# 25

# AI-Supported Brain–Computer Interfaces and the Emergence of 'Cyberbilities'

*Boris Essmann and Oliver Mueller*

## I. INTRODUCTION

Recent advances in brain–computer interfacing (BCI) technology hold out the prospect of technological intervention into the basis of human agency to supplement and restore functioning in agency-limited individuals and even augmenting and enhancing capacities for natural agency. By increasingly using Artificial Intelligence (AI), for example machine learning methods, a new generation of brain–computer interfaces aims to advance technological possibilities to intervene into agentive capacities even more, creating new forms of human–machine interaction in the process. This trend further accentuates concerns about the impact of neurotechnology on human agency, not only regarding far-reaching visions like the media-effective propositions by *Elon Musk* (Neuralink) but also with respect to current developments in medicine. Because these developments could be understood as (worrisome) 'fusions' of human, machinic, and software agency we investigate neurotechnology and AI-assisted brain–computer interfaces by directly focusing on agentive dimensions and potential changes of agency in these types of interactions. By providing a philosophical discussion of these topics we aim to capture the broad impact of this technology on our future and contribute valuable perspectives on its ethically and socially relevant dimensions. Although we adopt a philosophical approach, we do not restrict ourselves to a single disciplinary perspective, such as an exclusively ethical or neuroscience-oriented analysis. Given the potential to fundamentally reshape our individual and collective lives, the combination of neurotechnology and AI-technology may well create challenges that exceed disciplinary boundaries and which, therefore, cannot be met by a single discipline.

Our contribution to discussing the 'fusion' of human and artificial agency is the introduction of two neologisms – cyberbilities and hybrid agency – which we understand as concepts that integrate a range of disciplinary perspectives on this phenomenon. At a fundamental level, the concept loosely draws on *Amartya Sen*'s and *Martha Nussbaum*'s capabilities approach, but retools the notion of capabilities to analyze intricate human–machine interactions. We specifically adopt the normative core of capabilities – the ethical value of well-being opportunities – as a conceptual tool to evaluate risks and benefits of AI-supported brain–computer interfaces. However, like capabilities, cyberbilities presuppose a concept of human agency. Therefore, devising this concept requires a clarification of the underlying understanding of agency. Furthermore, because cyberbilities involve agency that is assisted by neurotechnology, we will also include an analysis of the various interactions between human and non-human elements involved.



https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

This chapter is divided into three main sections. In the first section, we present conceptual expositions of the terms capabilities, agency, and human–machine interaction which serve both as an illustration of the complex nature of BCI technology and some necessary background to motivate the following line of argument.[1] This section is not intended to exhaust the topic from a specific (e.g., ethical or neuroscientific) perspective, but rather to amalgamate three very different but – as we maintain – complementary approaches. Specifically, we draw on the work of capability theorists such as *Sen* and *Nussbaum*.[2] Also, since neurotechnology affects human agency on various levels, we discuss the notions of agency and human–machine interaction from the perspectives of neuroscience, philosophical action theory, and a sociological framework.[3] In the next section, we introduce the above-mentioned novel concepts of hybrid agency and cyberbilities which combine our preceding line of argument and denote new forms of agency resulting from 'agentive' technologies.[4] A cyberbility is a type of capability, in other words, it is a normative concept designed to gauge the various ways in which neurotechnology can lead to achievements of (or want of) well-being and contribute to (or detract from) human flourishing. In the last section, we propose a list of cyberbilities that illustrates ways in which neurotechnology can lead to well-being gains (or losses) and explores the personal, social, and political ramifications of neurotechnologically assisted (or, in our terms, hybrid) agency.[5] However, this list of cyberbilities should not be understood as a conclusive result of the preceding conceptual work, but rather as a tentative and incomplete catalogue of core claims and requirements that reflect how new kinds of technologies challenge our established understanding of agency and human–machine interaction. In this sense, we see the list of cyberbilities not as a completed ethical evaluation, but as a foray into mapping tentative points of normative orientation.[6] And finally, we want to discuss a potential objection regarding our approach.[7]

## II. FROM CAPABILITIES TO CYBERBILITIES

Let's start by anticipating our definition of cyberbilities: Cyberbilities are capabilities that originate from hybrid agency (i.e. human–machine interactions), in which agency is distributed across human and neurotechnological elements. As we will lay out in the following sections, this definition emphasizes that cyberbilities are embedded not only in personal aspects of agency, but also in a social environment that is shaped by the 'logic' of the respective technology and the institutions that deploy it (i.e. the 'technological condition').

In order to provide the necessary background for the notion of cyberbilities, we shall proceed in three steps. Firstly, we will briefly unfold in which way we retool the capabilities approach for our own purposes. Secondly, we argue that we need to revisit the concept of agency concerning its use in neuroscience and philosophy if we want to reliably describe the complex interactions between human and artificial elements, especially in the context of brain–computer interfaces. Lastly, we will draw on the notion of distributed agency introduced by sociologist *Werner Rammert*[8] to illuminate how technology affects agency and, consequently, human–machine

[1] See Section II.
[2] See Sub-section II 1.
[3] See Sub-section II 2.
[4] See Section III.
[5] See Sub-section IV 1.
[6] See Sub-section IV 2.
[7] See Section V.
[8] W Rammert, 'Where the Action Is: Distributed Agency between Humans, Machines, and Programs' in U Seifert, JH Kim, and A Moore (eds) *Paradoxes of Interactivity* (2008) (hereafter Rammert, 'Distributed Agency').

https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

interactions. All three steps serve to review current disciplinary views on the topics at hand and prepare our proposal of an extended and integrated perspective in Section III.

## 1. Capabilities

The capabilities approach, first introduced by *Sen*[9] and extended by *Nussbaum*[10], is a theoretical framework used in a number of fields to evaluate the well-being of individuals in relation to their social, political, and psychological circumstances. To capability theorists, each person can be described (and thus compared) in terms of their 'capabilities' to achieve and maintain well-being, and any restrictions of those capabilities are subject to ethical scrutiny. As a philosophical term, well-being does not mean, for example, happiness, wealth, or absence of negative emotions or circumstances. Rather, well-being is meant to encompass how well a person's life is going overall, not just in relation to available means to lead a comfortable life or to achieve temporary positive emotional states, but concerning that a person is understood as an end when we focus on the opportunities to lead a good life that are available to each person.[11]

There is a long history of debate on the capabilities approach, and *Sen* and *Nussbaum* themselves delivered further refinements of the approach. We are aware of the fact that there are a number of controversies and open questions, for example, that *Sen*'s account is overly individualistic[12], or regarding certain essentialist traits[13] of *Nussbaum*'s version of the capabilities approach. However, due to the explorative purpose of this paper, we do not want to engage in further discussions of these aspects. Rather, we draw on *Sen*'s and *Nussbaum*'s theories in a pragmatic way, adopting some of their core elements in order to develop a basis for our tentative list of cyberbilities, which we see as a conceptual means not only to grasp novel kinds of agency in the upcoming age of human–machine fusions but also to propose a perspective that could help to evaluate these human–machine mergers as well.

But what are capabilities? Loosely following *Sen*, a capability describes what a person is actually able to be and do to increase her well-being. To capability theorists, 'the freedom to achieve well-being is of primary moral importance'[14], and can therefore be used to evaluate if a person's social, political, and developmental circumstances support or hinder her well-being. In more technical terms, a capability is the real opportunity (or freedom) to achieve functionings, where functionings are beings and doings (or states) of a person, like 'being well-nourished' or 'taking the bus to work'. Both capabilities and functionings are treated as a measure of a person's well-being, and therefore allow us to compare people in terms of how well their life is going. They are distinguished, however, from resources like wealth or commodities, because those metrics arguably provide only limited or indirect information about how well the life of a person is going.

[9] E.g., A Sen, *Commodities and Capabilities* (1985) and A Sen, *Development as Freedom* (2001); as an introduction also cf. A Sen, 'Development as Capability Expansion' (1989) 19 *Journal of Development Planning* 41–58.

[10] E.g., M Nussbaum, *Women and Human Development: The Capabilities Approach* (2001) (hereafter Nussbaum, 'Capabilities Approach'); as an introduction cf. M Nussbaum, *Creating Capabilities: The Human Development Approach* (2013) (hereafter Nussbaum, 'Creating Capabilities').

[11] Nussbaum, 'Creating Capabilities', 18.

[12] C Gore, 'Irreducibly Social Goods and the Informational Bias of Amartya Sen's Capability Approach' (1997) 9(2) *Journal of International Development* 235–250.

[13] SM Okin, 'Poverty, Well-being, and Gender: What Counts, Who's Heard?' (2003) 31(3) *Philosophy & Public Affairs* 280–316.

[14] I Robeyns and M Fibieger Byskov, 'The Capability Approach' (2020) *The Stanford Encyclopedia of Philosophy Winter 2020 Edition* https://plato.stanford.edu/archives/win2020/entries/capability-approach.

https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

*Nussbaum* further developed the capabilities approach, specifically by extending the scope of *Sen*'s pragmatic and result-oriented theory.[15] For her, a functioning is 'an active realization of one or more capabilities (. . .). Functionings are beings and doings that are the outgrowths or realization of capabilities.'[16] *Nussbaum* stresses that she does not intend to deliver a theory on human nature as such. But she does understand the capabilities approach as an inherently evaluative and ethical theory that focuses on valuable capacities that human beings have reason to value and that a just society is obligated to nurture and support.[17] The normative criterion for valuableness is well-being as well (although quality of life or human flourishing are sometimes used synonymously). According to *Nussbaum*'s ambitious theory, the development of capabilities is connected to the notions of freedom (like in *Sen*'s theory) and dignity (by which she is going beyond *Sen*); she states: 'In general (. . .) the Capabilities Approach, in my version, focuses on the protection of areas of freedom so central that their removal makes a life not worthy for human dignity.'[18] Against this background *Nussbaum* famously compiled a list with ten central capabilities, ranging from life, bodily health, bodily integrity, up to the affiliation with others and the political and material control over one's environment.[19]

Our conception of cyberbilities shares not only *Sen*'s focus on well-being and functionings, but also *Nussbaum*'s idea to provide a list with core cyberbilities. However, we understand our list not as a substitution, but a supplement to *Nussbaum*'s, taking into account that AI-based brain–computer interfaces might change our understanding of both capabilities and agency.

Our reasoning is that modern technology is so complex and closely connected to human agency and well-being that it has the potential not only to subvert, but also to strengthen capabilities in complex ways. This relation will only become more intricate as neurotechnology and AI become more elaborate and integrated in our bodies, especially with human–machine fusions promised by future BCI technologies. Simply asking if such technologies contribute to or detract from well-being, or contradict or strengthen central capabilities, might be undercut by the impact they have on human agency as a whole. We could overlook subtle but unpreferable effects on agency if a technology grants certain well-being benefits, or miss beneficial effects on flourishing, for example in the case of capability-tradeoffs[20] realized by new types of technologically-assisted agency.

For this reason, we argue that evaluating current and future neurotechnology on the basis of the capabilities approach alone might fall short. Instead, we propose to combine the well-being and functioning focus of the capability approach with an extended perspective on agency that is tailored to identifying the impact of neurotechnology and AI on human agency as a whole. The specific challenge is that neurotechnological devices are not just another type of tool that human beings can use as an external means to realize capabilities and achieve well-being. By intervening into the brain of a person, neurotechnology interacts intimately with the basis of human agency, which opens the possibility to affect agency and capabilities in unforeseen ways. Because we may not be able to predict if this new kind of interaction relates positively or negatively to those dimensions, it seems prudent to develop a perspective that may accompany the coming neurotechnological developments with ethical scrutiny.

[15] Nussbaum, 'Creating Capabilities' (n 10).
[16] Ibid, 25.
[17] Ibid, 28.
[18] Ibid, 31.
[19] Ibid, 33–34.
[20] Cf. Section V.

https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

Hence, cyberbilities are an extension of the core tenets of the capability approach insofar as they are capabilities that arise from agency that is already enabled or affected by neuro- and/or AI-technology.

## 2. *Agency and Human–Machine Interactions*

After having briefly introduced the notion of capabilities we now focus on the conceptions of agency and human–machine interaction. This section will work towards an understanding of the ways in which human agency intersects and merges with machinic and software agency in technological contexts, a phenomenon which sociologist *Rammert* calls distributed agency.[21] The concept of hybrid agency, which we introduce in Section III, is a specific type of distributed agency which is also the core of the notion of a cyberbility.

There are two dimensions we consider to be central to human–machine interaction in general, and human–computer interaction in particular: Firstly, the causal efficacy of intentions, in other words, the idea that human intentions are the causal origin of technologically mediated actions, and secondly, the social aspect of acting in a technological context, especially when interacting with technological devices. We review established views on both agency and human–machine interaction in the context of BCI operation[22] and then go on to discuss these views in more depth.[23] While these two dimensions by no means exhaust the spectrum of relevant aspects in human–machine interaction, we see them as instructive starting points to develop our extended view that leads to introducing the novel concepts of hybrid agency and cyberbilities.

### a. The 'Standard View': Compensating Causality and Interactivity

Philosophically speaking, the concept of agency is connected with the phenomenon of intentionality and intention. An intention is a specific type of mental state that aggregates other action-related mental states (such as beliefs and desires), representing a concrete goal or plan and adding a stable commitment to actually perform actions aimed at realizing the respective goal or plan.[24] Theories that explain how intentions work conceptually are numerous[25], but the so-called standard view is that intentions govern and direct behavior through their specific causal efficacy.[26] In other words, intentions govern behavior by virtue of their direct and indirect causal effects on the chain of events from mental states to the execution of movements.[27] Hence, saying that a person 'has agency' amounts to saying that his intentions causally affect how the brain produces behavioral output, from cortical to spinal neural activity.

This view of agency is common not only in philosophy, but also in other disciplines, such as psychology and neuroscience. Principally, these disciplines agree that our behavior is governed by causally efficacious mental states, which emerge from the brain as their physiological basis. As a result, this view is compatible with a neuroscientific view of behavior and agency, and can

[21] Cf. Rammert, 'Distributed Agency' (n 8), 77–86.
[22] Cf. Section II 2(a).
[23] Cf. Section II 2(b) and II 2(c).
[24] Cf. M Bratman, *Intention, Plans, and Practical Reason* (1987).
[25] Cf. T O'Connor and C Sandis (eds), A *Companion to the Philosophy of Action* (2010).
[26] E.g., A Mele, *Springs of Action: Understanding Intentional Behavior* (1992); M Bratman, *Faces of Intention: Selected Essays on Intention and Agency* (1999).
[27] For an account detailing the effects of intentions not only on other mental states but also neurophysiological states underlying the execution of movements, cf. E Pacherie, 'The Phenomenology of Action: A Conceptual Framework' (2008) 107 *Cognition* (hereafter Pacherie, 'Action').

be used to describe the basic rationale of current brain–computer interfaces and neuromodulation technologies. In what follows, we will primarily focus on motoric neuroprostheses, as they provide a clear and instructive case of application. The rationale for motoric neuroprostheses reads: If an agent cannot perform actions and movements anymore because the causal chain from the brain to the extremities is, in some way or another, interrupted, disrupted, or limited the brain–computer interface can bridge causal gaps in this chain by (re-)connecting the neural correlates of intention with an artificial effector, such as a wheelchair or robotic arm.[28]

This basic rationale highlights a mainly restorative and supplemental quality of neurotechnology, which we call the compensatory view, as its main focus is to compensate for lost or limited neural function. The compensatory nature of neurotechnology is illustrated by *Walter Glannon* in his analysis of the specific interaction between brain–computer interface and user. Arguing that neurotechnologically assisted agency is comparable to natural agency, *Glannon* states: 'BCIs do not supplant, but supplement the agent's mental states in a model of shared control. Rather than undermining the subject's control of his behavior, they enable control by restoring the neural functions mediating the relevant mental and physical capacities'.[29] Besides drawing on the standard view of agency, in which the device compensates for the interrupted chain of events from intention to movement by bridging causal gaps, *Glannon* states that an 'extended embodiment'[30] is a further prerequisite: If the user fails to experience the device as part of her own body schema, she may not perceive the movements of the robotic arm as 'her own' which could 'undermine the feeling of being in control of one's behavior', thereby disrupting her sense of agency.[31]

According to *Glannon*, the restorative and supplemental character of brain–computer interfaces stems from the specific interaction between user and device, which creates the phenomenon of shared control, in other words, control over the course of action is partly on the side of the user, and partly delegated to the brain–computer interface. The interaction consists of the user directing her mental states in such a way that the interface can detect neural states which 'encode' her intentions. This kind of interaction is the basis of *Glannon*'s notion of shared control, and successful extended embodiment is necessary to sustain and improve this kind of interactive control.

Brain–computer interfaces based on these principles have been successfully implemented in human patients, and the technology clearly has the potential to compensate for limitations of agency in the way described above. However, it is important to note that while this conclusion is valid, it also stems from a specific understanding of technology, which might support the conclusion while also obscuring other relevant aspects. The compensatory view conceives neurotechnology as a type of instrumental technology and hence frames brain–computer interfaces as auxiliary devices. From this perspective, neurotechnological devices are conceptualized as tools which remain, by definition, fundamentally subordinate to human autonomy and intention. BCI operation appears as auxiliary in nature because the device takes over only partial segments of a course of action, and the overall goal and regulation of action remains governed by human agency.

[28] The same general rationale applies to many other use cases of neurotechnologies that alter, modulate, or monitor brain activity to, for example, enable the use of digital keyboards or cursors, neurofeedback systems, or brain stimulation devices such as deep-brain-stimulators.

[29] W Glannon, 'Neuromodulation, Agency and Autonomy' (2014) 46 *Brain Topography* 27 (hereafter Glannon, 'Neuromodulation').

[30] Ibid, 51.

[31] Ibid, 51. Note that experiencing control over one's behavior may be only one aspect of the sense of agency (cf. Subsection II 2(c)).

https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

In principle, many technological settings can be usefully described from the perspective of instrumental technology. But is this the case in BCI operation? After all, a brain–computer interface is not just an external object, but a device implanted into the brain, affecting and interacting with the origins of action rather than just the external locus of object manipulation. So, does this intimate characteristic distinguish a brain–computer interface from an external tool?

To address this question, we need to examine the effects of BCI operation concerning its causal and neurophysiological nature to see if brain–computer interfaces 'just bridge a causal gap', or if they do more than that.[32] This analysis will suggest that the compensatory view on BCI technology is an extension of the standard view on agency, thereby inheriting its conceptual limits. To counteract this limitation, we need to extend the vocabulary we use to describe agency, and we will do this by taking a closer look at the specific kind of interaction between user and device, taking into account certain social characteristics of this interaction.[33] The basic idea we need to address is that there are some human–machine interactions which are so intimate that it becomes hard to say where human agency ends and machine-agency starts: The interaction between human and machine is such that agency is actually distributed across both interaction partners, rather than ultimately remaining under the governance of human intention.

### b. Reframing Causality

Concerning the neurophysiological nature of a brain–computer interface operation, and shared control specifically, it should be noted that 'a brain–computer interface records brain activity' does not mean that it simply 'detects intentions in the brain'. A brain–computer interface is not like an ECG that detects a heartbeat. Rather, operating a brain–computer interface relies on a mutual learning process: Recently developed interfaces increasingly rely on machine learning to distinguish relevant from irrelevant information about intended movement from a narrow recording site that yields a stream of noisy and limited data.[34] At the same time, the user has to learn to influence his neural activity in such a way that the recording site provides enough information in the first place to successfully operate the external effector. This is achieved by passing through a lengthy training period in which user and interface gradually attune and adapt to each other.[35] Shared control over actions in *Glannon*'s sense is based on this kind of mutual adaptation.[36]

However, this attunement and adaptation between brain–computer interface and user also affects the brain as a whole, which mitigates the claim that in these user–computer interactions, control is merely partly delegated from user to device. As *Jonathan R. Wolpaw* and *Elizabeth Winter Wolpaw* note, natural (i.e. not neurotechnologically assisted) agency is a product of activity distributed across the whole central nervous system, which continually adapts and changes to produce appropriate behavioral responses to its environment.[37] Introducing a brain–computer interface basically creates a novel output modality for this complex system.

[32] See Sub-section II 2(b).

[33] See Sub-section II 2(c).

[34] For an overview of the principles of brain–computer interface operation see JR Wolpaw and EW Wolpaw (eds), *Brain-Computer Interfaces: Principles and Practice* (2012) (hereafter Wolpaw and Wolpaw, 'Brain-Computer Interfaces') or B Graimann, B Allison, and G Pfurtscheller (eds), *Brain-Computer-Interfaces: Revolutionizing Human-Computer-Interaction* (2010).

[35] For an exemplary case see JL Collinger and others, 'High-Performance Neuroprosthetic Control by an Individual with Tetraplegia' (2013) 381 *Lancet* 557–564.

[36] Cf. Wolpaw and Wolpaw, 'Brain-Computer Interfaces' (n 34) 7: 'BCI operation depends on the interaction of two adaptive controllers [brain and BCI]'.

[37] Wolpaw and Wolpaw, 'Brain-Computer Interfaces' (n 34) 6.

As a result, the central nervous system as a whole adapts and rearranges in order to learn to control this new way of interacting with its surroundings. And because brain–computer interfaces rely on a localized recording site and a specific type of neural signal, the user needs to retrain a small part of this extensive system to provide an output which normally is produced by the whole central nervous system, which in turn affects how the central nervous system works as a whole.

In our view, this speaks against the basic tenet of the compensatory view that a brain–computer interface just supplements the agent's mental states, as the whole system that is producing mental states is affected by neurotechnological interfacing. Specifically, it puts into question the view that a brain–computer interface simply bridges a causal gap in the action chain of its user, as a brain–computer interface does not carefully target a specific causal gap. Rather, it modulates the whole system to restore causal efficacy, restructuring the causal chain from intention to action in the process. While this does not mean that a brain–computer interface necessarily supplants a person's agency, we still claim that the compensatory view might easily miss important ramifications of the technology, even in terms of causal efficacy. Furthermore, we argue that the compensatory view also falls short of identifying more overarching agency-altering effects of neurotechnology. While *Glannon* discusses aspects of the sense of agency in terms of extended embodiment and experiencing control over one's behavior – important aspects that contribute to explaining the sense of agency – both embodiment and the sense of agency include further aspects. For example, it has been suggested that the sense of agency is an aggregation of at least three distinct phenomena, namely, the sense of intentional causation, the sense of initiation, and the sense of control.[38] The latter can be distinguished further into the sense of motor, situational, and rational control[39], raising the question of which aspects of control are actually shared between user and brain–computer interface. While the case of motor control seems quite clear, any effect of a neurotechnological device on rational or situational control over actions should be analyzed rigorously – the question is if an exclusively causal and neurophysiological vocabulary will suffice to explore these effects and their overarching consequences. It is to be expected that this situation will become even more pressing with the inclusion of increasingly complex and autonomous AI-technology. As outlined earlier, even current machine learning–supported brain–computer interfaces cannot be understood as simple 'translators' between brain and computer. Advanced AI-technologies will likely introduce additional dimensions of influence by establishing more sophisticated means of interaction between human and machine. We argue that this necessitates a framework that can capture not only specific causal effects, but also changes in interactivity between human and machine which might modulate the causal setting of agency altogether.

### c. Reframing Interactivity

The compensatory view addresses interactions between user and brain–computer interface by highlighting that both the causal compensation and the integration into the body schema is based on a reciprocal learning process. However, the interactions and adaptations between user and brain–computer interface also have a social dimension which is not addressed by the compensatory view. We argue that this is due to conceptual blind spots that result from its vocabulary, which treats agency and intentionality as purely biological functions. As a result, the compensatory view struggles with identifying and factoring in nonbiological (e.g., social and normative) and nonhuman (i.e. artificially intelligent) dimensions of agency.

[38] Cf. Pacherie, 'Action' (n 27) who integrates empirical studies in her theory. For phenomenological aspects see S Gallagher, 'Multiple Aspects in the Sense of Agency' 31(1) *New Ideas in Psychology*.

[39] Pacherie, 'Action' (n 27) 209–213. Also cf. J Shepherd, 'The Contours of Control' (2014) 170 *Philosophical Studies*.

https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

To counteract this shortcoming, it is necessary to extend the vocabulary of agency accordingly. Sociology, Science, and Technology Studies and Philosophy of Technology have a rich history of analyzing how technology permeates modern life and deeply affects and changes human agency. We will paradigmatically draw on a sociological theory called the gradualized concept of agency[40], which shifts the focus from agency as a biological capacity to agency as a phenomenon that emerges from various types of interactions between and among humans, machines, and software. Advanced technologies, it is argued, create a multitude of heterogeneous artificial 'agencies' which interact and influence not only each other, but also human agency in fundamental ways. Importantly, the gradualized concept of agency can be used to examine interactions between a brain–computer interface and its user on the level of human–machine interactions without contradicting the neurophysiological aspects of human agency discussed earlier. In fact, the gradualized concept of agency may help to emphasize that the compensatory view is not outright false by demonstrating its blind spots in a constructive manner.

As argued above, the compensatory view regards neurotechnology as a passive tool by arguing that its contributions to a course of instrumental action concern only partial sequences in the causal chain, while the order of causal events still is governed and regulated by human intention. Hence, the significance and involvement of technological contributions is derived primarily from human intention: The user and his intentions remain in control of the action.

By contrast, the gradualized concept of agency offers an analysis of this kind of relation that shows how advanced technology can subtly restructure instrumental action and lead to agency-altering consequences. It draws on an action-theoretic distinction between three dimensions of agency. The intentional dimension contains the rational capacity to set action goals and deliberate courses of action. Human intention embodies this capacity as an overarching mental state that governs action from planning to execution. The regulative dimension corresponds to control and monitoring of action courses. And the effective dimension describes the base level efficacy to causally affect the environment depending on intentional and regulative aspects.[41]

Based on this model, the gradualized concept of agency argues that technological involvement in the effective dimension can easily cascade from the effective to the regulative and even the intentional dimension. Three common motives of instrumental action illustrate this shift, as technology is often used to delegate effective and regulative aspects of actions in order to save time, improve action outcomes, and to realize action goals the agent could not realize herself. While these aspects may not seem noteworthy when using a conventional tool like a hammer or a common car, their significance and interconnectivity increases the more advanced a technological device is. This can be illustrated by way of two examples: Firstly, a navigation system not only saves time when planning a route, it also improves travel times by calculating and continuously adjusting the best route based on actual traffic data; and secondly, the Google search algorithm seems to be a simple tool to search for relevant information on the Internet. But by scanning billions of websites and documents in fractions of a seconds it is not only infinitely more efficient in finding information, but also autonomously regulates the search by ranking

[40] Cf. W Rammert and I Schulz-Schaeffer, 'Technik und Handeln. Wenn soziales Handeln sich auf menschliches Verhalten und technische Abläufe verteilt' in W Rammert and I Schulz-Schaeffer (eds) *Können Maschinen handeln?* 11–64 and I Schulz-Schaeffer and W Rammert. 'Technik, Handeln und Praxis. Das Konzept gradualisierten Handelns revisited' in C Schuber and I Schulz-Schaeffer (eds) *Berliner Schlüssel zur Techniksoziologie* 41–76. For further aspects also see I Schulz-Schaeffer, 'Technik und Handeln. Eine handlungstheoretische Analyse' in C Schuber and I Schulz-Schaeffer (eds) *Berliner Schlüssel zur Techniksoziologie* (hereafter Schulz-Schaffer, 'Technik und Handeln') and Rammert, 'Distributed Agency' (n 8).

[41] Schulz-Schaffer, 'Technik und Handeln' (n 40) 4–5. For an English version with slightly different terminology and line of argument cf. Rammert, 'Distributed Agency' (n 8) 74–77.

relevant information depending on context, which it determines dynamically. Google not only finds information; it evaluates which information is relevant.

It is noteworthy that technological artifacts themselves are the product of complex intentional actions, and that they embody the intentionality of their design: They are 'objectively materialized structures of meaning'[42]. In this perspective, artifacts carry normative weight which affects the structure of the actions they are involved in. Their designed versatility stems from being oriented towards typical rather than individual action, making them multipurpose and offering reliable repeatability of action. As a consequence, using an artifact requires that the agent adapts to its purpose rather than the other way around – particularly in cases where the artifact takes on partial actions which a human agent could not perform. This characteristic illustrates that technology not only improves or creates new courses of action, but that it is suggestive of certain action goals. Hence, artifacts have an active role in the intentional dimension as well. This effect is magnified when artifacts use software algorithms so that the user can delegate aspects of planning, monitoring, and control to the respective program.

These examples show that many interactions between user and advanced technology consist in various forms of delegation. In the context of AI-based neurotechnology, the combination of machines and software is of critical importance, as the involvement of machine learning and other AI-technology amounts to the inclusion of increasingly autonomous software agents in the equation which are capable of the self-generation of actions. Because software agents not only interact with human users, but also (and mostly) with other software agents, their 'intra-activities'[43] create open systems which lose the transparency of operation we usually expect from technological tools. Hence, when delegating actions to such intra-acting software agents, we do not use a tool, but interact with another type of agency. *Rammert* notes that '[w]hen human actions, machine operations and programmed activities are so closely knit together that they form a "seamless web", [we need to] analyze this hybrid constellation as a heterogeneous network of activities and interactivities.'[44] The gradualized concept of agency enables this kind of analysis by proposing the concept of distributed agency, which can be seen as a nondualist perspective[45] on the complex interactions between human and nonhuman contributors to agency. Of particular interest to us is the notion that agency can be (and often is) distributed across a hybrid constellation of entities, including (but not limited to) humans, machines, software, and AI. In this respect, being 'distributed' means that a simple observable movement performed by a patient with a BCI-enabled prosthesis is the result of a complex interplay of activities, interactivities, and intra-activities. So, who is acting in scenarios of neurotechnologically assisted agency? Following the gradualized concept of agency, not a singular agent, but a hybrid constellation of people, machines, and programs over all of which agency is distributed in complex ways.

The concept of distributed agency includes a further dimension which is of importance to our argument, namely the modern sociotechnological setting, or the 'technological condition' we mentioned in the introduction. With the concept of distributed agency, the gradualized concept

[42] Schulz-Schaeffer, 'Technik und Handeln' (n 41) 8, 18–19.

[43] In the gradualized concept of agency, intra-activity describes interactions among artificial (e.g., machinic and software) agents.

[44] Rammert, 'Distributed Agency' (n 8) 82. Note that the gradualized concept of agency defines interactivity as the specific case when human and nonhuman agencies intersect (ibid, 71).

[45] The traditional dualist or asymmetrical perspective on human–machine interaction asserts a dichotomy between 'human action' and 'machine operation', matching the former with the realm of autonomy and morality and the latter with heteronomy and causality (cf. instrumental theories of technology and the paradigm of tool use). The gradualized concept of agency directly opposes this perspective, at least in the case of complex technology.

of agency argues that technologically assisted agency emerges from 'many loci of agency'[46] rather than from singular instrumental actions (e.g., tool use) performed by an individual human agent. While the individual agent does contribute to agency, his contribution is only one activity in a stream of human interactions, machinic intra-activities, and human–machine interactivities. The sociotechnological setting can be addressed by further analyzing human interactions and machinic intra-activities.

*Rammert* notes that complex technological actions, such as flying tourists to Tenerife with a commercial airplane, include not only individual actions by the pilot, but also considerable contributions from a multitude of both human and nonhuman contributors.[47] On the human side, the pilot is fully dependent on the flight team on board (co-pilot) and on the ground (air traffic controllers, radio operators), as well as the airline company which planned and scheduled the flight, and also the passengers buying the tickets, and so on. On the technical side, the flight is also facilitated by the intra-activities of the various machines and programs integrated into the airplane as well as the respective facilities on the ground. Also, consider that the majority of the flight actions are performed by the auto-pilot, which consists of software programs which constantly measure, monitor, and adjust the mechanical parts of the airplane while checking back with the software networks on the ground which assist in planning, controlling, and navigating the airplane.

Coming back to the example of a movement performed by a patient with an AI-based, BCI-enabled prosthesis, we can apply the same perspective. At first glance, it is just the patient who directly performs the movement of the prosthesis. However, we need to acknowledge the different teams involved, for example, doctors and nurses who performed the initial surgery, and the researchers, technicians, and engineers who built the prosthesis, designed the clinical study, and maintain the device. Also, the hospital, healthcare system, and research and development are related associations of people. And lastly, funding agencies, policies, and social demands contribute to enabling the movement of the neuroprosthesis as well. On the technical side, a neuroprosthesis includes the 'decoder' which can be considered a piece of AI as it employs machine learning to interpret the neural data monitored by the implanted electrodes. While a science fiction example at the moment, the inclusion of more complex AI solutions in brain–computer interfaces may well be achievable in the near future.

## III. HYBRID AGENCY AS THE FOUNDATION OF CYBERBILITIES

The concept of distributed agency is a valuable tool to describe agency beyond the scope of the individual biological functions which underlie the human capacity to act in accordance with their intentions and plans. It shifts the perspective from the limited compensatory view of technological agency to the complex context in which technological agency not only takes place but emerges as the product of a broad spectrum of biological, psychological, social, and political factors. In this sense, the notion of distributed agency can be used as a viable philosophical tool to expose the conditions of possibility regarding concepts such as intention or capability.

### 1. *Distributed Agency and Hybrid Agency*

Because we aim to focus this critical potential on neurotechnologically-assisted agency in particular, we are faced with the challenge to address both its neurophysiological dimension –

[46] Rammert, 'Distributed Agency' (n 8) 78–81.
[47] Ibid, 78–80.

https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

because neurotechnological devices are directly 'wired' into a person's brain – and the socio-technological dimension – as such a device entails complex inter- and intra-activities between and among humans and machines. Thus, we introduce the concept of hybrid agency as a special case of distributed agency, namely as human–machine interactions in which agency is distributed across human and neurotechnological elements. This further emphasizes that neurotechnology – which, by definition, is technology that is directly connected to the brain – is not a conventional tool because it shapes agency not only by being used, but also by directly interacting with the origin of agency. Hence, hybrid agency describes intimate 'fusions' of human and machinic agency and requires direct human–neurotechnology interaction as a basis – but, of course, this does not exclude any biological, psychological, social, or political factors which are directly or indirectly related to neurotechnology as well. These related or indirect factors still shape the structures of neurotechnologically assisted agency, and can themselves be shaped by neurotechnology. And, importantly, hybrid agency specifically includes the various systems of intra-activities among technological and software-agents which neuro-technological devices imply.

The concept of hybrid agency directly opposes the compensatory view, which reduces these complex dimensions by drawing on the instrumental theory of technology, equating neuroprosthetics with conventional tool-use. In this model, neurotechnologically-assisted agency means that a single human agent uses a passive technological tool that compensates for limitations in the action chain, allowing the user to perform actions she would have performed anyway if she could have done so.

## 2. *Cyberbilities As Neurotechnological Capabilities*

Hybrid agency is the foundation of cyberbilities insofar as this kind of technologically-assisted agency creates specific types of capabilities (i.e. opportunities to gain functionings) which we call cyberbilities. A formal definition reads: 'cyberbilities are capabilities that originate from hybrid agency, i.e. human–machine interactions in which agency is distributed across human and neurotechnological elements.' Because capabilities are defined as real opportunities to achieve functionings – beings and doings that increase well-being – cyberbilities are real opportunities to achieve such functionings as the result of hybrid agency.

It is important to emphasize that cyberbilities are capabilities, not functionings. They are not specific skills or abilities a person may gain from neurotechnology. Rather, they denote the opportunities to gain all kinds of (neurotechnological or 'natural') functionings. And even functionings are not just skills or abilities (doings), but also include states of being (like having financial or social resources or being informed about a certain subject matter). If a paraplegic person uses a brain–computer interface to gain the ability to control her wheelchair, the resulting cyberbilities are related to the opportunities that are gained by this type of technological agency. The brain–computer interface opens up a spectrum of agency that was previously restricted, allowing this person, for example, to attend a wedding and thus participate in socializing, which potentially increases this person's well-being.

Hence, cyberbilities denote the opportunities opening up for users of neurotechnology. But because they are the result of hybrid agency, they are also the product of a technology that affects agency as a whole, in other words, not only on the level of causal efficacy, but also concerning psychological, social, and political factors. While a neurotechnological device may be designed to restore, facilitate, or enhance specific skills, gaining or regaining such skills has wider implications in that this can change how we conceptualize and live our lives. This is why

https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

neurotechnological agency cannot be reduced to gaining specific skills. We devised cyberbilities as a conceptual tool to reflect this important factor and provide a means of orientation concerning the potential developments entailed by the use of neurotechnology. Furthermore, cyberbilities are also concerned with the social ramifications of neurotechnological agency. The more the availability of neurotechnology increases, the more it affects all members of society.

## IV. CYBERBILITIES AND THE RESPONSIBLE DEVELOPMENT OF NEUROTECHNOLOGY

After having developed the concept of cyberbilities, we would like to propose a first tentative and incomplete list of cyberbilities, inspired by *Nussbaum*'s list of *capabilities*.[48] We consider our list to be incomplete because it is not meant to cover all basic needs of human beings, nor does it include any other holistic ambition. Therefore, the list presented in the following section should not be understood as a replacement of *Nussbaum*'s list. Rather, we merely aim to stimulate discussions about the implications of future neurotechnologies by drawing on core ideas of the capabilities approach. However, cyberbilities are comparable to capabilities in the following way: *Nussbaum*'s central capabilities describe opportunities which are based on personal and social circumstances which, if restricted or unattainable, would greatly reduce a person's chances to gain well-being-related functionings (to 'lead a good life'). Similarly, cyberbilities describe opportunities created by hybrid agency, which, if restricted or unattainable when using neurotechnology, would greatly reduce the chances to gain well-being-related functionings for a neurotechnologically assisted agent.

Our list of cyberbilities is also necessarily tentative: In order to address future neurotechnologies we have to work with a hypothetical view of neurotechnology that includes a type of AI-supported human–machine fusion that is yet to come. We base this view on current developments, where we can observe various endeavors aiming at advancing AI-assisted neurotechnology, from neuroprostheses for severely paralyzed patients, to sophisticated machine learning approaches, up to straightforward futuristic visions such as *Musk*'s neurotech company Neuralink.[49] Based on such enterprises we think of a future technology that is highly invasive and uses AI methods to generate a novel kind of human–machine fusion that goes far beyond traditional technological tools or machines. We assembled this list with this kind of future technology in mind. In the following, we first introduce our list of cyberbilities,[50] then provide some remarks on the responsible development of neurotechnology,[51] and finally discuss a potential objection against our proposal.[52]

### 1. *Introducing a List of Cyberbilities*

The five cyberbilities we introduce below fall on a spectrum that ranges from individual to social and political agency. While neurotechnological interventions can create specific neurotechnologically

[48] Cf. Nussbaum, 'Capabilities Approach' (n 10) 78–80.

[49] As a first application, Neuralink wants to develop brain–computer interfaces for patients with spinal cord injury, allowing them to control computers and mobile devices. Neuralink's vision includes constructing an automated robotic neurosurgery system that implants a fully integrated brian–computer interface with over 1000 channels for monitoring and stimulating neuronal activity in multiple brain regions. Neuralink ultimately wants to make this technology available for commercial use (cf. https://neuralink.com).

[50] See Sub-section IV 1.

[51] See Sub-section IV 2.

[52] See Section V.

https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

enabled functionings, they also affect a person in more general ways. New, enhanced, or restored functionings extend and shift a person's individual range of agency, and invasive or otherwise intimate interactions between human and machine may change how a person relates to their body. Both aspects can affect the identity and self-expression of a person, modulating their individual agency. But hybrid agency also affects social agency: On the one hand, neurotechnologies enable individual actions which can be the basis of social interactions and participation, potentially adding a social dimension even to the most basic movements.[53] On the other hand, hybrid agency itself is a type of interaction between human and neurotechnology which already includes various social aspects. Neurotechnology has the potential to support social agency, but some of its aspects may also radically reshape social engagement. Furthermore, hybrid agency has distinct political dimensions that range from enabling a person to take part in communal to political and democratic processes.

*Autonomy and self-endorsement*: Neurotechnological devices are often used with the intent to restore or increase a person's functionings (skills, abilities, states), which might also suggest that such devices generally support their autonomy as a more general capability. However, this view might be too simplistic if those functionings result from hybrid agency. Hybrid agency entails a relational dimension of autonomy because autonomy is no longer restricted to interactions between human beings but also concerns the interactivity between human and machine. A neurotechnologically-assisted person could retain autonomy in relation to human interactions while losing it in the context of human–machine interaction. Furthermore, due to the intimate fusion of human and machine, simply insisting that the human part must retain autonomy over the machinic part might be an oversimplified demand. Instead, we should address autonomy in this setting not in terms of the primacy and efficacy of human intention (i.e. the compensatory view), but in terms of 'self-endorsed agency'. Autonomy then denotes the extent to which a person experiences their behavior as volitional and self-endorsed as opposed to coerced, driven, or covertly directed by external forces. Understanding autonomy as a cyberbility that is focused on self-endorsed agency might be a viable way to safeguard and promote self-expression and identity.

*Embodiment and identity:* A technological device should restore or enhance a person's body in such a way that the person is able to integrate the device into her bodily experience, meaning that the person can, without disruptions, identify with the artificial 'part' of herself. She should be able to say 'I have acted like this with the support of the technology' or 'the device and I have acted together' or 'I have acted like this, and I did not experience the interference of the device', etc. Although a neurotechnological device may not be unperceivably 'merged' with the body (like, for instance, a deep brain stimulator), but rather remains separate from the body, the person should have the impression that the device 'behaves' in such a way that she can unreservedly identify with the actions she is performing with the support of the respective device. In other words: The person may not have a sense of ownership but should have a sense of agency. The technological tool should be integrated in the *body schema* of a person, even if the body image is radically changed, for example, in the case of neuroprostheses consisting of external artificial limbs which are 'wired' directly into the motor cortex while remaining clearly separated from the patient's body.

*Understandability and life-world*: Hybrid agency describes the fusion between a person and a neurotechnological device that is intimately connected with the brain and body of its user. Although a lay person may never entirely comprehend how such a device works exactly, a certain degree of understanding is indispensable. Complementing existing approaches to an

[53] Cf. W Wang and others, 'An Electrocorticographic Brain Interface in an Individual with Tetraplegia' (2013) 8(2) *PLoS ONE*; supplemental material shows the patient controlling an external robotic arm with a brain–computer interface and intentionally touching the hand of his girlfriend for the first time in years: UPMC, 'Paralyzed Man Moves Robotic Arm with His Thoughts' (*YouTube*, 7 October 2011) www.youtube.com/watch?v=yff20TlHv34&ab_channel=UPMC.

'explainable AI', a technological device should be 'understandable' in the sense that the user knows *that* the device creates a situation of hybrid agency and roughly *how* the device might affect her agency and behavior (e.g., knowing that a brain–computer interface complements the causal efficacy of her intentions and where the causal contribution lies, which might concern not only the execution of movements but also their planning or initiation). Furthermore, a person should be able to act in interplay with the device in such a way that she can always identify herself with the resulting joint action. While she does not need to be able to explain how the device works on a technical level, she rather needs to understand how the device contributes to hybrid actions and how the device creates well-being opportunities and, thus, becomes deeply integrated in the person's 'life world'.

*Social embeddedness and social experience*: Hybrid agency can create opportunities to engage with the social world, be it on the level of restoring mobility and allowing a person to meet other people or on the level of being able to express thoughts and feelings, for example via digital communication devices. Enabling, restoring, and extending such engagements – for example, in the case of severe paralysis, situations that restrict direct social contact (such as a pandemic), or when trying to socialize over long distances – hold the potential of significant well-being gains. At the same time, however, neurotechnology shapes and alters the basic conditions of social interactions, thereby influencing the way both neurotechnology users and nonusers are socially embedded in the first place. One possible way to capture such fundamental changes could be to focus on how our social experiences are affected by technology.

*Political engagement and participation*: By supporting individual and social agency, neurotechnology also opens up opportunities to engage in political activities on various levels and other forms of campaigning for the common good. Neurotechnological devices should be designed to foster participation in democratic processes such as voting, politicking, or running for office, and should also support engagement in local and global communities, organizations, and institutions.

## 2. *Remarks on the Responsible Development of Neurotechnology*

Because neurotechnologies are developed within a society and its always changing and shifting norms and regulations, cyberbilities are also linked to broad and ongoing societal, ethical, and legal questions. The keywords listed below are not to be understood as cyberbilities, but as indicators of more general questions surrounding cyberbilities. For example, due to usually limited resources we may encounter questions like which patient would benefit from this technology, meaning that not all persons may have the chance to alter their agency by gaining cyberbilities. Also, the neurotechnological engagement in certain activities may require laws that protect the user's personal data (e.g., online services, healthcare, marketing). Because neurotechnology is and will most likely continue to be heavily regulated, the use of neurotechnology on the individual and social level will inherit the legal and political aspects associated with the regulation of neurotechnology, potentially affecting neurotechnology users and their agency. These complex areas will require careful analysis in the coming years and the following remarks address some of the most basic requirements to safeguard the responsible development of neurotechnology. Furthermore, both the question of the trustworthiness of technological devices (especially regarding AI systems) in general and questions around data protection and informational self-determination will affect the future of neurotechnology and also how we evaluate cyberbilities in the future.

*Availability*: Market approval of neurotechnological devices is related to a host of important questions. Who will have access to neurotechnology? How is access regulated – via healthcare systems, or even the open market? And how does regulated access affect not only neurotechnology

https://doi.org/10.1017/9781009207898.033 Published online by Cambridge University Press

users, but also those who do not have access to neurotechnology and who have to interact or compete (e.g. in the job market) with those who do? Such questions indicate important consequences for well-being on multiple levels: If neurotechnology users are individually, socially, politically, or otherwise advantaged or disadvantaged, this circumstance generally affects neurotechnology-related opportunities to gain well-being – both for those who have and those who do not have access to neurotechnology. The question of availability specifically reveals that neurotechnology affects not only those who gain hybrid agency, but also those who do not. This aspect could even result in a 'feedback loop', as the relationship between neurotechnology users and nonusers might affect how norms and regulations develop, further changing this initial relation.

*Data protection*: Because neurotechnological devices monitor, record, and process neurophysiological (and potentially other biological or psychological) data, hybrid agency opens up a plethora of ways in which the data can be used and shared to create functionings or cyberbilities. But the same data could also be used for, among other things, political or commercial purposes. A neurotechnological device should be designed in such a way that it collects and uses personal data as conservatively as possible (e.g. restricted to momentary joint actions and activities), or at least implements particularly robust measures to prevent misuse of data (e.g. through encryption). Because AI (i.e. machine learning) is already implemented in neuroprostheses in order to interpret brain activity faster and more efficiently, such devices should be regarded as a genuine 'part' of the patient and thus be subject to the same legal and political protection concerning personal information and human rights as the user herself. Also, any further implementation of AI-technology needs to be carefully designed to safeguard both the data of its user and any human or nonhuman interaction partners.

*Trustworthiness*: A technological device should not only be reliable in a mere technological sense, but the person should be able to trust herself and the device, especially in cases when the device is merged with the human body or brain. This trust could be seen as a broad psychological foundation of neurotechnology usage, as it includes many of the other items on this list and the list of cyberbilities, like trusting that hybrid agency can be self-endorsed, confidence in the physiological safety and digital security (hacking, manipulation, privacy) of neurotechnology, and reliance on understanding, in principle, the ways in which the device modifies and influences one's natural capacity for agency.

## V. DISCUSSION AND CLOSING REMARKS

Neurotechnology will continue to afford us with astounding possibilities. While the application of neurotechnology is currently restricted to medical usage, we hope that we provided a convincing argument anticipating the future scope of this technology going above and beyond the therapeutic restoration of specific skills and abilities. The proposition of the concepts of hybrid agency and cyberbilities is directed at broadening our perspective so that the enormous potential and overarching impact of neurotechnology may come to the fore.

However, we want to discuss one objection that could be raised on this point, namely that the focus on well-being is too one-sided and may lead to disregarding the intrinsic value of human agency. After all, cyberbilities are not based on 'natural' agency, but hybrid agency. What if this novel kind of agency is in some way deficient, because its technological portion somehow detracts from the human part of agency? In some cases, then, well-being could be achieved at the price of losing aspects of 'natural' agency.

This reasonable objection raises questions about the relative normative weights of well-being and agency, a topic that also applies to the capability approach. There, capabilities and functionings are embedded in the more general concept of agency, and the latter itself has an intrinsic normative

value. But does the importance of agency outweigh the importance of well-being? If we transfer this question to the cyberbilities approach, we could ask: Could the pursuit of cyberbilities lead to justifying a loss of 'natural' agency for the sake of gaining well-being that is less connected to human agency, but rather grounded in technological agency? And to add a utopian twist, could an AI-based brain – computer interface at some point know better and decide itself whether human or technological agency leads to more well-being gains?

There probably is no clear answer to these questions. While it could be argued that this thought experiment warrants preserving 'natural' agency, our line of argument in previous sections hopefully demonstrated that 'natural agency' is not easy to define. Following the standard view, natural agency would mean that the intentions of the human agent systematically modulate which actions are carried out. But considering the gradualized concept of agency, we also saw that human agency is entangled in complex social, institutional, and political systems that influence which intentions are available to human agents in the first place. Human agency is already intrinsically affected by our use of technology and its sociopolitical context.

However, we want to address a point we think is related to this general question: the possibility of capability-tradeoffs. We argued that neurotechnology might not just compensate for causal gaps in the action chain, but rather has an influence on the entire action chain by modulating how the brain works as a whole. Furthermore, neuroechnology also affects, in various ways, the formation of intentions that lead to action chains in the first place. As a result, neurotechnology has the potential to lead to both gaining and losing capabilities.

Consider this example: A neurotechnological device might allow a person to achieve mobility-based functionings (like performing grasping movements with a robotic arm, or getting to work with a wheelchair controlled with the help of a brain–computer interface). If this device also has the effect that its user does not experience her movements as caused by herself (significant portions of grasping movements are controlled by the prosthesis; the wheelchair autonomously navigates to the workplace), then the well-being achievements (being self-sufficient at home and earning money) are realized at the cost of losing some portion of agency. This is a capability tradeoff: The capability (in this case, cyberbility) of neurotechnologically enabled mobility is traded off against the capability of controlling and planning one's movements (which is a part of 'natural' agency).

Of course, such tradeoffs are not necessarily adverse or harmful: In the case of grasping, delegating control to the device at the cost of the sense of control might be acceptable as long as a general sense of agency remains intact (for instance, if the prosthesis overall performs in line with the user's intentions). The case of the autonomous wheelchair is similar, although here the delegation of control goes much further because it includes planning and deciding how to navigate. Our argument is, there might be a point at which the 'cost' becomes unacceptable, for example, if significant portions of agency are traded off. Possible examples could be that the device increasingly detracts from agency, severely influences the decisions of users, or significantly affects the process of intention formation.

Naturally, determining the point at which capability tradeoffs become unacceptable is a difficult task as this is not a technical or scientific problem, but a normative one that needs to be addressed from ethical, legal, social, and political viewpoints. But this open question might help to conclude our line of argument, as we understand cyberbilities as a potential safeguard against unacceptable capability tradeoffs.[54]

[54] Funding acknowledgement: The work leading to this publication was supported by FUTUREBODY, funded by ERA-NET NEURON JTC2017.