context. Why not appeal to more familiar language, and say that the pattern of neural activity must "realize" the appropriate intention? The trouble with this formulation is that, even if intentions are in some sense *realized* in the brain, they are probably not realized by a pattern of neural activity whose spatial distribution overlaps exactly with the pattern that triggers the BCI. The largest implanted electrode arrays are only about a centimeter in width, and each individual electrode on that array can only acquire information from a spatial region about hair's width from its surface. The size of the triggering pattern is therefore tightly constrained by the size of the array. If we require the triggering pattern to realize the appropriate intention, then intentions will have to be localized to a narrowly circumscribed neural population in motor cortex. This *could* turn out to be true, but as Uithol et al. (2014) and Anderson (2014) have argued, many other brain regions are probably involved. At the very least, one should not assume this kind of strict localization on philosophical grounds.

Now that we are aware of the problem of localization, we can formulate the matching requirement a bit more precisely, with an eye to accommodating the view that, insofar as intentions are realized in the brain, they are likely realized in distributed patterns of activity.

> In order to for a BCI-user to have voluntary control over a BCI device, it must be the case that (i) the triggering pattern carries enough information about the realizing pattern to distinguish it from all

other realizing patterns, and (ii) the device must decode the relevant occurrent movement intention from the realizing pattern correctly.

5.1 The role of non-neural information

Now that we have a more sophisticated version of the matching requirement on the table, we can begin to evaluate it. The first reason to doubt the matching requirement is that it is probably impossible to decode intentions from the limited store of information available to BCI devices. I can offer two arguments for this conclusion.

The first argument, which is the more philosophical of the two, emphasizes the fact that intentions, unlike motor representations, play a systematic role in rational deliberation. As Bratman (1987) puts it, intentions serve both as *conclusions* of bouts of practical reasoning about ends, and as *premises* in bouts of practical reasoning about means. It is difficult to see how they could play this role unless they have propositional structure. Motor representations are not like this. They are representations, and therefore have content of some kind, but there is little reason to expect that content to come packaged in propositional format (Butterfill and Sinigaglia, 2014; Mylopoulos and Pacherie, 2017). For this reason, philosophers typically conceptualize intentions, but not motor representations, as propositional attitudes.

In order to know whether a BCI can decode intentions from the brain, therefore, we have to have an account of the relation between propositional attitudes and patterns of neural activity. Here it will be useful to divide the space of possible accounts into those that accept what I will call the *lossless encoding view*, and those that do not. According to the lossless encoding view, propositional attitudes are discrete, language-like mental representations that are losslessly encoded in neural activity. Here the term "losslessly encoded" is meant to express the idea that the relevant mental representations are encoded both completely and explicitly, such that all propositional attitudes that correctly describe the person's current configuration of mental states could - in principle - be recovered from the neural data without having to appeal to norms of rationality. That is to say, when working out whether the neural data encode some particular propositional attitude P, it is not necessary to consider any other putative attitudes, and to ask whether it *would make sense* for a person with those attitudes to also have attitude P. To put this point metaphorically, the

lossless encoding view says that the propositional attitudes are *written down* in the brain, symbol-for-symbol.

The lossless encoding view will be a natural view to take for those who accept both the language of thought hypothesis (LOT) and a reductionistic theory of mind-brain relations. The LOT hypothesis underwrites a commitment to the view that propositional attitudes are expressed as language-like mental symbols, while the reductionism provides the commitment to the view that those symbols will be unambiguously identifiable in patterns of neural activity. For our purposes, the upshot of the lossless encoding view, is that *if* we had sufficient knowledge of the mapping between propositions, on the one hand, and patterns of neural activity, on the other, it would be theoretically possible for a machine to decode intentions, despite not having recourse to rational norms.

There are many reasons one might be inclined to reject the lossless encoding view. Connectionists, for example, hold that the rules of thought are simply not language-like (Churchland, 1992; Shea, 2018). More relevant for our purposes are non-reductionistic philosophical views about the relationship between propositional attitudes and the brain. The most well-known reasons for taking an anti-reductionist view in this area have to do with the norms of rationality. Some philosophers hold that the norms of rationality that govern relations between the attitudes is not merely helpful in *discovering* which propositional attitudes correctly describe a person's current configuration of mental states. Instead, those norms are partially *constitutive* of the attitudes themselves.

According to such non-reductionist views, in order to work out which propositional attitudes are indicated by a pattern of neural activity, it is strictly necessary (rather than merely helpful) to reason about which attitudes it would make sense for the person to have. Crucially, this assent to normative reasoning necessarily involves appeal to at least two sources of *non-neural information*. One such source is biographical. According to one popular view, for example, it is necessary to take into consideration which attitudes the person is already entitled to and/or obligated to have, in virtue of their history of engagement with other people (Sellars, 1963; Brandom, 1994). According to another popular view, it is necessary to to take into consideration which clusters of attitudes best predict the person's past behavior (Dennett, 1989; Stalnaker, 1984) According to both views, it is also necessary to take the environmental context into consideration. To return to the opening example, if someone has an intention to throw a glass of water, then, barring hallucination, there must be a glass of water in the environment for her to throw. According to non-reductive views

of the relation between propositional attitudes and patterns of neural activity, one cannot extract propositional attitudes from brain data without taking such non-neural information into account (Rathkopf et al., 2022).

How does this bear on the question of whether BCIs can decode intentions from neural activity? If intentions are propositional attitudes, and if propositional attitudes cannot be recovered without consulting non-neural sources of information, then any device capable of "decoding" the attitudes will have to have, at the very least, access to those sources of information. As a matter of fact, BCIs do not have such access. They only have access to the averaged firing rates of a localized population of neurons.[4]

Although I am sympathetic to the non-reductionist views just canvassed, they might be wrong. And the lossless encoding view might be right. This brings us to the second argument for the claim that it is impossible to decode intentions from the limited store of information available to BCI devices. The argument is this: the lossless encoding view only says that decoding intentions is possible in principle. Given the actual design of BCI devices, however, decoding intentions is unlikely to be possible with anything like the kind of technology currently available. Of particular relevance is the fact that, in order to achieve the kind of decoding that is compatible with the lossless encoding view, it is necessary to know how to clearly delineate discrete mental symbols from dynamical neural data. This emphasis on the individuation of symbols stands in tension with the fact that the algorithms that drive BCI devices are trained by means of machine learning. They are not built to identify discrete vehicles of content. Instead, they are built to dredge up correlational information opportunistically. Nothing in their design forces these models to respect a theory about how neural symbols might be individuated. So, even if you hold the lossless encoding view, and think that it is in principle possible to decode intentions from patterns of neural activity alone, you should be wary of the view that current BCI technology can manage the task.

[4]Here I assume that the definition of BCIs restricts them to being something like the sorts of BCIs that exist today. We can imagine a more powerful device that, in addition to measuring neural data, also collects video and audio data over long periods of time, and feeds it all into a super-algorithm trained on rational propositional attitude ascription. I see no philosophical reason why such a device couldn't recover propositional attitudes accurately, but I assume that such a device would not count as a BCI.

5.2 Learning and the direction of fit

In the previous section, I argued that intentions probably cannot be decoded from the kind of information to which BCI devices have access. That was the first reason to doubt the matching requirement. I now turn to the second reason, which has to do with how we conceptualize the causal structure of the BCI.

Thus far, I have been talking about BCIs as if they are devices that are designed to *decode* the meaning of neural activity, and respond appropriately. On this view, the success of BCI applications depends on a successful episode of what is sometimes called *brain reading*. However, the reading metaphor is not forced upon us by the empirical facts.

To conceptualize the success conditions for BCI devices more realistically, we need to emphasize the role of learning and plasticity. When a BCI works well, its success is not explained merely by the clever design of the device, and the clever training of its underlying algorithm, which together enable the device to make accurate inferences from passive brain signals. Instead, its success is explained by the fact that the brain has learned how to exploit the affordances of the BCI for its own ends. That is, the explanation for successful BCI operation has a direction of fit that goes from the brain to the device, rather than the other way around. The coordination is *largely* an achievement of the brain, and of the person whose brain it is.

I do not mean to deny the claim that BCI models learn from neural data. Adaptive processes run in both directions, from device to brain, and from brain to device. Nevertheless, for the purposes of evaluating the matching requirement, the fact that the success of BCI operation is explained to some degree by the adaptive properties of the brain is consequential. The fact that the brain adapts itself to better exploit the limited receptivity of the BCI device implies that there is no fixed target for the BCI to adapt itself to. The successful use of a BCI is better described as causal entanglement than signal interpretation.

What is the evidence for thinking about the coordination between brain and BCI this way? One source of evidence is the fact that control over a BCI always requires a period of learning. In some cases, it is possible to acquire a modicum of control quickly, but it usually takes days before learning hits a plateau. During this time, the model that governs the BCI device need not, and often does not, undergo additional training. The learning period is necessary because the brain must work out how to adapt itself to the device.

Another source of evidence has to do with the fact that control over external

effectors can be achieved even if the electrodes are placed in improbable locations in cortex. One of the most striking illustrations of this idea that I have come across was an experiment done with mice in 2014 (Clancy et al., 2014). Using two-photon imaging, Clancy et al. showed that mice fitted with a BCI could learn to manipulate the pitch of an auditory stimulus by regulating the firing rate of a small population of neurons. The surprising thing about this study is that the two-photon imaging was done both in motor cortex and in somatosensory cortex, with approximately equal success. Because somatosensory cortex is, to put the point crudely, an *input* area of the brain, the fact that it could be used just as well as motor cortex to control a BCI shows that the coordination cannot rely on the BCI having been designed to interpret the meanings of signals that are native to motor cortex.

This experiment puts strain on the view BCI is a device for *reading*, *decoding* or *interpreting* neural activity. A BCI is a *tool* used by the brain to get a job done. Because the matching requirement demands that the BCI engage in an act of interpretation, the direction of fit argument also gives us a reason to think that the matching requirement cannot be satisfied.

We now have two reasons to think that the matching requirement cannot be satisfied. We also saw that the voluntary pattern requirement seemed likely to be trivial. So both versions the faithful interpretation requirement seem problematic. If, in the light of the foregoing arguments, you have lost your faith in the faithful interpretation requirement, you have two options. You can either accept the implication that BCI users *never* have voluntary control over their BCI devices, or we can reject the matching requirement itself. If you watch BCI-users expend tremendous effort trying to move a robotic arm, the former view seems absurd. I conclude that we should reject the matching requirement. To put the point positively, we can - in principle - show that someone is culpable for some BCI-mediated action without having to show that the BCI decoded some neural signal correctly.

6 Compromised control through technological mediation

To say that it is possible to blame someone for some BCI-mediated action without saying anything about decoding accuracy is not to say that all cases of BCI-mediated action are equally blameworthy. Blameworthiness depends on

voluntary control, and the degree of voluntary control over BCI-mediated action can vary. In this section, I hope to explain why BCI-mediated action is not always under voluntary control, and consequently why it is not always blameworthy.

In the previous section, I claimed that a BCI is a tool wielded by the brain. That is the foundation for the positive view. We should handle matters of control and culpability in BCI-mediated action just as we handle such matters in the case of action that is mediated by ordinary tools. Rather than making this recommendation the conclusion of an explicit argument, I will try to show that it helps us sort through relevant kinds of cases in ways that seem insightful.

The first observation is that ordinary tools afford variable degrees of control. Let us draw a coarse-grained distinction between cases in which it is easy to judge that an action was under control, and therefore intentional, and cases in which it is more difficult. Here is a paradigmatic case of the former sort. A hunter spends hours stalking a deer through a dense forest. He finally finds a favorable position, aims with great care, fires his rifle, and hits the deer in the heart. If that hunter receives a fine for hunting without a license, it will not help for him to claim that his bullet hit the deer accidentally. Two features of the case make that claim implausible. The first is that if you fire a shot at random in the forest, the shot is much less likely to hit a deer than a tree, and even more unlikely to hit a deer square in the heart. The other feature of the case is that rifles are highly predictable, easily controllable tools. Bullets travel in approximately straight lines, rifles only fire one round, and they have a safety mechanism built in.

In other cases of technology-mediated action, matters of control and culpability are more nuanced. Think of a person flying a kite in a fickle wind in a crowded park. The wind suddenly dies, the kite falls, and hits a passer-by on the head. The kite flyer is not culpable for assaulting the passer-by in the same way that the hunter is culpable for having killed the deer. Again, two features of the case seem relevant to this judgment, and they run parallel to the two features I highlighted in the hunting case. First, because the park was crowded, the probability of the kite hitting someone on the head is considerable. Second, because the wind is fickle, one cannot easily predict how long the kite will stay aloft, or where it will be by the time it reaches the ground. Under these circumstances, the claim that the collision was unintentional is not so easy to dismiss.

Cases of BCI-mediated action come in many varieties. Some will be like the

hunting case, and others will be like the kite case. So, with these two cases as guides, let us return to the vignette with which the chapter began. Recall that, in that rather fanciful story, one colleague uses a BCI-controlled robotic arm to throw a glass of water in the face of another colleague. I submit that this case is analogous to the deer hunting case, and should be treated similarly. Achieving a very specific and complex movement trajectory by means of a BCI-controlled robotic arm requires intense concentration. If the concentration is broken, the action will fizzle. Moreover, if the glass of water were tossed in a randomly-chosen direction, it is unlikely that the water would just happen to land in the face of the only person in the room with whom you are currently angry. In this case, therefore, the defense that "I was not in control of my own arm" is vanishingly improbable. The water thrower is straightforwardly culpable for her action, regardless of which patterns of neural activity were causally involved in bringing that action about.

The water-throwing case is fanciful. Many BCI-mediated actions will be more like the kite case. Imagine a person learning how to shake hands by means of her robotic arm. She grasps the proffered hand too tightly, and injures it. We know from experiments in robotics that exerting an appropriate amount of pressure in a human handshake requires very fine-grained motor control. If you were to guess at the appropriate level of pressure, in the absence of sensory feedback, you would probably get it wrong. The probability of accidentally causing harm is therefore considerable. Moreover, in the absence of sensory feedback, predicting how much pressure your robotic hand is currently delivering is also quite difficult. In a case like this, it seems reasonable to judge that the BCI user is not culpable for assault in the same sense that the water-thrower is.[5]

Making judgements about culpability in kite-like cases is difficult because you have partial control, and there is no algorithm for determining how much control is sufficient for culpability. Nevertheless, I think the moral uncertainty built in to such cases has nothing to do with the brain, per se, and everything to do with the fact that some tools are intrinsically difficult to control. Let's call this phenomenon *compromised control through technological mediation*, or

[5]Experiments with BCI-governed robotic arms that do incorporate sensory feedback are just getting off the ground now. They are difficult to justify, because they involve drilling two holes in a person's head, rather than just one. In any case, it is worth asking how my argument would fare if accurate, high-bandwidth sensory feedback could easily be achieved. That would have the effect of adding more control to the system, and thereby making the exculpatory strategy inappropriate in a wider variety of cases.

CCTM. As the kite case shows, cases of CCTM need not be high-tech. Nevertheless, high-tech tools will often give rise to cases of CCTM. Think of driving a car, operating a drone, or performing robot-assisted surgery. BCI-mediated action is like these other forms of technology-mediated action. When something goes wrong and harm is caused, it can be hard to work out whether the tool user is directly culpable for the harm, or whether some degree of uncontrollability in the technology itself should be taken into account, in which case the tool user is guilty, at worst, of negligence.[6]

On the view I have articulated here, concerns about control and culpability for BCI-mediated action are substantially different from the corresponding concerns about action carried out by the native body (like, for example, when you punch someone with your fist.) But the reason that these concerns are different is not because we are using a technology that has crossed the barrier of skin and skull, as Clark and Chalmers (1998) famously put it. Rather, it is because BCIs are complex tools that will often give rise to cases of compromised control.

[6]In German, the phrase "technologisches Versagen" captures the phenomenon I am trying to describe. It is often used to describe circumstances in which (i) the technological artifact performs as it was designed (it did not malfunction), (ii) the user of the technology had "good intentions," and, nevertheless, (iii) something bad happens. I cannot think of a good English translation of this phrase. "Technological failure" is a poor choice because it suggests malfunction. "Technological shortcoming" would be better, but seems to be used primarily to refer to circumstances in which technological progress in some domain has been slower than expected, rather than to a particular technology.

References

Anderson, M. L. (2014). *After phrenology*. MIT Press.

Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard university press.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Cambridge: Cambridge, MA: Harvard University Press.

Butterfill, S. A. and Sinigaglia, C. (2014). Intention and motor representation in purposive action. *Philosophy and Phenomenological Research*, 88(1):119–145.

Churchland, P. M. (1992). *A neurocomputational perspective: The nature of mind and the structure of science*. MIT press.

Clancy, K. B., Koralek, A. C., Costa, R. M., Feldman, D. E., and Carmena, J. M. (2014). Volitional modulation of optically recorded calcium signals during neuroprosthetic learning. *Nature neuroscience*, 17(6):807–809.

Clark, A. and Chalmers, D. (1998). The extended mind. *analysis*, 58(1):7–19.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Miller, D. J. (2022). Two problems of moral luck for brain-computer interfaces. *Journal of Applied Philosophy*, 39(2):266–281.

Mylopoulos, M. and Pacherie, E. (2017). Intentions and motor representations: The interface challenge. *Review of Philosophy and Psychology*, 8(2):317–336.

Rainey, S., Maslen, H., and Savulescu, J. (2020). When thinking is doing: responsibility for bci-mediated action. *AJOB neuroscience*, 11(1):46–58.

Rathkopf, C., Heinrichs, J.-H., and Heinrichs, B. (2022). Can we read minds by imaging brains? *Philosophical Psychology*, 36(2):221–246.

Rudy-Hiller, F. (2022). The Epistemic Condition for Moral Responsibility. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Sellars, W. (1963). Science, perception, and reality.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

Stalnaker, R. C. (1984). *Inquiry*. Cambridge University Press.

Steinert, S., Bublitz, C., Jox, R., and Friedrich, O. (2019). Doing things with thoughts: Brain-computer interfaces and disembodied agency. *Philosophy and Technology*, 32(3):457–482.

Uithol, S., Burnston, D. C., and Haselager, P. (2014). Why we may not find intentions in the brain. *Neuropsychologia*, 56:129–139.

Wegner, D. (2002). *The Illusion of Conscious Will*. MIT Press.