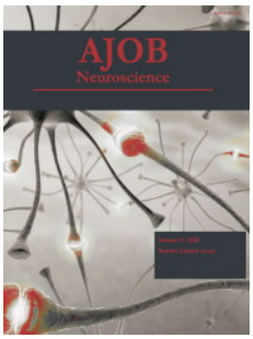

AJOB Neuroscience

ISSN: 2150-7740 (Print) 2150-7759 (Online) Journal homepage: www.tandfonline.com/journals/uabn20

Closed-Loop Neurotechnologies, Agency and Mental Interference

Vera Tesink, Thomas Douglas, Lisa Forsberg, Sjors Ligthart & Gerben Meynen

To cite this article: Vera Tesink, Thomas Douglas, Lisa Forsberg, Sjors Ligthart & Gerben Meynen (11 Jun 2026): Closed-Loop Neurotechnologies, Agency and Mental Interference, AJOB Neuroscience, DOI: 10.1080/21507740.2026.2678800

To link to this article: https://doi.org/10.1080/21507740.2026.2678800

Published online: 11 Jun 2026.

AJOB NEUROSCIENCE
https://doi.org/10.1080/21507740.2026.2678800

TARGET ARTICLE

OPEN ACCESS Check for updates

Closed-Loop Neurotechnologies, Agency and Mental Interference

Vera Tesink[a], Thomas Douglas[b], Lisa Forsberg[b], Sjors Ligthart[c,d] and Gerben Meynen[a,d]

[a]Vrije Universiteit Amsterdam; [b]University of Oxford; [c]Tilburg University; [d]Utrecht University

ABSTRACT

Closed-loop neurotechnologies bring great promise for treating neurological and psychiatric disorders. However, the use of artificial intelligence (AI) in their application raises ethical concerns, since AI-driven closed-loop devices may cause unforeseen mental interference that might, absent consent, infringe the user's mental rights. Whether such worries are warranted, however, may depend on whether closed-loop neurotechnologies qualify as moral agents, and on whether they are distinct from the moral agent on whose brain they act. If they are not moral agents, or are not separate moral agents, they will arguably be incapable of infringing the user's mental rights. In this article, we explore different possible agential relationships between the human user and closed-loop neurotechnologies and consider the implications for the protection that our mental rights provide.

KEYWORDS

Closed-loop neurotechnology; agency; artificial intelligence; mental interference; moral rights

INTRODUCTION

Closed-loop neurotechnologies read and analyze brain activity and adapt their output to modify brain activity accordingly—without any human intervention (Kohler et al. 2017).[1] In the future, these technologies are likely to employ artificial intelligence (AI) in the form of adaptive machine learning to analyze brain data and determine the output (Zhu et al. 2020; Kellmeyer 2021).[2] Closed-loop neurotechnologies are of great interest to neurology and psychiatry, as some studies suggest they might be effective in alleviating symptoms in certain disorders such as major depressive disorder and obsessive-compulsive disorder (Sullivan et al. 2021) and do not present the same disadvantages as "traditional" open-loop devices, such as the need to infer optimal stimulation settings through trial and error (Klein et al. 2016). Closed-loop neurotechnologies are also being developed for non-therapeutic purposes, such as cognitive enhancement (Valeriani et al. 2022).

While closed-loop neurotechnologies thus hold promise, their use also raises ethical concerns regarding mental interference (Kellmeyer et al. 2016; Ligthart et al. 2021). Neurotechnologies can affect the user's mental life in ways that may constitute mental interference and, if not consented to, might infringe moral and/or legal rights over the mind—henceforth referred to as "mental rights"—such as the right to mental integrity (Craig 2016; Shaw 2022), the right to mental self-determination (Bublitz and Merkel 2014; Bublitz 2020) and the right to freedom of thought (Alegre 2017; McCarthy-Jones 2019). These mental rights are meant to protect the mind from undue influence, for instance by neurotechnologies, and are usually thought to be held against other moral agents.

In cases of open-loop neurostimulation, it is generally a human (moral) agent (typically a healthcare professional) who is in control of the neurostimulation and who may infringe the user's mental rights. For closed-loop neurotechnologies, however, this is less clear. Some argue that closed-loop neurotechnologies, because they are driven by AI, might qualify as (minimal) moral agents, meaning they would be capable

CONTACT Vera Tesink v.tesink@vu.nl Faculty of Social Sciences and Humanities, Department of Philosophy, Vrije Universiteit Amsterdam, Amsterdam, Netherlands.

[1]Unlike "open-loop" devices where a medical doctor actively controls the neurostimulation.

[2]The output might vary between different kinds of closed-loop neurotechnologies, and will depend on whether the loop fully takes place within the skull, such as closed-loop deep brain stimulation (Parastarfeizabadi and Kouzani 2017), or the loop extends beyond the skull, such as brain-computer interfaces that direct their output at external devices such as prostheses or wheelchairs (Belkacem et al. 2023). Another difference is that the latter may require active involvement of the patient in the form of, for instance, forming intentions (Wolpaw and Wolpaw 2012), while the former generally does not require any such mental participation. Our focus in this article will be on the "within-skull" closed-loop neurotechnologies that are most useful in psychiatric contexts.

of infringing rights.[3] If so, when closed-loop neurotechnologies "autonomously" and adaptively intervene in the user's brain in ways that go beyond initial consent, they might, in doing so, infringe the user's mental rights, suggesting that closed-loop neurotechnologies could pose a serious threat to users' minds.

Still, the extent of the threat that closed-loop neurotechnologies pose to the mind depends not only on their status as moral agents, but also on the degree to which the agency of the device remains sufficiently separate from the agency of the user. Closed-loop neurotechnologies can become intricately involved in modifying and creating mental states that underlie the user's actions. When the neurotechnology becomes closely involved in the user's agency, it could be argued that they should appropriately be viewed as together forming a *hybrid* agent, or even *one single* agent (Kellmeyer et al. 2016; Goering et al. 2017). If the closed-loop neurotechnology is no longer a separate (moral) agent, but in some way part of the user's agency, it may no longer be capable of interfering with the user's mind in a way that might infringe mental rights, as these rights are usually understood to be held only against *others*.[4]

Some research has been done into the relationship between (closed-loop) neurotechnologies and a user's *sense* of agency (Haselager 2013; Goering et al. 2017). However, "sense" of agency is established on a phenomenological level and does not necessarily translate to "actual" agency (as the capacity to act in the world) and vice versa (Vukov 2017). For instance, an implanted neurotechnology might improve motor functioning in a patient with Parkinson's disease, thus increasing her agency as the capacity to act, but the patient may feel alienated from the device and not identify with it, therefore reducing her sense of agency. As for the agential relation between closed-loop neurotechnologies and users on an ontological level, thus far little research has been done.[5] Yet, how this relation is understood may have important implications for the extent to which neurotechnologies form a risk to a user's mental rights, which may not be threatened to the same degree by a closed-loop neurotechnology if the neurotechnology in question becomes part of the agency of the user.

In this article, we will explore different potential agential relations between closed-loop neurotechnologies and users and consider their implications for the possibility of mental interference and mental rights infringements.[6] To do so, in the next section, we briefly explain how we will understand mental rights infringements and the role of (external) moral agents. In the second section, we consider whether closed-loop neurotechnologies can qualify as moral agents capable of infringing mental rights. In the third section, we discuss an example case of a mental rights infringement by a closed-loop neurotechnology as a moral agent. In the fourth section, we explore three different possible agential relationships between closed-loop neurotechnologies and their human users, and examine what this might imply for mental interference and mental rights infringement. We conclude that, assuming that closed-loop neurotechnologies are moral agents capable of infringing mental rights, whether closed-loop neurotechnologies threaten the user's mental rights depends on the extent of agential distinctness that exists between the two, as significant agential integration seems to prevent infringements of rights that are usually held only against other agents.

NEUROTECHNOLOGIES THREATEN MENTAL RIGHTS

Neurotechnologies are often thought to pose a threat to the mind. Upon altering mental states of users by modifying neural activity, they can interfere with users' mental states. While there is no agreed upon definition of mental interference, based on accounts in the literature, inducing significant mental alterations via direct physical alteration of brain states is usually considered a mental interference (Bublitz and Merkel 2014; Craig 2016; Shaw 2022; Lavazza and Giorgi 2023; Ratoff 2024; Zuk 2024; Douglas 2025). If such mental interference is not consented to by the user, it would typically infringe the user's mental rights—or rights over the mind. More specifically, it would plausibly infringe at least one of the right to mental integrity (Craig 2016; Shaw 2022), the right

[3]We will discuss the possibility of moral agency of closed-loop neurotechnologies in Section 2.

[4]We will elaborate on our account of mental interference and mental rights infringement, and the role of external moral agents, in the next section.

[5]An article by Steinert et al. (2019) discussed whether and which BCI-mediated events might qualify as actions according to standard accounts of action in philosophy and law.

[6]Identifying rights infringements is often considered one step within a broader assessment of liability and responsibility, but mental rights infringements are also normatively significant in their own right: they mark a distinctive wrong—crossing a boundary of a person's mental life—conceptually separable from blame or compensation, and they can ground immediate reasons for action (e.g., halting ongoing interference, strengthening safeguards) even before responsibility is assigned. The focus of our article is on questions about rights infringements by closed-loop neurotechnologies, and we bracket downstream responsibility issues, which, while important, require separate addressing in future work.

against mental interference (Douglas 2026), the right to mental self-determination (Bublitz and Merkel 2014; Bublitz 2020) or the right to freedom of thought (Alegre 2017; McCarthy-Jones 2019). Infringements of such rights over the mind mark a distinct kind of normative wrong, where a protected boundary over a person's mental life is crossed. The right to mental self-determination, for instance, is thought to protect "freedom from severe interferences by the state and third parties" (Bublitz and Merkel 2014, p. 58), and the right to mental integrity can be infringed by "intentionally interfering with a person's mental states through non-rational means" (Shaw 2022, p. 1418). Rights to protect the mind seem especially relevant in the context of neurotechnologies that can influence mental states through physically altering brain activity, where the rights thus should provide a protective barrier against non-consensual mental interference.

Typically, infringements of (mental) rights presuppose a moral agent that does the infringing. While non-agents or non-moral agents may arguably also interfere with the mind—flashes of light due to lightning might cause a person to experience an epileptic seizure or a cat might urinate in a drinking water tank, causing a person who drinks the water to experience delirium—such interference is unlikely to infringe any moral rights, as this requires the involvement of moral agents that can act according to moral rules (Warren 2000). Also unlikely to infringe moral rights is mental interference by oneself, which might for instance occur when one recalls a traumatic memory. In such cases the agent of interference is also (a part of) the rightholder, and it is doubtful that one can hold rights against oneself. Even if one can, the rightholder is also in the position to waive the right, so in interfering with oneself, it is plausible that one (implicitly) waives her own right against such interference (Cholbi 2015; Muñoz 2024).

Often when speaking about a mental rights infringement by an open-loop neurotechnology, it is assumed that involved human (moral) agents, such as a doctor, interfere with the mind *through* the neurotechnology. For instance, if a medical doctor uses a noninvasive neurotechnology such as transcranial magnetic stimulation to magnetically stimulate certain areas of a patient's brain and causes significant mental effects that were not consented to, we would say that the doctor is infringing the patient's mental rights *through* the neurotechnology. With closed-loop neurotechnologies, however, this might be different; there is no human moral agent directly controlling the neurostimulation, but an AI algorithm that continuously decides when and how strongly to stimulate the brain (Schopp et al. 2025). This raises the question of whether the device itself might be the moral agent that infringes the patient's mental rights.

CLOSED-LOOP NEUROTECHNOLOGIES AND MORAL AGENCY

The involvement of an AI system that drives neurostimulation decisions could suggest that closed-loop neurotechnologies can be agents, and might, on some accounts, even suggest that they possess some level of moral agency. The view that technological devices employing AI such as closed-loop neurotechnologies can be considered *agents* is supported by a substantial strand of the literature. In very broad terms, an agent is a being that has the capacity to act, and agency is the *exercise* of this capacity (Schlosser 2019). On standard causal theories of action, events are actions if they are caused by intentional mental states in the right way (Davidson 1980; Bratman 1987; Schlosser 2019), where "in the right way" means that actions are caused by mental states via non-deviant causal chains (Bishop 1989; Mele 2017).[7] While more basic technological entities that need human input for every output appear to lack such intentional mental states, several scholars argue that AI systems that display behaviors and cognitive capacities similar to those traditionally associated with human beings might have this ability to form intentional mental states and therefore qualify as agents (Floridi and Sanders 2004; List 2021; Ziemke 2023). Moreover, others argue that even without possessing "genuine intentional states like any biological being," they might still possess "minimal agency" that indicates the capacity for goal-directed behavior without presupposing any (human-like) mental representation (Manna and Nath 2021, p. 2).

This opens up the possibility that AI systems might also be considered *moral* agents. On what Himma (2009) refers to as the "standard account" of moral agency, a being is a moral agent when it acts in accordance with moral rules and laws. It is generally assumed that having a faculty for moral reasoning is a prerequisite for moral agency, and accordingly, consciousness and rationality are often considered key features of moral agency (Behdadi and Munthe 2020; Manna and Nath 2021). It is not clear that AI systems

[7] An example of a deviant causal chain would be a case where the intention to pull a trigger causes a fear-induced tremor that causes one's finger to pull the trigger; this would not be an action proper on the standard theory (Mayr 2011).

can satisfy the putative consciousness requirement (Himma 2009).

Nevertheless, some authors contend that AI systems can still be regarded as moral agents when the criteria for moral agency are defined in less stringent, non-anthropocentric terms. Floridi and Sanders' (2004) influential functionalist account makes this move explicit by opting for adopting a "higher level of abstraction" from paradigmatic human cases while still preserving the key features that define human agents as moral agents. On their account of moral agency, then, an agent is a system that is autonomous, interactive and adaptive, and a moral agent is an agent that performs actions with moral consequences—which does not require any internal mental states (Floridi and Sanders 2004). In a similar functionalist vein, Sullins (2006) holds that moral agency requires autonomy (in the "engineering" sense of not being controlled by another agent), intentionality (in the sense that its behavior can be explained and predicted by ascribing beliefs and desires to it, regardless of whether it actually has them) and role-responsibility (as fulfilling a social role that assumes responsibilities toward other moral agents), and he contends that robots or AI systems can possess this and can thus be moral agents (Sullins 2006). On both of these accounts, it seems that AI systems such as those involved in closed-loop neurotechnologies could qualify as moral agents. Fossa (2018) also underwrites the functionalist approach and suggests that moral agency can be a matter of degrees, emphasizing that while human agents and AI agents can display different degrees of moral interactivity, autonomy and adaptability, both can be "different instances of a discrete and formal behavior called 'moral agency'" (Fossa 2018, p. 117).

As for closed-loop neurotechnologies specifically, Kellmeyer et al. (2016) have argued that these technologies possess moral agency if they involve sophisticated and complex computational systems. They claim that when neurotechnologies use machine learning algorithms that make their decision-making sufficiently sophisticated and so complex that even engineers or designers cannot predict their behavior, "this type of adaptive decisionmaking capacity implies both intelligence and autonomy—it certainly seems to satisfy our criteria for moral agency" (Kellmeyer et al. 2016, p. 626).

It has also been suggested that ascribing moral agency to AI systems such as closed-loop neurotechnologies may be pragmatically valuable. Floridi (2013), for instance, claims that it can be practical to treat certain systems as moral agents when it helps to understand and govern their effects, and Fossa (2018) argues that treating a device as a moral agent can help clarify its behavior—selecting, updating and executing policies—and give a measure for normative assessment, whereas a framing as tools risks misallocating the locus of morally salient decision-making. In a similar vein, a growing legal-philosophical literature suggests that attributing agency to advanced AI systems can play an important practical role in contexts where such systems make autonomous, normatively significant decisions. When AI behavior cannot be straightforwardly traced back to the intentions of any single human actor, responsibility risks becoming diffused across designers, programmers, physicians and users, and some authors propose that treating the AI system behavior as rule-governed action rather than an extension of human intentions can help explain norm-salient outputs that the standard "mere tool" model obscures (Hallevy 2010; Mulligan 2017; Lima 2018; Abbott and Sarch 2024). Similarly, for closed-loop neurotechnologies, a tool-model that explains the outcomes of the devices solely by reference to individual human intentions may be explanatorily unsatisfactory because it cannot predict or manage system-level affordances and failure modes that arise from interactions between the device and the user. Human intentions may ultimately become poor guides to the system's adaptive "behavior"; treating the device as an agent can capture the device's evolving decisions as it acts in the moment, and treating it as a moral agent allows for designating the neurotechnology as a source of normatively salient events (rather than assigning such events to absent human agents).

Some commentators doubt whether AI systems can be moral agents because it is not clear that they can be held morally responsible for their actions (Sparrow 2007; Asaro 2011; Constantinescu et al. 2022). Yet, there is disagreement as to whether moral agency is sufficient for moral responsibility; some argue that it is necessary but not sufficient (Parthemore and Whitby 2013; Behdadi and Munthe 2020). Floridi and Sanders (2004) and Symons and Abumusab (2024) maintain that autonomy, interactivity and adaptivity are sufficient for moral agency, while blameworthiness is a separate, higher-order attribute. Others underline this distinction by emphasizing the potential for so-called "responsibility gaps," where artificial systems can perform moral wrongs that seem to require moral assessment even when no involved moral agent (human or machine) seems to satisfy the conditions for blame (Matthias 2004; Danaher 2016). These lines of thought suggest that moral agency need not always imply moral responsibility.

Whereas the foregoing offers some reasons to treat AI systems as moral agents, the position remains contested, especially by those who hold more demanding accounts of moral agency that, for instance, require some form of consciousness (Himma 2009; Purves et al. 2015; Brożek and Janik 2019; Chakraborty and Bhuyan 2024). While acknowledging that the debate is far from settled, in this article we will nevertheless assume, in line with the less demanding functionalist accounts of moral agency just discussed, that AI systems such as those involved in closed-loop neurotechnologies can possess at least a "minimal" form of moral agency. Such "minimal" moral agency does not presuppose the presence of human-like intentionality or consciousness, but merely that a system acts autonomously (in the engineering sense of this term), interactively and adaptively. Moreover, this kind of moral agency would allow for closed-loop neurotechnologies to infringe moral rights without implying that they can be held morally responsible for their actions, as the sort of moral agency that is required for infringing moral rights is arguably weaker than the one required for moral responsibility (Behdadi and Munthe 2020).

CLOSED-LOOP NEUROTECHNOLOGIES AND MENTAL RIGHTS INFRINGEMENTS

If closed-loop neurotechnologies can qualify as moral agents, they may be capable of infringing users' mental rights. Consider the following example of a patient with depressive symptoms—we will call her patient X. She has a closed-loop neurotechnology implanted into her brain to regulate her mood (Widge 2023). The device records brain data and interprets this data, and it implements, via a self-learning algorithm, a pattern of neurostimulation. Before the neurotechnology was inserted, patient X consented to the placing of the device. As part of the consent procedure, she was informed that the device employs a self-learning algorithm capable of adapting stimulation parameters in real time on the basis of ongoing neural data, and that this adaptive capacity may lead the device to adjust stimulation parameters over time in response to neural patterns associated with depressive states. She was also informed that the intention was that the device would stabilize her mood through using neurostimulation to modify the serotonergic aspects of her nervous system—the same aspects influenced by antidepressant medications that X has tried in the past. The aim is to produce similar effects to these antidepressant medications, but tailored to X's real-time neural state.

Now, some time after implantation, the neurotechnology detects abnormal brain activity in X that signifies the onset of a depressive episode and initiates, based on algorithmic calculations, a novel stimulation pattern. While patient X consented to the use of adaptive neurostimulation aimed at regulating her depressive symptoms, the specific intervention implemented by the device in this instance departs significantly from the therapeutic strategy that formed the basis of the original treatment plan. The system had been calibrated to modulate neural circuits associated with serotonergic mood regulation, but after several months of adaptive learning, the algorithm begins stimulating nodes within the dopaminergic reward circuitry to induce a drug-like euphoric state in patient X. As a result of this stimulation, the depressive episode is averted, but patient X experiences pronounced euphoria accompanied by heightened impulsivity and diminished capacity for self-control. The stimulation strategy thus differs markedly from the therapeutic approach that originally motivated the treatment plan, and on the basis of which X had consented to insertion of the device. We suggest that this could constitute an infringement of X's mental rights, since the dopaminergic stimulation was arguably not covered by her initial consent.

If the closed-loop neurotechnology is a moral agent, we might say that the neurotechnology infringes the patient's mental rights, as the neurotechnology causes the mental interference by modifying neural activity. Some may object to this. For instance, some might argue that the designer of the neurotechnology, who designed the algorithm that ultimately decides when to stimulate, is the one who infringes the patient's rights. While the designer may have devised the algorithm that makes stimulation decisions, inherent to self-learning algorithms is that any external oversight—and thus control—over algorithmic decisions diminishes with every new input, rendering it less plausible that the designer is sufficiently implicated in the mental interference to infringe patient X's mental rights (Danaher 2016).

Others may hold that it is the doctor who placed the neurotechnology in the patient's brain who infringes her mental rights, seeing that the doctor implanted and activated the device. What might make this less likely is that the closed-loop neurotechnology is self-governed and decisions about when to stimulate are not made directly by the doctor, and this lack of direct involvement in stimulation decisions—and thus lack of control over the mental alteration that constitutes the mental interference—seemingly makes it less plausible that the doctor is infringing the patient's

mental rights (Danaher 2016).[8] Moreover, the doctor may not know, at the moment of stimulation, what the device has decided or what mental effects will follow, and may only become aware of a rights-infringing mental effect after it has already occurred, weakening the claim that the doctor is the infringer. While the doctor does have the capacity to turn off the device and stop the ongoing rights infringement, this need not imply that she is the one doing the infringing—similar to how a nurse stopping a surgeon while performing an unconsented to procedure on a patient under anesthesia does not imply that the nurse is the one who infringes the patient's bodily rights. Having the capacity to halt the infringement could imply that the doctor is *complicit* in the infringement (Mellema 2016), but complicity still implies there is another moral agent that is the "primary" infringer of the rights.[9]

Still, it may be argued that the doctor bears (at least some) responsibility for the infringement if she has the capacity to stop it.[10] Moreover, this could be so even if the neurotechnology itself cannot be responsible for the direct rights infringement, for example, because it does not possess the capacities required for responsibility. As mentioned earlier, whether or not technological artifacts, even if considered (minimal) moral agents, can be held responsible for actions is contested. However, while rights infringements are typically followed by questions about blameworthiness and compensation for the wrong, recognizing an infringement is a first step that can be separated from assigning responsibility. The separation of rights infringements from questions about responsibility also applies to responsibility ascribed to those who fail to stop an ongoing infringement, as whether for instance the doctor should have deactivated the neurotechnology depends on additional facts—such as notice, foreseeability and availability—that go beyond the "mere" question of whether a rights boundary was crossed. Our analysis thus treats rights infringement as separate from questions about responsibility (and omissions), and this means that the doctor being responsible does not by itself make her the infringer, and the device's inability to bear responsibility does not imply it cannot infringe the patient's mental rights.

Lastly, it might be argued that there is no mental rights infringement to begin with because the patient consented to the implantation of the closed-loop neurotechnology, and also to its adaptive decision-making and (unforeseen) mental effects. While consenting to the use and effects of the neurotechnology can indeed prevent rights infringements, consent may sometimes fall short, especially if we think of the closed-loop neurotechnology as a moral agent. In particular, consenting to a device that autonomously adjusts stimulation does not necessarily amount to consenting to every intervention the device may generate. To illustrate, consider a "traditional" medical case where during surgery to remove an ovarian cyst (which the patient consented to), a surgeon decides to also perform another procedure in the abdomen (because, say, she observes some spots on it that might be an indication of a non-life-threatening disease). This would likely not be covered by the initial consent and if the doctor does not return to the patient to ask for consent for this second intervention, she would infringe the patient's bodily rights. If a closed-loop neurotechnology is a moral agent, we might have to regard it as we do the doctor in this case, where the stimulation decisions made autonomously by the neurotechnology can be additional interventions that may require new consent.[11] Otherwise, the initial consent would effectively grant the device the kind of "carte

[8]This also seems to become less likely the more time has passed between implantation and interference, as the control of the doctor typically diminishes over time (Llorca Albareda et al. 2023).

[9]It might be argued that, since the doctor can stop an ongoing infringement, it may be practically irrelevant whether the neurotechnology is the primary infringer: in any case, we would address the physician to halt the interference. However, in cases where it is clear that the doctor is not the infringer, identifying the neurotechnology as the infringer can still matter for complicity. If the device does infringe, the doctor's knowing inaction that enables continuation can ground complicity-based responsibility, but if no distinct primary infringer exists (e.g., under agential integration, as discussed later), there is no complicity or complicity-based responsibility for the doctor as this specifically presupposes a primary wrong by another agent (though responsibility based on omission or negligence by the doctor may still be assessed on separate grounds). Therefore, whether the neurotechnology is the primary infringer can be relevant for downstream responsibility assignment.

[10]The same may apply to the patient herself who may also have the capacity to stop an ongoing rights infringement by deactivating the neurotechnology, and this might mean that she also bears some responsibility for the infringement. However, similar to the doctor, this does not mean that she is the infringer of rights, as this can be separated from responsibility questions. Also similar to the doctor, the patient will likely be unaware of what the device has decided and what the consequences of such changes in stimulation may be due to the device's grounds for decisions and the decisions themselves as well as their (side-)effects being to a significant degree be opaque. So, while both the doctor and the patient may physically be in a position to turn off the stimulation, they need not be epistemically in the position to do so in order to prevent a rights infringement from occurring.

[11]Some may suggest that such unforeseen events could be prevented by implementing algorithmic safeguards, such as coding an algorithm in such a way that it does not perform certain actions without consulting a doctor first. However, this may not be a realistic solution in many circumstances. For example, in cases where quick intervention by a neurotechnology is required to avert the onset of acute episodes or seizures, waiting for approval by a doctor would be highly inefficient and possibly harmful to the patient.

blanche" that is generally not accepted in medical ethics, where each new procedure or significant deviation from an agreed treatment plan normally requires specific consent from the patient. Importantly, unlike open-loop devices that deliver fixed stimulation set by a doctor, a closed-loop neurotechnology autonomously selects and adapts neurostimulation in real time on the basis of opaque—to both the doctor and the patient—algorithmic computations, allowing neurostimulation to diverge quickly and significantly from what was originally consented to. For a closed-loop device, adaptive stimulation decisions may thus sometimes be best understood as new interventions, rather than mere execution of a prior human plan, and can therefore cross rights boundaries.[12]

It thus seems that closed-loop neurotechnologies, if they qualify as moral agents, can pose a serious threat to the mind of the user by potentially infringing her mental rights. Yet, the extent of the threat might not depend only on its status as moral agent, but also on the extent to which the agency of the device remains sufficiently distinct from the user's agency. In the next section, we explore different ways in which closed-loop neurotechnologies could become involved in the agency of the person in whom they are implanted, and consider the implications for the possibility of mental rights infringement by the neurotechnology.

AGENTIAL RELATIONS AND MENTAL RIGHTS INFRINGEMENTS

On the view that closed-loop neurotechnologies can possess the kind of moral agency required to infringe another's rights, they could threaten the user's mental rights. However, the same AI-component of closed-loop neurotechnologies that seems integral to seeing them as moral agents capable of infringing mental rights might cast doubt on whether we can still appropriately view closed-loop neurotechnologies as *separate* agents when interacting with the brain—mainly in light of the role they may play in the agency of the user.

As well as exercising their own agency, closed-loop neurotechnologies may also influence the agency of the person in whom they are implanted. To illustrate, consider an example of a person exercising her agency by physically moving her body, such as moving her arm to wave at an acquaintance. There are neuronal mechanisms that underlie such an action; neurons in the motor cortex become active—are excited or inhibited—and send a signal down to the spinal cord, which is subsequently relayed via motor neurons to the muscles in the arm causing them to contract, leading to movement of the arm. Usually, the initial neuronal activation in the motor cortex leading to the muscle movement is caused by a cascade of earlier neural states initiated by the visual perception of the acquaintance.[13] Now, a closed-loop neurotechnology could initiate the same movement in the person, but without the preceding cascade of biologically-induced neural signaling that would otherwise have been involved. By stimulating neurons, the device could induce the activation of neurons in the motor cortex that results in the person forming the intention to move her arm, with this intention then causing the arm movement in the usual way. The electrical stimulation by the closed-loop neurotechnology would signify it exercising its agency, but by doing so, it also plays a role in the agency of the person—it causes the human agent to perform an action and exercise her agency.[14]

This raises questions regarding the agential relationship between a closed-loop neurotechnology and the person in whom it is implanted. Are there two separate agents inside one body? Does the agency of the person extend to the neurotechnology, the former subsuming the agency of the latter, becoming one single agent? Or does there arise some sort of hybrid agency between the person and the neurotechnology, where they perform actions together but are still two separate agents? The kind of agential relationship that exists between the human user and the closed-loop neurotechnology has implications for the extent to which the neurotechnology is aptly seen as a separate

[12]It may seem unusual that, unlike in traditional cases where a patient typically consents to the moral agent who might infringe her rights (e.g., the surgeon), in closed-loop cases the patient provides consent to the doctor and to implantation, but not also to the putative rights-threatening agent—namely, the autonomous neurotechnology. However, valid consent usually authorizes specific interventions within constraints, not only a particular person—the patient authorizes the physician to deploy the device within these constraints, and the physician in turn delegates execution to the device, but the device need not be part of the consent transaction itself.

[13]We will leave open whether this cascade involves neural *and* mental states, as we assume the widely held physicalist view of the mind that considers mental states to either supervene on or be identical to neural states.

[14]Note that our focus will be on the class of actions that the standard theory of action generally focuses on, which are actions that involve overt bodily movements. This means we will leave aside the class of actions referred to as mental actions (Fiebich and Michael 2015), as there is some discussion as to whether these actions can be seen as actions proper within the standard theory (see, for instance, Ruben and Philosophy Documentation Center 1995; Strawson 2003), as well as omissions (Clarke 2014).

moral agent capable of mental interference that might infringe mental rights. Below, we will explore three potential ways to view this relationship, and consider their implications for the possibility of mental rights infringement by the neurotechnology.

Two Separate Agents

The first sort of relationship that may exist between the closed-loop neurotechnology and the user is a two-agent relationship, where the closed-loop neurotechnology and the user remain two separate (moral) agents that perform separate acts. The neurotechnology and the user are stand-alone agents that can act by producing changes in their environment. The neurotechnology can act by, for instance, stimulating certain neurons, which is based on algorithmic computations involving antecedent states such as a representational state ("belief" that stimulation is necessary to prevent or promote a certain mental state) and a motivational state ("intention" to stimulate in a certain way). Such neurostimulation might result in mental states—representational and motivational states—in the user that cause her to act. This seems to be how we generally consider closed-loop neurotechnologies to work: a loop between two separate (agential) entities, the neurotechnology and the person (or their brain).

If the closed-loop neurotechnology and the user are separate agents acting separately, the closed-loop neurotechnology can interfere with the user's mental states just as another person could. If such interferences go beyond consent, the neurotechnology can infringe the user's mental rights. For example, in the case of patient X, the dopaminergic intervention and subsequent mental alterations by the closed-loop neurotechnology would constitute an interference with X's mind, and if not consented to, this could constitute a mental rights infringement by the neurotechnology.

One Agent

A second sort of relationship is a one-agent relationship, where the closed-loop neurotechnology is viewed as *part of* the human user's agency. What might motivate such a view is that, certainly in cases in which a neurotechnology is very actively involved in determining the mental states of the user, the neurotechnology may lie at the root of many actions performed by the user. When a neurotechnology creates many "new" neural states that lead—in a closed loop—to a multitude of mental states—representational and motivational states—the neurotechnology seems to fulfill a role similar to "input" processes within the brain—for example, inputs from the visual cortex. In cases where the neurotechnology is doing much of the causal work in "co-producing" the user's mental states and actions, it might be that it is no longer possible to clearly discern between agential processes of the neurotechnology and the user. Consequently, one might say that the agency of the user *extends* to the neurotechnology—and they together should be viewed as one agent.

Kellmeyer et al. (2016) seem to acknowledge this possibility of agency extension when they suggest that closed-loop neurotechnologies that implement AI learning systems "should be considered to have at least a version of autonomy" and that we therefore "must ask ourselves whether the autonomy of the human part of the BCI system should be extended to include the algorithmic part" (p. 626).[15] Moreover, with regard to neurotechnological behavioral control, Glannon (2017) argues that "the shared behavior control between the conscious subject and the artificial device is not fundamentally different from the shared behavior control between the conscious subject and naturally occurring unconscious processes in his normally functioning brain" (2017, p. 323). Such cases of (seamless) shared control of actions may well support the view that the neurotechnology and the user are one agent rather than two separate ones.[16]

If the closed-loop neurotechnology and the user are one agent, the neurotechnology is no longer capable of infringing the user's mental rights, because it can no longer be viewed as a *separate* moral agent. If the neurotechnology is part of the user's agency, it would imply that any mental effects that the neurotechnology causes that might have "ordinarily" been construed as interference would have to be treated on par with mental effects caused by "regular" neuronal events. Such effects could perhaps still be

[15]While the authors speak of "autonomy," autonomy generally presupposes the more "basic" component of agency, so we take this claim to also imply agency.

[16]Against this one-agent view, one could argue that actions caused by neurotechnology-induced mental states are not proper actions because they involve deviant causal chains, thus problematizing viewing them as one agent. That is, it could be argued that in cases where a "mental" state by the neurotechnology causes a person to act, this is not a proper action because it is not caused by the "right" kind of mental state—i.e., her own mental state. Therefore, what might be required for accepting this one-agent view is assuming that the closed-loop neurotechnology is also in some way part of the mind, in the sense that its "mental" states can also be attributed to the person. While some have argued for the emergence of "hybrid minds" between neurotechnologies and persons (Bublitz et al. 2022; Soekadar et al. 2021), the question of mind extension is one that is part of a large metaphysical debate in philosophy of mind that we will not broach here.

considered mental interference—assuming that one part of an agent can interfere with another—but not infringements of mental rights, as in such cases the moral agent supposedly infringing mental rights coincides with the moral agent that is the holder of those rights. As mentioned earlier, it is generally assumed that rights against mental interference are held only against *others*. Moreover, even if such rights are held also against oneself, it is plausible that in performing an action that would otherwise infringe the right, one implicitly waives the right (Cholbi 2015; Muñoz 2024). So, this would mean that in the case of patient X, inducing the drug-like euphoric state by neurostimulation would not be any more of a mental rights infringement than a memory popping into one's mind without this being sought.

Hybrid agency

A third possibility, one that rejects the one-agent view but still appreciates the intimate role that the closed-loop neurotechnology can play in the user's agency, is that the user and the neurotechnology display something like "hybrid agency." Hybrid agency would mean that they are both separate agents *and* (sometimes) part of a joint agent, where the closed-loop neurotechnology and the user can exercise their agency independently of each other, but occasionally their agency also intertwines so that they act as one agent.

Some in the literature on closed-loop neurotechnologies explicitly discuss the possibility of hybrid agency. For example, Rainey and Erden (2020) argue that the way closed-loop neurotechnologies affect our agency may promote a "kind of hybrid control at play" (p. 2443). Others have similarly endorsed such a "cooperative" form of agency between neurotechnologies and persons. Goering et al. (2021) refer to this as "relational agency." They argue that agency in general has a relational nature that "often involves receiving uptake for one's intentions and having assistance in enacting one's intentions," and that we may learn to consider neurotechnologies "as part of a co-agency, or as a supportive helper aiming to help enact and make her intention legible" (Goering et al. 2021, p. 6). Soekadar and colleagues make similar claims but with a focus on cognition, as they argue that interaction between AI-infused neurotechnologies (specifically brain-computer interfaces or "BCIs") and the brain "generates a hybrid cognitive system that runs on, or is fed by inputs from, the organic hardware of the brain as well as the AI implementing BCI" (2023, p. 77).

It is not obvious, however, what such a relationship would exactly look like. Hybrid agency between a closed-loop neurotechnology and a person might resemble something like what is called "shared" or "collective" agency in the literature, which refers to when two or more agents can act together to collectively perform actions (Misselhorn 2015). Hakli and Mäkelä (2019), for instance, suggest that "human beings and technological artifacts acting in interaction can form so-called hybrid agents, which could be seen as collective agents satisfying the conditions of moral agency" (p. 261). Shared or collective action generally involves agents collectively completing actions by having "joint intentions" (Ekins 2012). Whether closed-loop neurotechnologies and users can have joint intentions, however, is not clear, as this seems to require some form of awareness of the intentions of the other (Bratman 1993) that the user is unlikely to have of the intentions of the closed-loop neurotechnology due to its opaque decision-making processes.[17]

The analogy with collective agency might still be useful, however, to consider the implications of hybrid agency for mental rights infringements by closed-loop neurotechnologies. A characteristic feature of collective agency is that distinct agents—agents in their own right—can act collectively as one agent in some instances. We might think that when the neurotechnology and the user act as a collective, or one agent, effects arising out of integrated agential control might not count as the neurotechnology infringing the user's rights—similar to what we saw in the one-agent relationship. Still, in those instances that the neurotechnology and the user exercise their agency separately, this would seemingly allow for the neurotechnology to infringe the user's mental rights—similar to what we saw in the two-agent relationship. However, distinguishing between cases of acting separately and acting together might not be easy. The hybrid agency view challenges the traditional idea of clearly individuated agents and therefore does not invoke clear intuitions regarding rights infringements. Returning to patient X, if the closed-loop neurotechnology and X together constitute a hybrid agent, whether or not the

[17]One could argue that when a neurotechnology acts in a closed loop, constantly anticipating and adapting to neural and mental states of the user, there might be a sense in which the neurotechnology can be "aware" of the intentions of the user. However, even if we were to assume that closed-loop neurotechnologies could be "aware" in the right way, there is unlikely to be awareness of the neurotechnology's intentions by the user due to its non-transparent computational processes, and reciprocal awareness seems essential for an intention to be considered properly "shared" or "joint" (Bratman 1993).

dopaminergic intervention by the neurotechnology would infringe her mental rights will depend on whether or not they were acting as separate agents or as one agent.

Implications for Rights-Based Protection of the Mind

The applicability of rights over the mind (implicitly) presupposes a stance on agential distinctness. Rights are ordinarily held against *others*, and when the neurotechnology and user act as two agents, the "otherness" condition is met and there is a distinct moral agent capable of infringing a right. When there is more agential integration, such as on the hybrid or one-agent views, that otherness can diminish or entirely disappear, and with it the applicability of rights held against others (assuming that no other person is infringing the user's rights). The point is not merely a semantic one, but it ties the normative category of rights to an ontological view of agential distinctness (or how many agents are in play at a particular time). This suggests that questions about the applicability of mental rights in cases of closed-loop neurotechnology cannot be resolved by rights theory alone; we also need an account of agency and how to individuate agents. The current discussion thus makes explicit that the applicability of rights tracks agential distinctness, not just harm or location in the brain.

This matters practically, because new technologies such as closed-loop neurotechnologies might increasingly defy the usual agent-tool categorizations. Especially the AI component that may confer moral agency on closed-loop neurotechnologies challenges the "standard" rights application generally seen in medicine. Traditionally, there is a clearly distinct human agent (the doctor) who initiates an intervention, so rights apply straightforwardly to that agent, but this is less straightforward for adaptive closed-loop neurotechnologies that select interventions without human direction and that might become integrated in a user's agency. As a result, rights may be clearly applicable when the device operates as a distinct agent, clearly inapplicable when the device functions as part of the agent and have uncertain applicability when agency is shared.

As variations in agential distinctness do not fit neatly into traditional, person-centered rights analysis, there may be a risk of both over-protection and under-protection of the mind (insofar as protection is solely rights-based). That is, treating every intervention as a potential mental rights threat can yield over-protection—e.g., misplaced accusations when there is no agential distinctness—whereas treating no interventions as such risks under-protection when there might be (full) agential distinctness and rights should be triggered. This risk seems most acute in hybrid-agency cases, where it is unclear whether, and when, the user and device are acting as one or as separate agents. Moreover, determining whether the neurotechnology can itself infringe rights is normatively relevant for downstream responsibility: if the device is a distinct rights-infringing agent, then a doctor's or user's knowing failure to intervene can ground complicity-based responsibility, whereas where no such separate agent exists (e.g., under agential integration), complicity does not apply. Examining when rights genuinely apply in light of agential relations—and when other safeguards should take over—could help preserve meaningful protection of the mind.

CONCLUSION

Closed-loop neurotechnologies hold potential for treating some neurological and psychiatric disorders. However, those that employ sophisticated AI algorithms may raise worries regarding mental rights infringements, since the devices might cause mental effects in users that are not covered by initial consent. For such a mental rights infringement to be perpetrated by the AI system itself, it is plausibly required that the AI system is a moral agent, and one that is separate from the human user. Therefore, we first explored whether these closed-loop AI systems can be moral agents. As it turns out, on several accounts of moral agency, they can. This implies that, on such accounts, closed-loop neurotechnologies implanted in people's brains are in principle capable of infringing mental rights.

However, the close involvement of closed-loop neurotechnologies in the agency of the human user in whom they are implanted might suggest that they may not always be sufficiently distinct from the user to be *separate* moral agents—which in turn could undermine their capacity to infringe mental rights. That is, closed-loop neurotechnologies can become intimately involved in modifying and creating the mental and neural states that underlie a person's actions, and therefore potentially integrated with the agency of the person to such an extent that we should view them as together becoming a "hybrid" agent or even one single agent. If so, the closed-loop neurotechnology may no longer be capable of infringing

the person's mental rights due to it no longer being a separate moral agent toward which such rights apply. Thus, it appears that whether the risk of mental rights infringements due to the involvement of an AI component in closed-loop neurotechnologies is realized depends to a considerable degree on the nature of the agential relationship between the neurotechnology and the human user. The ontological framing of this agential relationship might therefore be of relevance to shaping safeguards to protect users' minds in the context of neurotechnology use.

CONTRIBUTORS

VT is the main author of the article. TD, LF, SL and GM contributed to the drafting of the initial manuscript and commented on later versions of the manuscript. All authors approved the final manuscript. GM acts as guarantor.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

FUNDING

This research is funded by the Dutch Research Council (NWO) (grant number VI.C.201.067), the European Research Council Consolidator Award (grant number 819757) and the Uehiro Foundation on Ethics and Education.

ORCID

Vera Tesink http://orcid.org/0000-0003-2293-8860
Thomas Douglas http://orcid.org/0000-0002-6788-3608
Lisa Forsberg http://orcid.org/0000-0002-5239-393X
Sjors Ligthart http://orcid.org/0000-0001-6458-4058
Gerben Meynen http://orcid.org/0000-0001-7298-8407

REFERENCES

Abbott R, Sarch A. 2024. Punishing artificial intelligence: legal fiction or science fiction. In: Moura Vicente D, Soares Pereira R, Alves Leal A, editors. Legal aspects of autonomous systems. Springer International Publishing. p. 83–115. https://doi.org/10.1007/978-3-031-47946-5_6

Alegre S. 2017. Rethinking freedom of thought for the 21st century. Eur Hum Rights Law Rev. 3:13.

Asaro PM. 2011. A body to kick, but still no soul to damn: legal perspectives on robotics. In: Lin P, Abney K, Bekey GA, editors. Robot ethics ethical social implications robotics. MIT Press. p. 169–186.

Behdadi D, Munthe C. 2020. A normative approach to artificial moral agency. Minds & Machines. 30(2):195–218. https://doi.org/10.1007/s11023-020-09525-8

Belkacem AN, Jamil N, Khalid S, Alnajjar F. 2023. On closed-loop brain stimulation systems for improving the quality of life of patients with neurological disorders. Front Hum Neurosci. 17:1085173. https://doi.org/10.3389/fnhum.2023.1085173

Bishop JC. 1989. Natural agency: an essay on the causal theory of action. Cambridge University Press.

Bratman M. 1987. Intention, plans, and practical reason. Harvard University Press.

Bratman ME. 1993. Shared intention. Ethics. 104(1):97–113. https://doi.org/10.1086/293577

Brożek B, Janik B. 2019. Can artificial intelligences be moral agents? New Ideas Psychol. 54:101–106. https://doi.org/10.1016/j.newideapsych.2018.12.002

Bublitz C, Chandler J, Ienca M. 2022. Human–machine symbiosis and the hybrid mind: implications for ethics, law and human rights. In: Ienca M, Pollicino O, Liguori L, Stefanini E, Andorno R, editors. The Cambridge handbook of information technology, life sciences and human rights. 1st ed. Cambridge University Press p. 286–303. https://doi.org/10.1017/9781108775038.024

Bublitz JC, Merkel R. 2014. Crimes against minds: on mental manipulations, harms and a human right to mental self-determination. Crim Law Philos. 8(1):51–77. https://doi.org/10.1007/s11572-012-9172-y

Bublitz J-C. 2020. The nascent right to psychological integrity and mental self-determination. In: von Arnauld A, von der Decken K, Susi M, editors. The Cambridge handbook of new human rights: recognition, novelty, rhetoric. Cambridge University Press. p. 387–403. https://doi.org/10.1017/9781108676106.031

Chakraborty A, Bhuyan N. 2024. Can artificial intelligence be a Kantian moral agent? On moral autonomy of AI system. AI Ethics. 4(2):325–331. https://doi.org/10.1007/s43681-023-00269-6

Cholbi M. 2015. On Marcus singer's "on duties to oneself". Ethics. 125(3):851–853. https://doi.org/10.1086/679554

Clarke R. 2014. Omissions: agency, metaphysics, and responsibility. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199347520.001.0001

Constantinescu M, Vică C, Uszkai R, Voinea C. 2022. Blame it on the AI? On the moral responsibility of artificial moral advisors. Philos Technol. 35(2):35. https://doi.org/10.1007/s13347-022-00529-z

Craig JN. 2016. Incarceration, direct brain intervention, and the right to mental integrity – a reply to Thomas Douglas. Neuroethics. 9(2):107–118. https://doi.org/10.1007/s12152-016-9255-x

Danaher J. 2016. Robots, law and the retribution gap. Ethics Inf Technol. 18(4):299–309. https://doi.org/10.1007/s10676-016-9403-3

Davidson D. 1980. Toward a unified theory of meaning and action. Grazer Philos Stud. 11(1):1–12. https://doi.org/10.1163/18756735-90000093

Douglas T. 2025. An intuitive, abductive argument for a right against mental interference. J Ethics. 29(1):133–154. https://doi.org/10.1007/s10892-024-09476-7

Douglas T. 2026. Protecting minds: the right against mental interference. Oxford University Press.
Ekins R. 2012. Joint intention and group agency. In: Ekins R, editor. The nature of legislative intent. Oxford University Press. p. 47–76. https://doi.org/10.1093/acprof:oso/9780199646999.003.0003
Fiebich A, Michael J. 2015. Mental actions and mental agency. Rev Philos Psych. 6(4):683–693. https://doi.org/10.1007/s13164-015-0289-5
Floridi L. 2013. Distributed morality in an information society. Sci Eng Ethics. 19(3):727–743. https://doi.org/10.1007/s11948-012-9413-4
Floridi L, Sanders JW. 2004. On the morality of artificial agents. Minds Mach. 14(3):349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d
Fossa F. 2018. Artificial moral agents: moral mentors or sensible tools? Ethics Inf Technol. 20(2):115–126. https://doi.org/10.1007/s10676-018-9451-y
Glannon W. 2017. Brain implants: implications for free will. In: The Routledge handbook of neuroethics. Routledge. p. 319–334.
Goering S, Brown T, Klein E. 2021. Neurotechnology ethics and relational agency. Philos Compass. 16(4):e12734. https://doi.org/10.1111/phc3.12734
Goering S, Klein E, Dougherty DD, Widge AS. 2017. Staying in the loop: relational agency and identity in next-generation DBS for psychiatry. AJOB Neurosci. 8(2):59–70. https://doi.org/10.1080/21507740.2017.1320320
Hakli R, Mäkelä P. 2019. Moral responsibility of robots and hybrid agents. Monist. 102(2):259–275. https://doi.org/10.1093/monist/onz009
Hallevy PG. 2010. The criminal liability of artificial intelligence entities. (SSRN Scholarly Paper 1564096). Social Science Research Network. https://doi.org/10.2139/ssrn.1564096
Haselager P. 2013. Did I do that? Brain–computer interfacing and the sense of agency. Minds Mach. 23(3):405–418. https://doi.org/10.1007/s11023-012-9298-7
Himma KE. 2009. Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics Inf Technol. 11(1):19–29. https://doi.org/10.1007/s10676-008-9167-5
Kellmeyer P. 2021. Big brain data: on the responsible use of brain data from clinical and consumer-directed neurotechnological devices. Neuroethics. 14(1):83–98. https://doi.org/10.1007/s12152-018-9371-x
Kellmeyer P et al. 2016. The effects of closed-loop medical devices on the autonomy and accountability of persons and systems. Cambridge Q Healthcare Ethics. 25(4):623–633. https://doi.org/10.1017/S0963180116000359
Klein E et al. 2016. Brain-computer interface-based control of closed-loop brain stimulation: attitudes and ethical considerations. Brain-Comput Interfaces. 3(3):140–148. https://doi.org/10.1080/2326263X.2016.1207497
Kohler F et al. 2017. Closed-loop interaction with the cerebral cortex: a review of wireless implant technology§. Brain-Comput Interfaces. 4(3):146–154. https://doi.org/10.1080/2326263X.2017.1338011
Lavazza A, Giorgi R. 2023. Philosophical foundation of the right to mental integrity in the age of neurotechnologies. Neuroethics. 16(1):10. https://doi.org/10.1007/s12152-023-09517-2
Ligthart S, Kooijmans T, Douglas T, Meynen G. 2021. Closed-loop brain devices in offender rehabilitation: autonomy, human rights, and accountability. Cambridge Q Healthcare Ethics. 30(4):669–680. https://doi.org/10.1017/S0963180121000141
Lima D. 2018. Could AI agents be held criminally liable? artificial intelligence and the challenges for criminal law. South Carolina Law Rev. 69:677–696. https://durham-repository.worktribe.com/output/1186493/could-ai-agents-be-held-criminally-liable-artificial-intelligence-and-the-challenges-for-criminal-law.
List C. 2021. Group agency and artificial intelligence. Philos Technol. 34(4):1213–1242. https://doi.org/10.1007/s13347-021-00454-7
Llorca Albareda J, García P, Lara F. 2023. The moral status of AI entities. In: Lara F, Deckers J, editors. Ethics of artificial intelligence. Springer Nature Switzerland. p. 59–83. https://doi.org/10.1007/978-3-031-48135-2_4
Manna R, Nath R. 2021. The problem of moral agency in artificial intelligence. In: 2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW). p. 1–4. https://doi.org/10.1109/21CW48944.2021.9532549
Matthias A. 2004. The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol. 6(3):175–183. https://doi.org/10.1007/s10676-004-3422-1
Mayr E. 2011. Deviant causal chains. In: Mayr E, editor. Understanding human agency. Oxford University Press. p. 104–141. https://doi.org/10.1093/acprof:oso/9780199606214.003.0006
McCarthy-Jones S. 2019. The autonomous mind: the right to freedom of thought in the twenty-first century. Front Artif Intell. 2:19. https://doi.org/10.3389/frai.2019.00019
Mele AR. 2017. Actions, explanations, and causes. In: Mele AR, editor. Aspects of agency: decisions, abilities, explanations, and free will. Oxford University Press. p. 27–62. https://doi.org/10.1093/acprof:oso/9780190659974.003.0003
Mellema G. 2016. Complicity and moral accountability. University of Notre Dame Press. https://doi.org/10.2307/j.ctvpj78ss
Misselhorn C. 2015. Collective agency and cooperation in natural and artificial systems. In: Misselhorn C, editor. Collective agency and cooperation in natural and artificial systems: explanation, implementation and simulation. Springer International Publishing. p. 3–24. https://doi.org/10.1007/978-3-319-15515-9_1
Mulligan C. 2017. Revenge against robots. (SSRN Scholarly Paper 3016048). Social Science Research Network. https://doi.org/10.2139/ssrn.3016048
Muñoz D. 2024. Obligations to oneself. In: Zalta EN, Nodelman U, editors. The Stanford encyclopedia of philosophy. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2024/entries/self-obligations/
Parastarfeizabadi M, Kouzani AZ. 2017. Advances in closed-loop deep brain stimulation devices. J Neuroeng Rehabil. 14(1):79. https://doi.org/10.1186/s12984-017-0295-1
Parthemore J, Whitby B. 2013. What makes any agent a moral agent? Reflections on machine consciousness and moral agency. Int J Mach Conscious. 5(2):105–129. https://doi.org/10.1142/S1793843013500017
Purves D, Jenkins R, Strawser BJ. 2015. Autonomous machines, moral judgment, and acting for the right reasons.

Ethic Theory Moral Pract. 18(4):851–872. https://doi.org/10.1007/s10677-015-9563-y

Rainey S, Erden YJ. 2020. Correcting the brain? The convergence of neuroscience, neurotechnology, psychiatry, and artificial intelligence. Sci Eng Ethics. 26(5):2439–2454. https://doi.org/10.1007/s11948-020-00240-2

Ratoff W. 2024. The right to mental autonomy: its nature and scope. J Ethics Soc Philos. 27(2):257–286. https://doi.org/10.26556/jesp.v27i2.2907

Ruben D-H, Philosophy Documentation Center. 1995. Mental overpopulation and the problem of action. J Philos Res. 20:511–524. https://doi.org/10.5840/jpr_1995_14

Schlosser M. 2019. Agency. In: Zalta EN, editor. The Stanford encyclopedia of philosophy. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

Schopp L, Starke G, Ienca M. 2025. Clinician perspectives on explainability in AI-driven closed-loop neurotechnology. Sci Rep. 15(1):34638. https://doi.org/10.1038/s41598-025-19510-9

Shaw E. 2022. Neuroscience, criminal sentencing, and human rights. William Mary Law Rev. 63:36.

Soekadar S, Chandler J, Ienca M, Bublitz C. 2021. On the verge of the hybrid mind. Morals Mach. 1(1):30–43. https://doi.org/10.5771/2747-5174-2021-1-30

Soekadar SR et al. 2023. Future developments in brain/neural–computer interface technology. In: V Dubljević, Coin A, editors. Policy, identity, and neurotechnology: the neuroethics of brain-computer interfaces. Springer International Publishing. p. 65–85. https://doi.org/10.1007/978-3-031-26801-4_5

Sparrow R. 2007. Killer robots. J Applied Philosophy. 24(1):62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Steinert S, Bublitz C, Jox R, Friedrich O. 2019. Doing things with thoughts: brain-computer interfaces and disembodied agency. Philos Technol. 32(3):457–482. https://doi.org/10.1007/s13347-018-0308-4

Strawson G. 2003. XI—Mental ballistics or the involuntariness of spontaneity. Proc Aristotelian Soc. 103(1):227–256. https://doi.org/10.1111/j.0066-7372.2003.00071.x

Sullins JP. 2006. When is a robot a moral agent. Int Rev Inf Ethics. 6(12):23–30.

Sullivan CRP, Olsen S, Widge AS. 2021. Deep brain stimulation for psychiatric disorders: from focal brain targets to cognitive networks. Neuroimage. 225:117515. https://doi.org/10.1016/j.neuroimage.2020.117515

Symons J, Abumusab S. 2024. Social agency for artifacts: chatbots and the ethics of artificial intelligence. Digital Soc. 3(1):1–28. https://doi.org/10.1007/s44206-023-00086-8

Valeriani D, Santoro F, Ienca M. 2022. The present and future of neural interfaces. Front Neurorobot. 16:953968. https://doi.org/10.3389/fnbot.2022.953968

Vukov JM. 2017. Three kinds of agency and closed-loop neural devices. AJOB Neurosci. 8(2):90–91. https://doi.org/10.1080/21507740.2017.1320324

Warren MA. 2000. Personhood and moral rights. In: Moral status: obligations to persons and other living things. Oxford University Press. p. 90–121.

Widge AS. 2023. Closed loop deep brain stimulation for psychiatric disorders. Harv Rev Psychiatry. 31(3):162–171. https://doi.org/10.1097/HRP.0000000000000367

Wolpaw J, Wolpaw EW. 2012. Brain-computer interfaces: principles and practice. Oxford University Press.

Zhu B, Shin U, Shoaran M. 2020. Closed-loop neural interfaces with embedded machine learning. In: 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS). https://doi.org/10.1109/ICECS49266.2020.9294844

Ziemke T. 2023. Understanding social robots: attribution of intentional agency to artificial and biological bodies. Artif Life. 29(3):351–366. https://doi.org/10.1162/artl_a_00404

Zuk P. 2024. Mental integrity, autonomy, and fundamental interests. J Med Ethics. 50(10):676–683. https://doi.org/10.1136/jme-2023-109732