Neuroethics (2023) 16:10
https://doi.org/10.1007/s12152-023-09517-2

ORIGINAL PAPER

Philosophical foundation of the right to mental integrity in the age of neurotechnologies

Andrea Lavazza · **Rodolfo Giorgi**

Received: 5 August 2022 / Accepted: 2 March 2023 / Published online: 22 March 2023

Abstract Neurotechnologies broadly understood are tools that have the capability to read, record and modify our mental activity by acting on its brain correlates. The emergence of increasingly powerful and sophisticated techniques has given rise to the proposal to introduce new rights specifically directed to protect mental privacy, freedom of thought, and mental integrity. These rights, also proposed as basic human rights, are conceived in direct relation to tools that threaten mental privacy, freedom of thought, mental integrity, and personal identity. In this paper, our goal is to give a philosophical foundation to a specific right that we will call right to mental integrity. It encapsulates both the classical concepts of privacy and non-interference in our mind/brain. Such a philosophical foundation refers to certain features of the mind that hitherto could not be reached directly from the outside: intentionality, first-person perspective, personal autonomy in moral choices and in the construction of one's narrative, and relational identity. A variety of neurotechnologies or other tools, including artificial intelligence, alone or in combination can, by their very availability, threaten our mental integrity. Therefore, it is necessary to posit a specific right and provide it with a theoretical foundation and justification. It will be up to a subsequent treatment to define the moral and legal boundaries of such a right and its application.

Keywords Mental privacy · Neurorights · Intentionality · Autonomy · First-person perspective · Narrative identity

A. Lavazza (✉)
University of Pavia, Pavia, Italy
e-mail: lavazza67@gmail.com

A. Lavazza
Centro Universitario Internazionale, Arezzo, Italy

R. Giorgi
University of Pisa, Pisa, Italy
e-mail: rodolfo.giorgi@libero.it

Privacy and mental integrity as one concept and one right

The concept of privacy is relatively recent and has taken on different connotations and nuances over time. The concept of mental integrity is even more recent and has to do with the technological advances of the last few decades. In this paper, our aim is to contribute to the current debate with an attempt at conceptual clarification and philosophical foundation of a single right to be attributed to individuals, namely the right to mental integrity, which also incorporates a component of mental privacy, to defend the individual in the face of risks posed by new technologies in general, and neurotechnologies in particular.

A subjective right is a recognition by the legal system (i.e., objective right) of a claim, which implies an obligation of others not to do or to do something. It involves the power to act to defend a recognized and

possibly threatened interest. Rights and obligations are thus correlative. So, by right here we mean a general concept that is moral first and can form the basis for legal implementation at various levels, for example as a reference for codes of conduct but also as the specific content of a legal norm, as happened recently with an attempt to include certain neurorights in the Chilean Constitution. In this paper, we are not interested in defending a new specific right vis a vis already established rights to better defend the mental integrity of individuals (cf. [1–3]. Our aim is to provide a basis and justification for the right to mental integrity (as described below) by referring to the fundamental characteristics of our mind as philosophical research has identified and described them – a task that the debate on neurotechnologies and neurorights does not yet seem to have considered carefully enough (cf. [4, 5].

When considering both mental privacy and mental integrity in the face of neurotechnologies, it is useful to make some clarifications with respect to the different levels at which interference with the personal sphere of the individual can occur. This allows us to better qualify the kind, content and rank of rights that are relevant to potential violations of mental privacy and integrity.

It is often claimed that a great deal of personal information is disseminated by individuals either voluntarily (for example, through social media interactions) or unintentionally, through many electronic tracking systems that we implicitly consent to or do not care about at all. In this sense, the right to privacy in its classical meaning should certainly be recognized and protected. However, it appears that it is difficult to propose a rigid application of it in a context in which information circulates in large quantities and at great speed. The type of information that can be collected thanks to digital profiling (be it lawful or illegal) makes it possible to track, analyse and predict many individual attitudes and behaviours – even those that are more sensitive, such as sexual, political, or health-related ones.

Mental privacy as a specific right vis a vis neurotechnology is characterised differently by its kind and content and, consequently, is a candidate for a higher rank. This is justified by the fact that new neurotools, through brain processes decoding, are potentially capable of accessing personal information (thoughts, judgments, desires, intentions) that the individual has never manifested and could never manifest externally, although they can be roughly inferred, as we discuss below. This can happen today, for instance with communication neurotechnology based on detecting and interpreting neural signals to produce intelligible speech, writing, or typing [6].

It can therefore be argued that the elements residing in the individual's mind/brain are potentially more important and are in any case subjectively more relevant to the individual, as they consider them inaccessible to the direct knowledge of others. The mind/brain is thus, even in comparison with the difficulty of keeping other data confidential, the ultimate seat of personal information and the individual's reserve of privacy, to which a special value and, therefore, special protection is to be attributed.

It follows that the right to mental privacy can be distinguished from the general right to privacy by kind, content, and rank. It qualifies as a different kind because it specifically concerns a special access to the mind/brain and some of its characteristic elements (thoughts, judgments, desires, intentions). It has a different content because it focuses on what can be inferred from the decoding of brain/mental states. It has a superordinate and special rank because it concerns the elements of the mind and life of the individual that are the most personal and precious and were until now considered pragmatically inviolable (or almost inviolable) by definition and worth being protected in principle but are nowadays accessible in different degrees through neurotechnologies. Indeed, so far freedom of speech has been taken as a freedom to be defended, while freedom of thought cannot be literally threatened, as thoughts are inaccessible and even a prisoner devoid of all their rights might continue to enjoy free thoughts. But neurotechnologies are poised to be a gamechanger in this scenario since they are in the process of decodifying the neural bases of thought.

A similar argument can be extended to founding the right to mental integrity. Mental integrity roughly refers to the ability to instantiate one's mental/brain states and realise one's mental/brain states processes without interference, including unauthorised observation (more on this below). In this sense, tools that have long been available as psychoactive substances, if taken against the will of the individual, are able to heavily interfere with their mental integrity. However, the novelty brought by neurotechnologies is given by

the specificity of possible interferences and by the fact that such interferences can be realised in ways that the individual is not always able to perceive at any stage of the process. In this sense, the right to mental integrity should be updated and reviewed to include protection from interference that the incorrect or malicious use of neurotechnologies could imply.

Why we must defend mental privacy and integrity

Threats posed by all the unsupervised or unjustified use of new neurotechnologies are often mentioned in recent scientific literature. In attempting to ground the concept of and the right to mental integrity, we do not intend here to give an exhaustive list of all currently available technologies, nor to describe how their specific uses may threaten mental integrity, nor to propose ways in which a neuroright to mental integrity may be codified and implemented. What we are interested in here is to point out with the utmost clarity that the mere possibility of detecting and interfering with aspects of the mind/brain that were hitherto completely unattainable changes the scenario of people's individuality and autonomy as we have conceived it so far [7].

If specific instruments, which are available and usable, can potentially read our thoughts, reconstruct at least partially our point of view, modify our mental processes, influence our personal moral processing, and our identity building, then one should define precisely what is at stake and what must be granted special protection.

Many techniques are available today to monitor the brain: from EEG to brain imaging for clinical use up to consumer applications based on the same methods (gaming, education, and meditation). Work, warfare, and criminal justice are also fields in which neurotechnologies have been applied [8, 9]. The interventions concern the modulation and/or stimulation of brain activity with drugs and variously invasive devices: from transcranial direct current stimulation (tDCS) to transcranial magnetic stimulation (TMS) acting from outside the skull, to deep brain stimulation (DBS, with electrodes implanted in the brain) (cf. [10].

The monitoring activity is based on the general concept of brain data, the definition of which can be borrowed from a recent work. "Human brain data are quantitative data about human brain structure, activity, and function. These include direct measurements of brain structure, activity and/or function (e.g., neuronal firing or bioelectric signals from EEG) and indirect functional indicators (i.e., blood flow in fMRI and fNIRS). These types of brain data can be combined with non-neural contextual data, such as voice recordings, smartphone usage data or neuropsychological assessments, that can be used to support inferences about mental processes in a broader sense" [11].

As the authors explain, brain data are the most direct correlates of mental states, both cognition and emotions. A number of technologies already in use make it possible, by means of reverse inference, to monitor perceptual and cognitive processes from patterns of brain activation [12]. In animal models, mind/brain activity could not only be monitored but also actively modified [13–15]. fMRI scans and high-density electrocorticographic signals are used to decode mental imagery and silent speech [16, 17]. Intracranial EEG recordings have made it possible to identify brain activity patterns related to inner speech [18]. Machine learning has also made it possible to reconstruct the cognitive processes of individuals under examination from the EEG [19].

In addition, "technologies such as neural interfaces, affective computing systems, and digital behavioural technologies enable increasingly reliable statistical associations between certain data patterns and mental activities such as memories, intentions, and emotions. Furthermore, Artificial Intelligence enables the exploration of these activities not just retrospectively but also in a real-time and predictive manner" [20].

Although it is still unfeasible to precisely decode the full content of mental states, research in this field is advancing at a very fast pace. Recently, the first brain implant was authorised,[1] and the neurotech company Neuralink has made bold announcements concerning brain-computer interfaces and brain implants in humans, not only aimed at restoring impaired functions but also at making it possible to merge human cognition and artificial intelligence.[2]

Neuralink has begun experiments to create images into the brains of monkeys without them having external visual stimulation. This is an attempt that

[1] Cf. https://www.fiercebiotech.com/medtech/synchron-implants-brain-computer-interface-first-us-patient-paralysis-trial.

[2] https://www.nytimes.com/2022/11/30/health/elon-musk-neuralink-brain-device.html?partner=slack&smid=sl-share.

has already been carried out [21]. Using electrodes implanted in the visual cortex, it is feasible to create the illusion of seeing a dot of light even in total darkness. Precise stimulation of visual field maps could make it possible to paint a detailed scene that does not exist in the environment but is perceived by the subject. Something similar to the evil genie that could deceive us with respect to the external world imagined by Decartes in his *Metaphysical Meditations* (1641).

This is the reason why our philosophical foundation of the right to mental integrity resorts to basic properties of the mental domain so far considered as unattainable.

Founding and justifying the right to mental integrity

As said, the aim of this paper is to defend an idea of the integrity of the mind which covers different aspects regarding the individual's rights, from the right to privacy to freedom of thought and consciousness. We can start from a relevant definition of mental integrity that has been proposed by Inglese and Lavazza [22]: "Mental integrity is the ability to formulate thoughts, judgments, and intentions, make plans and implement them without direct external interference of any kind due to neurotechnology". This definition can be combined with another definition of mental integrity as the "individual's mastery of his mental states and his brain data so that, without his consent, no one can read, spread, or alter such states and data in order to condition the individual in any way" [23].

One of the key features that we wish to include in the concept/right of mental integrity is that of *mental privacy*. The notion of privacy is generally applied to what is described as the intimate sphere of one's personal space. It is commonly accepted that everyone must have a personal space that ought to be respected and must not be violated by any form of monitoring and dissemination of contents by an external agent.

Accordingly, we define mental integrity *as the protection of and non-interference in certain mental and brain states and processes (correlates of overt mental functions) that are central to an individual's identity, autonomy and worth.* These mental and brain states, processes and data are in the head before they can be manifested and encompass everything that an individual typically does not want to be revealed or, at least, to fully control with regards to their dissemination. Brain and mental domains do not imply, however, any form of ontological dualism between mind and brain. The distinction between mind and brain is conducted on an analytical and functional level, and it relates both to cognition (with psychological assessment and training tools) and to the biochemical mechanisms that regulate mood (with brain imaging and the administration of specific drugs).

It can be argued that our experience of the world and our mental events necessarily depend on the subject and that the subject should be the sole responsible of the direction of their mental activity. The way in which the subject's mental processes are developed and oriented should be defended from whatever undesired or unjustified external intrusion or influence determined by any agent or subject who is not the *owner* of those mental processes.

Given the specificity of the mental processes that we will describe, it makes sense to include mental privacy in mental integrity, in the broad sense of the term, since making an individual's mental processes public through technological means is tantamount to damaging that individual and undermining their identity, autonomy, and value.

For instance, suppose an individual is captured by an extremist group of a particular religion and is forced through violent interrogation to reveal their: the prisoner is then subjected to forms of indoctrination/manipulation/coercion until they convert to the captors' religion. The individual may eventually give in and profess the new faith, even if it is contrary to their most deeply held beliefs, but still mentally retain their original creed. Yet, by resorting to fMRI and an expert software trained on other captives, the group could verify whether the profession of the new faith is sincere and act accordingly against the individual.[3]

Thus, the condition for which every mental event or subjective experience is *free* from the intrusion or influence of agents external to the mental domain is what we define as the *privacy of the mind*. And the difference between earlier forms of privacy violations

[3] We thank an anonymous reviewer for pressing us on this point.

and current and future ones seems to be qualitative, and not only a matter of degree.

In the following subsections, we shall explore in detail the three main concepts that constitute the core of the mental domain: intentionality; first-person perspective; moral autonomy, and identity as a self-narrative. This third aspect that we will consider more briefly concerns the autonomy of the subject in the original sense of the Greek term from which the word is derived. Autonomy, in fact, is the capacity to make one's own law. We shall use Kantian constructivist metaethics as an example of individual ethical elaboration that needs mental integrity in order to be able to manifest itself and contribute effectively to the formation of a shared morality.

These are peculiar aspects that are analytically decomposable and analysable at the mental level and that have neural correlates predominantly in the brain (with possible cognitive extensions, but which must be constitutive in order to be assumed as part of the extended mind; [24]. Such aspects characterise the human being and constitute central features of a human being's identity and ability to relate to the world.

Intentionality

The main way to investigate the role of intentionality is to consider our mental domain. We might say that the *mental domain* is one's private realm where a person can be free to think, make decisions and make individual choices without letting anyone discover or know in advance what is happening in their mind. Apart from that, there is a sense in which the mental domain can be applied to the special and direct relationship between a subject and their *intentional states* (cf. [25], for a recent discussion on this topic). The kind of privacy we aim to describe here regards the subject's possession of intentional states which belong uniquely to that subject.

Our thoughts are naturally inclined towards something, namely we are always able to think about a thing or another. A mental state cannot exist without the object to which it is directed. Indeed, a peculiar feature of the mental domain is the *aboutness* of mental states. Every state of mind is necessarily a state of mind which is directed towards something, and this appears to be a characteristic of our way of introspectively instantiating the elements that constitute all our mental domains (cf. [26], for the classic perspective).

Intentionality regards our way of thinking about the world. The representations of the elements of the world conceived as intentional objects are considered as the different objects of our thoughts [27]. A human mind naturally possesses an ability to instantiate objects as *mental things*, such as representations of physical items, states of affairs, concepts, and propositional attitudes. These are the main aspects that constitute what we define as *thoughts*. We are immediately able to direct a mental state towards a mental thing and to possess it in such a way that we can automatically define that mental thing as our *mental content or state* [28, 29].

This ability appears to be the basis on which our cognitive capacities are grounded, and it is the immediate source of our thoughts since we can perform the instantiation of a mental content or state instantly and effortlessly [25, 26]. Moreover, this ability is pre-theoretical, in that we cannot deduce it from any other concept or learn to do it through any primitive education. We are born with it, and we become gradually aware of this capacity. We acquire this awareness since our childhood, by applying it on different levels of instantiation, from the basic ones to the most complex levels of thought. The process of *ostension* could be plainly represented as a primitive form of meaning association that is realised when, for the first time, we simply point out a thing that captures our attention. Instead, intentionality appears to be a faculty that is rooted in our private mental space.

The nature of the relationship between intentionality and the physical structures of our brain and how the mental domain is in some way connected to the physical domain is still being researched and not yet fully understood. This problem regards the ontological basis of this natural and pre-theoretical ability that belongs to our mind. Our aim here is not to enter this debate; we assume as a matter of fact that intentionality exists as a primary faculty of mental domain and that most of the mental states have intentional properties (this is the traditional idea found in Brentano, 1974).[4] In other words, the human mind works as a natural system that operates through the ability of

[4] Some contemporary philosophers point out that these intentional states necessarily lie in consciousness; cf. e.g., [30, 31],[32]. For different views on this topic, cf. [33],[34],[35]

the self to select *particular abstract objects* (such as propositional attitudes, representations, etc.) and consider them as their *own particular abstract objects*.

In this sense, mental intentionality appears to be the fundamental capacity that characterises the human mind and the main aspect distinguishing it from the kind of intelligence found, as far as we know, in other living entities and artificial systems (cf. [36]. As there cannot be any experience of the world without a subject who makes that experience, in the same way thoughts and mental contents do not exist if there is not an entity that instantiates specific thoughts and particular mental objects. The mere fact that we are assuming the existence of thoughts and discussing their properties entails that mental intentionality is a human ability as a peculiarity of the mental domain.

This suggests that human intentionality is an evident property per se, since it is the only way we can establish a distinction between mental and physical items. To define a certain thing as a mental object, we need a sufficient criterion to distinguish that object from any physical object. We cannot only rely on instances of mental objects, as particular thoughts, or beliefs, but we have to find a *common property* to what we define as mental objects. The sufficient criterion apt for this scope is given by the notion of intentionality conceived as the property of possessing an abstract content within a mental space by directing our attention to it. If this property did not exist, there would not be any distinction between what we call mental objects and physical objects, and we would not be able to refer or simply to describe what happens in our mind when we instantiate a mental state.

An additional proof that shows the existence of intentionality as a natural ability has been derived a posteriori through empirical research. For instance, the studies by Owen [37] and Monti [38] disclosed the persistence of some form of intentionality in patients allegedly in a vegetative state. These studies show the essential role played by the intentional structures in our mental domain and how they relate to their neural correlates. The researchers told vegetative patients to imagine certain familiar situations by realising intentional states, namely to picture performing a sport or making specific movements in a familiar environment.

Although researchers did not know if the patients would even be able to hear their commands, after these instructions were given, the findings were surprising. Monitoring the brain activity of the patients by checking the cerebral areas that are generally activated when engaging with mental representations and spatial navigation, the researchers found that those areas were activated in the vegetative patients in the same way as the brain areas of healthy volunteers who received the same instructions.

According to these studies, brain imaging makes it possible to detect intentional mental processes in individuals who are seemingly unconscious and totally disconnected from the external environment. Even in these cases, some form of intentionality is radically integrated in our *mind* and *survives* in a pathological condition by continuing to operate independently from the activation of other areas of the brain. It seems plausible to conclude that there is a certain *correlation* between the activation of specific brain areas and the intentional states instantiated by the subjects in response to the external stimuli. Here we are not interested in stating that those brain areas causally determine some mental states or that some mental states are precisely identical or reducible to those areas. As a matter of fact, the researchers found that something *happened* in the brain once the instructions were given but they were not sure of what was *felt* by the patients or experienced at the subjective and private level. Indeed, one might argue that these data suggest that a complete activation of the brain is not a sufficient condition for the instantiation of mental states and, consequently, intentional states are not reducible to neural states. However, by considering this experiment we do not aim to draw metaphysical implications concerning the mind–body problem or to advocate for a rigid identity between mental and physical states. Our goal is only to show that the property of intentionality, as a fundamental aspect of our mental privacy, is at risk of being manipulated by new technologies.

In this scenario, vegetative patients tend to react when they are stimulated in the same way as the healthy volunteers. Since the healthy volunteers were instantiating an intentional state following the instructions, it is plausible to say that the vegetative patients were instantiating some intentional states when they received the same stimuli. Here the most interesting point is that the empirical evidence shows the fundamental role of intentionality in the human mind/brain. The fact that, despite a pathological condition, one can *intentionally* react to an external stimulus implies

that our mental/cerebral domain holds a clear intentional structure that tends to persist. In other words, we are capable of developing our thoughts through a primary intentional capacity that can be discovered even through an empirical investigation, not just by our introspective faculty or individual awareness.

So, a mental process begins with the intentional act of a subject who *internally* adheres to a belief or to a representation by considering them as their own mental states (cf. [39],and [40] on the internalist perspective on intentionality). This form of adhesion is automatic, non-inferential and immediate. This does not entail that we cannot have a mental state without having derived this state from other ideas or inferred it from a deep reflection. However, the outcome of a mental process is the possession of a mental state which is only *our,* privately held *mental state*.

Hence, the basis from which we can explain what intentionality is lies in the unique relation that exists between a subject and the object to which the former's mind is directed. One might say that there is a *private dimension* of the subject's adhesion to their mental states when they think about something. A subject can recognize a mental state as their own only if this mental state belongs solely to this subject and not to other subjects. If there is an internal process of this sort, any physical process or tool able to make this process visible or to alter its course is a violation of our mental domain.

The tools by which researchers could detect the presence of some form of intentionality in an allegedly vegetative patient opened up an unprecedented window in our mind/brain. At the same time, they are a dangerous threat to our right to mental integrity, since they can detect one of our most precious and private mental properties.

The first-person perspective

The second fundamental aspect of mental integrity is the first-person perspective, which makes our mental domain a unique source of knowledge, value, and freedom. It seems clear that through our subjectivity we can experience some features of the world and have instances of sensations that are inexplicable through a third-person perspective. If I observe my maths book from the multiple points of view of a scientific approach I can measure its weight, the number of pages, its shape and all the physical aspects that belong to that object. However, one of the aspects that I can also consider is the "what it is like" of seeing the colour of its cover. The specific subjective sensation I have when I see that book is totally within my private mental domain and is not derivable from any physical description of the book. That means that all the physical facts can be described through a third-person perspective, whereas the aspects of our perceptual experience are the objects of a *first-person perspective* (cf. [41, 42].

A third-person perspective offers the possibility to apprehend physical descriptions, measurable data, and the observable facts regarding an object in an intersubjective way which is *prima facie* accessible to every individual. By contrast, a first-person perspective offers only one point of view which gives a privileged access to the *qualitative instances* of an object, namely what we define as *phenomenal facts*. Phenomenal facts possess a private character which qualifies and defines these states as such [43]. For instance, a perceptual experience, like seeing a red apple or tasting a hamburger, seems to have a special *quality* that determines the intrinsic character of that experience. This character has been defined as the "what it is like" of a subjective experience [44] and can include feelings, emotions, and bodily sensations.

In order to comprehend the distinction between a third-person perspective and a first-person one, we should take into account the classic distinction between primary and secondary qualities as it was developed in modern philosophy.[5] Primary qualities refer to every aspect that belongs to the object and can be observed intersubjectively: they are qualities such as size, shape, number, mass, etc. On the contrary, secondary qualities include all the aspects that belong to the subject's experience of a given object and that pertain to their private perception. Indeed, this concept is inevitably linked to the state of consciousness of the self, which constitutes a *substratum*

[5] It can be helpful to mention some contemporary views that introduce qualities not included in the dichotomy between primary and secondary ones. One of the most notable theories is that proposed by Naess [45], who devised the idea of *tertiary* qualities, namely properties that depend neither on the subject nor on the object but belong to *concrete contents* which are related one-to-one to an irreducible constellation of factors conceived in terms of subject, object and medium. Concrete contents and abstract structures comprise what we define "reality". This view can help fill the gap between an objectivist perspective and a subjectivist one regarding perceptual experience. (We thank an anonymous reviewer for suggesting this integration).

as the necessary condition for one to experience any possible feelings or bodily states (for a comprehensive description of these qualities, cf. [46–49].

So, we might state that the first-person perspective allows us to grasp some specific qualities, and this represents a specific and very peculiar way to know some fundamental facts of reality which are inaccessible through the third-person perspective. Based on our subjective experience we can collect a set of data regarding the object that are connected to the consciousness of being an individual who makes experiences. Moreover, these data are not publicly observable as they are not accessible by multiple points of view like the elements analysed in a scientific approach. They remain confined in one's own consciousness. The qualities we apprehend by consciousness are within one's mental domain and are possessed exclusively by the subject (cf. [46, 50–54]. In light of this, we can consider two main properties characterising our subjective experience:

- *the private dimension*: we cannot verbally communicate to others *what it is like* to perceive the instances of qualities we have when we are in specific states of experience. These instances are "locked" in our mental domain, and they are not *translatable* from an epistemic point of view. There is an epistemic boundary between our personal access to the qualitative instances of things and the intersubjective descriptions we can use to convey the contents of experience to other subjects: all the effects we have when we collect phenomenal data are *ineffable* and not determinate [55–57]. In this way, it is not possible to linguistically translate the kind of qualitative character we have perceived to make other individuals understand what exactly we feel when we experience something.
- *the intrinsic character*: the main feature of our experience is determined by the existence of *qualities*. A quality is something whose existence does not depend on anything; its being is a fundamental fact that we are forced to accept without the possibility of further ontological investigations. When we consider something that possesses a quality, we cannot define this quality by using other terms or concepts connected to other qualities. A quality does not exist as the result of a deduction or as an intuition, it is not derivable from an inference. It is an essential aspect that exists independently from any other thing and does not receive its reference from an attribution external from itself. So, it is impossible for anyone to infer the sense in which a quality is experienced or perceived from a physical description of something.

All that said, the subjective experience is a fundamental aspect of the right to mental integrity since it determines our individual perspective of the world. Through our subjective experience we are not only subjects who perceive instances of qualities, but we are also *agents* who have their own personal point of view of the world and decide accordingly. If qualities are the things that make the subjective experience unique, the individual perspective we develop from this exclusive and specific point of view is the basis for our freedom of choice and possibility to act without any external interference.

Concerning the value of the specific point of view of each sentient individual, it makes sense that the characteristic by which a being can be granted a full moral status is their first-person perspective or phenomenal consciousness, that is, the unique and specific ability of an individual to have conscious experiences, such that no one else can have those same experiences. From this condition one can deduce the inviolable dignity of that living being [58].

In this vein, we need to defend the unique character of first-person perspective because it is a prerequisite for our freedom. The very fact that we possess something unique and exclusively "ours" within subjective experience helps us to preserve our freedom as human beings and to resist any hostile influence or power. We may notice that if our subjective experience could be revealed and disclosed, our interests and experiences would be individuated, and we would suddenly become an *object* that risks being turned into a pure instrument. If our subjectivity is violated, we lose the capacity of keeping our experiences private and we lose the most precious and unique characteristic we have as individuals. Moreover, if all the mental processes confined in our mental domain were disclosed and became *public* for someone else who has found a way to penetrate our mind/brain, our freedom would be at risk.

An individual's first-person perspective long seemed like something that cannot be unveiled, precisely because philosophical analysis has always regarded it as a unique experience that cannot be

reproduced from another point of view, being ineffable by definition. What it is like to be in that state remained confined to the mind/brain of the individual. Specific basic sensations, such as seeing a particular shade of colour associated with a relevant memory, or a particular mix of sensations at a given time, are probably impossible to observe and reconstruct externally in their full detail.

However, what is becoming feasible today is the analysis and combination of proxy indicators that, in principle, comes close to an outside reading of first-person perspectives. Brain imaging studies make it possible to decode the content of thoughts [59], to predict choices [60], to estimate the likelihood of behaviours such as suicide [61], to translate depression into patterns of biochemical imbalances [62]. Algorithms that interpret facial micro-mimicry make it possible to draw a fairly accurate picture of the subject's emotions and intentions. The information gleaned from all deeds and choices that are digitally traceable by increasingly widespread cameras offer a huge amount of data that, with the automated processing capability and combined with the set of other data from the neurotechnologies mentioned earlier, could lead to a privacy breach into our first-person perspective, exposing us to an unprecedented manipulation of our mental lives.

For example, the gaslighting that can be accomplished with generic tools today could soon be achieved with far greater effectiveness thanks to new technologies, bypassing all available defences.

The fundamental prerequisite for being autonomous and free from any external intervention that can manipulate our mental states is to protect the private features of the mental domain, which needs to remain impenetrable from the monitoring of whatever device or external actions. Everyone has the right to preserve their ability to intentionally instantiate a mental state and protect their subjective experience without the interference of a machine, device or system that can decode, decipher, interpret, or change their private mental domain.

In the light of this, we can define the integrity of the mind as the necessary condition for which the property of a certain mental state x is not altered or intentionally changed by a state y. State y can be seen as the state of any agent or entity external to the subject of x, namely the subject who possesses x. These agents or entities might be, for instance, devices designed to manipulate x, by reading or decoding the subject's thoughts, or any treatment capable of altering the intrinsic characteristics of x.

In this sense, if I am the subject that instantiates x, I am the only subject who has that state, and x should not be made public to other subjects. Moreover, x should not be altered or changed in its fundamental characteristics, otherwise this would constitute a violation of my mental integrity. One should be able to have a mental state freely and without the presence of any external factor that can disturb this process.

A kind of technology that might alter or orient our mental states is, for instance, an advanced system of virtual reality, such as the implementation of the *Metaverse,* which could redefine social interactions and create a new digital environment without any need of physical contact with *real* and *present* individuals. Even brain implants, brain scanners and any medical treatment aimed to alter mental processes risk being used to make thoughts readable and modifiable. Very recently, a clinical trial for a permanently implantable brain-computer interface was approved in the US by the FDA [63]. This device eavesdrops on the signals from the brain (potentially not only those from the motor cortex) and converts them into commands that enact an action, like moving a robotic arm or a cursor on a screen.

An artificial implant might orient the course of a mental state directed to a particular object by moving it to a different object, maybe suggested and preferred by the subject that is managing this process. The novel structures of artificial intelligence are based on advanced algorithms projected to predict user preferences and interests and manipulate their choices. In this way, it seems that these patterns can also change our behaviour conceived in terms of individual freedom of thought. This risk could be also exemplified by deep learning machines that aim to reproduce the functions of a human brain so as to codify what happens in our mind. However, the problem is not given only by the *reproducibility* of our brain functions but also relies on the possibility of orienting and directing mental processes directly in our mental domain.

So, the main current risk is the attempt to apply some new neurotechnologies to orient the internal states and events of the mental domain and to violate the first-person perspective. Thus, it is necessary to defend the integrity of the mind as a new specific right, for human beings to be able to freely exercise their mental capacities without being manipulated or conditioned by the interference of any external agent.

Autonomy in moral decision and identity-building

Philosophical reflection, especially of Kantian orientation, claims that the fundamental values of personal dignity and autonomy are based on the possibility of self-determination, that is, of being one's own legislator. This metaethical process seems to presuppose a mental/cerebral space devoid of interference – both at the level of mental privacy and at the level of mental integrity – in which to be able to make the necessary reflection to freely embrace certain values and make autonomous choices. Having one's brain/mental states be observed in real time or experiencing direct interference when deciding one's behaviour in sensitive contexts, as could happen with unauthorised, unjustified, or malicious uses of neurotechnologies, would leave no room for the individual to conduct the processes necessary to self-determine morally.

In particular, moral constructivism holds that moral principles and duties arise from the rational justification that each person gives of those principles in a process that is first individual and then intersubjective [64, 65]. Respect for minimum criteria of rationality and freedom in one's reflections is the prerequisite for this activity, through which the individual adopts a rule and shares it with others to arrive at a satisfactory convergence. The autonomy of the individual thus understood is the condition of possibility of moral life itself and must be protected from any interference that threatens this fundamental process.

It is therefore not a question of simply protecting privacy and mental integrity in general, but of ensuring that the fundamental process of moral self-determination, which needs specific guarantees of non-interference, is protected by specific rights when neurotechnologies capable of violating the privacy and mental integrity of the individual become available. Obviously, moral constructivism is not the only metaethical approach on the market, but it is a proposal that draws on Kant and Rawls and is attracting increasing attention in the ethical field. It exemplifies how the defence of mental integrity also involves an aspect that is often overlooked when discussing the risks posed by neurotechnologies and the domains that can be seriously interfered with by their use.

In addition, some threats to identity understood in a narrative[6] and relational sense may also fall within this context. A strong idea of identity is one that presupposes metaphysical essences or stable criteria of individuation. But such a conception of personal identity is challenged by the cognitive sciences. Consider for example the idea of self-conception proposed by Mackenzie and Walker [68] as something that is created rather than discovered (cf. [69]. It is not about searching within oneself for a pre-defined identity, a real inner self that cannot change or can do so only to a small extent. The idea is instead that we construct our personal identity through self-narratives that we produce over time based on the environment in which we find ourselves. "There are cognitive structures through which we interpret our personal histories and psychological traits and shape our intentions and plans" [69].

We are also "dynamic complex co-creations informed by the perspectives and creative intentions of others" [70], 118). Narrative identity construction is therefore conceptualised as a negotiation between the internal perspective and the (both correct and misleading) interpretations coming from the outside. Identity finds stability when the self-narrative is balanced between the two perspectives that communicate through interpersonal exchanges.

This exchange becomes particularly delicate when it is functional to the construction of narrative and relational identity. The right to mental integrity in this sense takes on a particular relevance and value because it concerns the access that the subject wants to or can give to both the input and the output of their narratives. Ultimately, the individual must be able to protect their own personal space in which to construct their identity without violations of privacy, which in this case amount to real interference, because the involuntary publicity of certain inner processes irreparably alters the processes themselves.

One can think of the inner construction of identity as a process of trial and error that may include antisocial, aggressive, or self-destructive aspects – all attitudes that could be misinterpreted and affect judgement about the individual. Instead, such judgement

[6] Cf. Macintyre [66] and Carr [67] for a reflection on the notion of a *narrative self* which constructs one's identity based on the relevant facts of one's biography. According to this view, personal identity is mainly determined by the past and present events experienced in one's life.

should be built on the actual choices resulting from that inner process. An analogy may come from the psychotherapy sessions that an individual may take to resolve some issues that hinder the construction of a pacified narrative identity. Reports of such sessions already enjoy the highest legal protection about privacy, both from the perspective of the psychotherapist's professional ethics and from the law. But psychotherapy is based on the free narrative of the subject.

Neurotechnological violations of privacy could overcome even the ultimate protection of one's inner self and make public individual mental/cerebral states that could be seriously misunderstood. For example, the activation of brain areas that are associated with violent behaviour or racist prejudice (an analysis that is partly already feasible; cf. [71, 72] does not *ipso facto* mean that the individual's identity is characterised by those behaviours. They might be in the process of controlling and overcoming those tendencies that have "spontaneously" arisen in them due to the evolutionary history of the species or to environmental influences.

If we knew that our thoughts are being observed, our moral deliberation could be influenced and distorted by that monitoring, and we would not really be autonomous in the literal sense. Moreover, the observation of our path toward moral deliberation could be misunderstood or misjudged, whereas what matters is the result, that is, our actual judgement or moral choice. Monitoring the process could then condition the outcome in a way that could lead to, for example, conformism.

Conclusion

The potential issues related to identity and self-narrative raise questions about the introduction of specific new rights related to neurotechnology. Indeed, it is not clear how to distinguish between soft and hard interventions, where the latter amount to new potential violations. These novel threats to privacy and mental integrity, in fact, cannot be compared to those posed by longer-standing practices like espionage or the administration of psychoactive substances against one's will.

All the examples so far seem to show that new technologies create a qualitative discontinuity, which is not brought about by the specific technique as such but by its ability to overcome a diaphragm that was hitherto considered insurmountable. This is our mental life. The tools and behaviour of others could interfere with our mental life even before, but not to the extent of reaching the most 'personal' and profound core that is now potentially threatened by the presence of neurotechnologies in the strict sense of the word.

In this vein, the philosophical foundation of a right to mental integrity performs the function of clarifying which aspects are directly exposed, for the first time, to such far-reaching and direct external interference. Intentionality, the first-person perspective, moral choice, and the construction of one's identity are concepts and processes that need as precise a theoretical definition as possible, although the boundaries between them cannot be drawn as sharply as those, for example, between different organic functions or biological tissues.

In this paper, we proposed philosophical foundations for a right to mental integrity that encompasses both privacy classically understood and protection from direct interference in mind/brain states and processes. Such foundations focus on aspects that are well known within philosophy of mind but not commonly considered in the literature on neurotechnology and neurorights, namely intentionality, the first-person perspective, and moral autonomy as related to constructing norms of conduct as well as one's narrative and relational identity. These are all basic aspects of the value, freedom, and dignity of every human being that a convergence of new technologies can expose to violations of an entirely unprecedented kind.

Hence the need to establish a right that will then have to be characterised more precisely in its ethical and legal coordinates. In our perspective, such a right should not be understood as a guarantee against malicious uses of technologies, but as a general warning against the availability of means that potentially endanger a fundamental dimension of the human being. Therefore, the recognition of the existence of the right to mental integrity takes the form of a necessary first step, even prior to its potential specific applications.

Funding Open access funding provided by Università degli Studi di Pavia within the CRUI-CARE Agreement.

Declarations

Conflicts of interests The authors declare that they have no relevant financial or non-financial interests to disclose.

References

1. Ienca, M. 2021. On Neurorights. *Frontiers in Human Neuroscience* 15: 701258.
2. Ligthart, S. (2020) "Freedom of thought in Europe: do advances in 'brain-reading' technology call for revision?", *Journal of Law and the Biosciences*, 7(1), lsaa048.
3. Bublitz, J.C. 2022. Novel Neurorights: From Nonsense to Substance. *Neuroethics* 15 (1): 1–15.
4. Hertz, N. 2023. Neurorights–Do We Need New Human Rights? A Reconsideration of the Right to Freedom of Thought. *Neuroethics* 16 (1): 1–15.
5. Wajnerman Paz, A. 2022. Is Your Neural Data Part of Your Mind? Exploring the Conceptual Basis of Mental Privacy. *Minds & Machines* 32: 395–415.
6. Chandler, J.A., K.I. Van der Loos, S.E. Boehnke, J.S. Beaudry, D.Z. Buchman, and J. Illes. 2021. Building communication neurotechnology for high stakes communications. *Nature Reviews Neuroscience* 22 (10): 587–588.
7. Drew, L. 2022. The brain-reading devices helping paralysed people to move, talk and touch. *Nature* 604 (7906): 416–419.
8. Delfin, C., H. Krona, P. Andiné, E. Ryding, M. Wallinius, and B. Hofvander. 2019. Prediction of recidivism in a long-term follow-up of forensic psychiatric patients: Incremental effects of neuroimaging data. *PLoS ONE* 14 (5): e0217127.
9. Williamson, B. 2019. Brain data: Scanning, scraping and sculpting the plastic learning brain through neurotechnology. *Postdigital Science and Education* 1 (1): 65–86.
10. Lavazza, A. 2022. Free Will and Autonomy in the Age of Neurotechnologies. In *Protecting the Mind: Challenges in Law, Neuroprotection, and Neurorights*, ed. P. López-Silva and L. Valera, 41–58. Cham: Springer.
11. Ienca, M., J.J. Fins, R.J. Jox, F. Jotterand, S. Voeneky, R. Andorno, et al. 2022. Towards a Governance Framework for Brain Data. *Neuroethics* 15 (2): 1–14.
12. Poldrack, R.A. 2011. Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron* 72 (5): 692–697.
13. Carrillo-Reid, L., S. Han, W. Yang, A. Akrouh, and R. Yuste. 2019. Controlling visually guided behavior by holographic recalling of cortical ensembles. *Cell* 178 (2): 447–457.
14. Marshel, J.H., Y.S. Kim, T.A. Machado, S. Quirin, B. Benson, J. Kadmon, et al. 2019. Cortical layer-specific critical dynamics triggering perception. *Science* 365 (6453): eaaw5202.
15. Ramirez, S., X. Liu, P.A. Lin, J. Suh, M. Pignatelli, R.L. Redondo, et al. 2013. Creating a false memory in the hippocampus. *Science* 341 (6144): 387–391.
16. Kay, K.N., T. Naselaris, R.J. Prenger, and J.L. Gallant. 2008. "Identifying natural images from human brain activity. *Nature* 452 (7185): 352–355.
17. Horikawa, T., M. Tamaki, Y. Miyawaki, and Y. Kamitani. 2013. Neural decoding of visual imagery during sleep. *Science* 340 (6132): 639–642.
18. Moses, D.A., M.K. Leonard, J.G. Makin, and E.F. Chang. 2019. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature Communications* 10 (1): 1–14.
19. Omurtag, A., H. Aghajani, and H.O. Keles. 2017. Decoding human mental states by whole-head EEG+ fNIRS during category fluency task performance. *Journal of Neural Engineering* 14 (6): 066003.
20. Ienca, M., Malgieri, G. (2022) "Mental data protection and the GDPR", *Journal of Law and the Biosciences*, 9(1), lsac006.
21. Chen, X., F. Wang, E. Fernandez, and P.R. Roelfsema. 2020. Shape perception via a high-channel-count neuroprosthesis in monkey visual cortex. *Science* 370 (6521): 1191–1196.
22. Inglese, S., and A. Lavazza. 2021. What Should We Do With People Who Cannot or Do Not Want to Be Protected From Neurotechnological Threats? *Frontiers in Human Neuroscience* 15: 703092.
23. Lavazza, A. 2018. Freedom of thought and mental integrity: The moral requirements for any neural prosthesis. *Frontiers in Neuroscience* 12: 82.
24. Clark, A., and D. Chalmers. 1998. The extended mind. *Analysis* 58 (1): 7–19.
25. Crane, T. (1998) "Intentionality as the mark of the mental", in A. O'Hear (ed.), *Contemporary Issues in the Philosophy of Mind.* Cambridge: Cambridge University Press.
26. Brentano, F. (1874) [1911, 1973] *Psychology from an Empirical Standpoint.* London: Routledge and Kegan Paul
27. Searle, J. 1983. *Intentionality*. Cambridge: Cambridge University Press.
28. Crane, T. 2003. *Elements of Mind, an Introduction to the Philosophy of Mind*. Oxford: Oxford University Press.
29. Crane, T. (2007) "Intentionalism", in A. Beckermann and P. McLaughlin (eds.) *Oxford Handbook to the Philosophy of Mind.* Oxford: Oxford University Press.
30. Searle, J. 1990. Collective Intentions and Actions. In *Intentions in Communication*, ed. Philip R. Cohen, Jerry Morgan, and Martha Pollack, 401–415. MIT Press.
31. Searle, J. 1992. *The Rediscovery of the Mind*. Cambridge, Mass.: MIT Press.
32. Strawson, G. 1994. *Mental Reality*. Cambridge, Mass.: MIT Press.
33. Dretske, F. 1995. *Naturalizing the Mind*. Cambridge, Mass.: MIT Press.
34. Harman, G. 1990. The intrinsic quality of experience. *Philosophical Perspectives* 4: 31–52.
35. Lycan, W. 1996. *Consciousness and Experience*. Cambridge (MA): MIT Press.
36. Searle, J. 1980. Minds, brains and programs. *The Behavioral and Brain Sciences* 3 (3): 417–424.

37. Owen, A.M., M.R. Coleman, M. Boly, M.H. Davis, S. Laureys, and J.D. Pickard. 2006. Detecting awareness in the vegetative state. *Science* 5792: 1402–1402.
38. Monti, M.M., A. Vanhaudenhuyse, M.R. Coleman, M. Boly, J.D. Pickard, L. Tshibanda, et al. 2010. Willful modulation of brain activity in disorders of consciousness. *New England Journal of Medicine* 362 (7): 579–589.
39. Kriegel, U. 2002. Phenomenal content. *Erkenntnis* 57 (2): 175–198.
40. Levine, J. 2003. Experience and representation. In *Consciousness: New Philosophical Perspectives*, ed. Quentin Smith and Aleksandar Jokic, 57–76. Oxford University Press.
41. Madell, G. 1988. *Mind and Materialism*. Edinburgh: Edinburgh University Press.
42. Shoemaker, S. 1996. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.
43. Farrell, B.A. 1950. Experience. *Mind* 59: 170–198.
44. Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 83: 435–450.
45. Naess, A. 1985. The world of concrete contents. *Inquiry* 28 (1–4): 417–428.
46. Block, N. (1990) "Inverted earth" In *Philosophical Perspectives* 4, ed J. Tomberlin. Ridgeview.
47. Block, N. 2002. The Harder Problem of Consciousness. *The Journal of Philosophy XCIX* 8: 1–35.
48. Dennett, D. (1988) "Quining Qualia", in A. Marcel & E. Bisiach (eds) *Consciousness in Contemporary Society*. Oxford University Press: Oxford.
49. Peacocke, C. 1983. *Sense and Content*. Oxford: Oxford University Press.
50. Locke, J. (1690) *An Essay Concerning Human Understanding*.
51. Palmer, S. 1999. Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences* 22 (6): 1–21.
52. Rey, G. 1993. Sensational Sentences Switched. *Philosophical Studies* 70: 1.
53. Shoemaker, S. 1982. The Inverted Spectrum. *Journal of Philosophy* 79: 357–381.
54. White, S. L (1995) "Color and the narrow contents of experience", *Philosophical Topics* 23.
55. Block, N. (1978) "Troubles with functionalism", reprinted in (N. Block, ed.) *Readings in the Philosophy of Psychology*, Vol 1. Harvard University Press, 1980.
56. Chalmers, D. 1996. *The Conscious Mind*. New York: Oxford University Press.
57. Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32: 127–136.
58. Kriegel, U. 2019. The value of consciousness. *Analysis* 79 (3): 503–520.
59. Mason, R.A., and M.A. Just. 2016. Neural representations of physics concepts. *Psychological Science* 27 (6): 904–913.
60. Soon, C.S., A.H. He, S. Bode, and J.D. Haynes. 2013. Predicting free choices for abstract intentions. *Proceedings of the National Academy of Sciences* 110 (15): 6217–6222.
61. Just, M.A., L. Pan, V.L. Cherkassky, D.L. McMakin, C. Cha, M.K. Nock, and D. Brent. 2017. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour* 1 (12): 911–919.
62. Deacon, B.J., and G.L. Baird. 2009. The chemical imbalance explanation of depression: Reducing blame at what cost? *Journal of Social and Clinical Psychology* 28 (4): 415–435.
63. Browne, G. (2022) "The Age of Brain-Computer Interfaces Is on the Horizon", *Wired*, https://www.wired.co.uk/article/synchron-brain-computer-interface
64. Bagnoli C (2021) Constructivism in Metaethics, *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.), https://plato.stanford.edu/archives/spr2021/entries/constructivism-metaethics.
65. Rawls, J. 1980. Kantian Constructivism in Moral Theory: The Dewey Lectures 1980. *Journal of Philosophy* 77 (9): 515–572.
66. Macintyre, A. 1992. Plain Persons and Moral Philosophy: Rules, Virtues and Goods. *American Catholic Philosophical Quarterly* 66 (1): 3–19.
67. Carr, D. 2021. Personal identity is social identity. *Phenomenology and the Cognitive Sciences* 20 (2): 341–351.
68. Mackenzie, C., Walker, M. (2015) "Neurotechnologies, personal identity, and the ethics of authenticity", J. Clausen, N. Levy (eds), *Handbook of Neuroethics*, Dordrecht: Springer, 373–392.
69. Wajnerman Paz, A. 2021. Is Mental Privacy a Component of Personal Identity? *Frontiers in Human Neuroscience* 15: 773441.
70. Baylis, F. (2012) "The self in situ: a relational account of personal identity", J. Downie and J. Llewellyn (eds.), *Being Relational: Reflections on Relational Theory and Health Law*, Vancouver: UBC Press, 109–131.
71. Levy, N. 2017. Am I a racist? Implicit bias and the ascription of racism. *The Philosophical Quarterly* 67 (268): 534–551.
72. Poldrack, R.A., J. Monahan, P.B. Imrey, V. Reyna, M.E. Raichle, D. Faigman, and J.W. Buckholtz. 2018. Predicting violent behavior: What can neuroscience add? *Trends in Cognitive Sciences* 22 (2): 111–123.