C Chiurco[1], A Favaro[1], S F Storti[1], L Brusini[1], A M Salih[1], I Boscolo Galazzo[1], S Plis[1], and Gloria Menegaz [1]

[1]Affiliation not available

January 26, 2026

Abstract

The dual concepts of neurotechnology and artificial intelligence (AI) form an intriguing but also potentially explosive mixture because of its many ethical and legal implications. The advent of AI and the progress in neurotechnologies are reshaping the landscape not only in all scientific fields but also in everyday life both individually and collectively, ushering in a new era where the centrality, integrity and identity of humans is no longer a fact. Such tumultuous progress has implications at all levels, individual, societal, economical and political. Without the pretension of exploring the whole set of relevant aspects, we aim at providing a multidisciplinary view on the main ethical, legal and societal issues stemming from neurotechnology and AI, by assessing them using keywords like trustworthiness, fairness, awareness, security, and privacy. In this paper, we propose an overview on the current scenario, taking a philosophical perspective in the light of ethics, and boiling it down to aspects closely related to the technological developments and the regulatory measures that are currently in-place and called for.

The marriage of neurotechnologies and AI: ethical, regulatory and technological aspects

C. Chiurco, A. Favaro, S.F. Storti *Member, IEEE*, L. Brusini *Member, IEEE*, A.M. Salih, I. Boscolo Galazzo *Member, IEEE*, S. Plis *Member, IEEE*, and G. Menegaz *Senior Member, IEEE*

Abstract

The dual concepts of neurotechnology and artificial intelligence (AI) form an intriguing but also potentially explosive mixture because of its many ethical and legal implications. The advent of AI and the progress in neurotechnologies are reshaping the landscape not only in all scientific fields but also in everyday life both individually and collectively, ushering in a new era where the centrality, integrity and identity of humans is no longer a fact. Such tumultuous progress has implications at all levels, individual, societal, economical and political. Without the pretension of exploring the whole set of relevant aspects, we aim at providing a multi-disciplinary view on the main ethical, legal and societal issues stemming from neurotechnology and AI, by assessing them using keywords like trustworthiness, fairness, awareness, security, and privacy. In this paper, we propose an overview on the current scenario, taking a philosophical perspective in the light of ethics, and boiling it down to aspects closely related to the technological developments and the regulatory measures that are currently in-place and called for.

C. Chiurco, A. Favaro, S.F. Storti, L. Brusini, I. Boscolo Galazzo, and G. Menegaz are with the Dept. of Engineering for Innovation Medicine, University of Verona, Verona, Italy. E-mail: {carlo.chiurco, andrea.favaro, silviafrancesca.storti, lorenza.brusini, ilaria.boscologalazzo, gloria.menegaz}@univr.it.

Ahmed. M. Salih is with the Department of Population Health Sciences, University of Leicester, Leicester, UK and with William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University London, London, UK and with PRIME Lab, Scientific Research Center, University of Zakho, Kurdistan Region, Iraq. E-mail: a.salih@leicester.ac.uk

S. Plis is with the Center for Translational Research in Neuroimaging and Data Science, Georgia State University E-mail: splis@gsu.edu

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes a Table summarizing some explainable artificial intelligence methods (pdf file). Contact Prof. Gloria Menegaz (gloria.menegaz@univr.it) for further questions about this work.

I. INTRODUCTION

Neurotechnology encompasses any method and device that interfaces with the nervous system to probe, monitor or modulate neural activity. Though it has a long history behind, it has acquired a new flavor thanks to the advent of recent technological developments, while the marriage with artificial intelligence (AI) has made available tools to probe in a more powerful way the brain structure and function, also enabling the processing of unprecedented amounts of neural data. In consequence, neurotechnology is particularly critical as it touches all the dimensions of humans' being, perceiving and interacting. This shift is accompanied by new ethical, legal and societal concerns. The progress in technological developments allows increasingly faster and global communication means for sharing, gathering and exploiting information in such a way that exposes individuals to many risk factors, previously confined to clinical and experimental environments. This, in turn, has shifted the focus in neurotechnologies, from the clinical to the real world, bringing in ethical, societal, and regulatory issues regarding data privacy, mental autonomy, and cognitive liberty, where regulatory frameworks lag behind technological advancements. Unlike medical settings, where ethical oversight and informed consent are well-established, consumer neurotechnology can operate unregulated, using unprotected data sharing channels and in the absence of institutional oversight and professional supervision.

Today, consumer-grade neurotechnological devices are proliferating, from wearable brain computer interfaces (BCIs) and portable electroencephalography (EEG) systems to mobile apps and tools. While these technologies hold potential benefits, such as helping individuals to develop better focus and relaxation techniques, they also raise important discussions related to the commercialization of brain data and the risk of behavioral manipulation. This becomes an issue when such data may be shared with third parties, sold to marketers, or used in contexts not originally disclosed to end users. In addition, AI adds to the problem of limited traceability especially for complex black box models. Indeed, the term "black box models" includes all those AIs whose internal mechanisms (i.e., the way how the inputs are transformed into outputs) leading to their final decision are not easy to interpret or understand for humans, even if inputs and outputs are known. Thus, ethical and legal safeguards are needed to prevent people from unauthorized cognitive intrusions and potential manipulation [1]. On top of this, the risks are exacerbated by the uncertainty in the measures and AI model outcomes, which bring in difficulties of their own. In response to such challenges, neuroethics and neurorights have been proposed as the branches of ethics and law targeting the specific field of neurotechnologies.

In this work, starting from the illustration of the ethical and regulatory canvas, some relevant technical aspects will be discussed in the light of the implications of the ethical and legal points, with the aim of

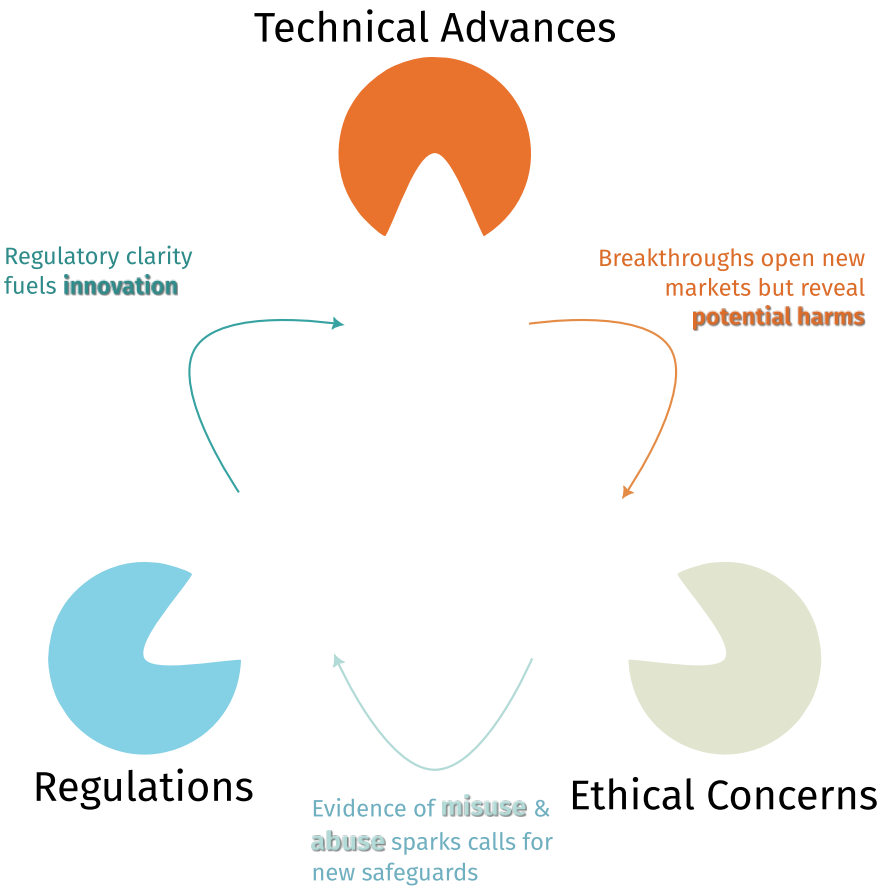

Fig. 1: Schematic representation of the interconnection between the key concepts. Regulation, technology, and ethics are cyclically linked to promote mindful innovation.

providing a comprehensive overview on the so-defined landscape highlighting the inherent complexity and the criticality of the involved issues. Particular attention will be devoted to all the facets concerning specific key terms like “mental manipulation” and “explainability”, in an attempt to provide an exhaustive definition with respect to the field in which they are used. As it is peculiar of philosophical sciences, we also aim at highlighting some ethical and regulatory issues that anticipate technological developments as seeds and food for thoughts for the future. Fig. 1 illustrates the virtuous innovation cycle that interlinks regulation, technology, and ethics. Clear and forward-looking regulations create a stable ecosystem that sparks innovation, driving new technical advances. These breakthroughs unlock opportunities but also reveal potential harms, surfacing fresh ethical concerns. Evidence of misuse and abuse then triggers updated safeguards, refining the regulatory landscape and completing the cycle. The dynamic loop underscores how balanced governance sustains responsible technological progress.

II. ETHICAL CANVAS

A. Toward a trustworthy neurotechnology

Every technology must be dependable, meaning with that it must not only be safe and technically reliable, but also trustworthy. Technical aspects, or functional requisites of trustworthiness, are reliability (for instance, the system is capable to fully operate for a prolonged time, to cope with sudden power shortages, etc.), the capacity to generate a low number of false alarms, the capacity to execute complex operations and safety (meaning with that a low capacity to harm and kill). Each failure in functionality may result in damages affecting the dignity of persons, by reducing the quality of their lives, or even putting them at risk. However, even if functional factors may trigger ethically relevant consequences, they are not intrinsically ethical. A trustworthy technology completes this picture, because it is designed with an attention to ethical factors from the very beginning. For instance, drawing inspiration from the four classical ethical principles governing the field of medical bioethics as devised by T.L. Beauchamp and J.F. Childress, Floridi et al. [2] have envisaged five principles that, if fulfilled, could make AI trustworthy, namely beneficence, non-maleficence, autonomy, justice, and explicability. Therefore, a trustworthy AI respects dignity by not inflicting harm neither directly or indirectly, and it does so by respecting the principles of beneficence and non-maleficence, while, at the same time, respects human autonomy and even possibly enhances it, such as in the case of people suffering from Alzheimer or depression. For the latter to be achieved, it is the level of autonomy accorded to machines and the possibility to explain the decisions they autonomously take that matter most.

A principlist approach, such as that proposed by Floridi et al., is better suited to achieve the desirable goal of a trustworthy (AI-powered) neurotechnology, unlike normative ones, such as consequentialism and deontology. Principlism offers two advantages: i) its tenets mirror the four principles of the well-established tradition of biomedical ethics, which are application-oriented and not merely speculative; ii) the introduction of explicability as the fifth principle provides the ethical rationale for eXplainable AI (XAI); moreover, this notion is currently incorporated in legislation (for instance, in European Union [EU]'s AI Act).

Neurotechnology, AI-powered or not, poses two main ethical issues, namely a) respect of dignity and b) possible challenges to autonomy. Respect of dignity is endangered either by bias leading to discrimination, or by breaching privacy, while respect of autonomy may face manipulation, control, and the intrinsic opacity of sub-symbolic languages. Two possible comprehensive solutions are XAI or the introduction of a new ethical and legal framework known as neuroexceptionalism, calling for the introduction of new rights (neurorights). We will analyse the pros and cons of the latter, before drawing a list of four guiding

principles.

B. Neurotechnology and dignity: bias and privacy

Respect of dignity involves notions such as personhood and personal identity, which include characteristics such as "self-consciousness, responsibility, planning of the individual future, and similar dimensions" [3]. The problem, here, is to list these meaningful characteristics and acknowledge them so that they may be guaranteed. However, another important feature of personal identity is the notion of sameness, whose relation to neurotechnology is ambiguous. Sameness is an important feature of moral evaluation: for instance, we deem someone virtuous because she responds with the same behaviour in similar situations (for instances, she behaves courageously whenever courage is needed). However, neurotechnology may provoke changes in the personality and the personal identity of people: sometimes these changes may constitute the desired result of a therapy, but unusual or unwanted ones could occur after an intervention, such as the insurgence of depression or disproportionate euphoria. International debate maintains that even in such cases personal identity is not compromised [4]. Here the concept of narrative identity can be helpful: the person remains "the same", provided she still recognises herself as the author of her life, even if others (such as her family or friends) cannot recognise her as the person they knew before. Also, neurotechnology always implies a condition where human and machines interact at a very deep level, or are even fused, thus reducing the scope for objections concerning sameness.

However, this does not mean dignity is entirely safe from the challenges posed by neurotechnology. On the contrary, dignity can be indirectly but severely affected by biases and privacy issues, and this is even truer in the case of AI-powered neurotechnology, since AI has repeatedly been found consistently porous in both fields. The principles affected here are beneficence and non-maleficence, because situations and outcomes disrespectful of the dignity of a person or a group often result in damages or unnecessary risks affecting them.

1) Biases: Bias issues especially concern diagnosis. AI may either produce biased results, or these results may induce biases. Both may result in discrimination, affecting the dignity and the life quality of persons. In most cases, discriminatory biases produced by AI are due to the defective quality and quantity of collected data employed by machines in their training. Data that are not homogeneous, or not homogeneously collected, or related only to a partial population may result in discriminatory machine decisional output: for instance, studies on programs used in the field of predictive justice have shown that black population in the United States (US) is known to be subject to a far higher probability to be mistakenly classified as more crime-prone than their white counterparts (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; accessed: 2025-05-27), a risk that the introduction

of AI-powered brain imaging techniques to contribute to more evidence-based decisions in criminal justice (from investigation and the assessment of criminal responsibility, to punishment, rehabilitation of offenders, and the evaluation of their risk of recidivism) could significantly worsen. Other discriminatory output may include a significant degradation in algorithm performances when characteristics such as skin, colour or gender are considered.

These biases, of course, may also occur when AI is embedded in neurotechnology, especially with clinical or therapeutical use. It is well known that existing data mostly concern Caucasian men from Europe and north America, leading to defective generalisation and inaccurate, even wrong diagnosis. As a result, underrepresented and already socially penalised groups are disproportionally affected; moreover, this leads to bigger inequalities and their systematisation, as machines use past results for their training (for instance, black population in the US is disproportionately at risk of being mis-diagnosed with schizophrenia). Similar problems may arise if models are not generalised enough, excluding groups on the basis of age, gender, or health status.

2) Privacy: In neurotechnology, privacy issues are mostly found in monitoring health, and they may arise from the construction of digital phenotypes (DP) with predictive function. Patterns arising in the user-device interaction create digital biomarkers that help track anomalies: for instance, by monitoring repetitive reaction times to a simple task on senior-friendly tablet computers we can predict with good certainty early signs of incoming decline of cognitive functions [5]. However, DP require enormous amounts of personal data, whose collection and storage must follow the rules of informed consent. The people involved must also be allowed to exert control over information concerning them, for instance by rejecting profiling, a practice that violates privacy and hampers (at least indirectly) autonomy as well (see section II-C below). Public authorities cannot hamper individual prerogatives concerning access to collected and stored data and must also protect them from unauthorized access. Such violations are also ethically, not just legally, relevant, since they constitute a violation of personal dignity. However, while informed consent for clinical reasons follows high standards, the same is not true for wearables and apps. Apart from ethically preoccupied legislation, privacy may also be protected by technical features that could prevent damage from unauthorized intrusions, such as avoiding storing all data in one place or using encryption.

C. Neurotechnology, autonomy, and the problems of manipulation and control

Unlike personal identity, neurotechnology poses a direct challenge to autonomy, because it is capable to exert its influence also at an unconscious level, a dimension of human beings normally not considered in normative ethics and principlism. This raises the case for new conceptual frameworks and categories,

because the intrinsically deep, intimate nature of brain data makes existing philosophical, ethical and legal notions of privacy and personhood look increasingly external, while mind-reading and brain-mining techniques force us to rethink the very notion of the inner reality of human beings. Existing conceptions of autonomous subjectivity, together with the ethical and legal frameworks they entail, are likely unfit to fend off the challenge posed by neurotechnology, be it AI-powered or not, because, being all deeply entrenched in the Cartesian mind-body divide, they fail to provide a more faithful picture of the mixed, even contradictory complexity of the human being, whereas subliminal activity is always present at all levels of neural activity presiding over our desires and decisions [6]. All existing ethical paradigms are affected by this capacity of neurotechnology: as for normative ethics, both consequentialism and deontology presuppose the notion of a perfectly rational and fully sovereign individual as essential to decision-making. Principlism, while eschewing the debate about what ethical norms we should follow, cannot escape the problem of whether we are autonomous enough to express our informed consent when it resorts to its notion of Beauchamp & Childress. Therefore, ethical reflection should rethink the notion of subjectivity by considering non-conscious sphere in a way that still allows it to retain its autonomy—a necessary condition in order to consider subliminal manipulation, potentially brought about AI-powered neurotechnology, ethically as well as legally blameworthy. The same difficulties are also found in existing legislation. For instance, in the initial considerations of the EU's AI Act (numbers 28-29), despite AIs are clearly listed as technologies that may be used to limit the autonomy of individuals, legal provisions so far only cover explicit or (partly or fully) aware forms of coercion and deceit, not subliminal ones: article 5 only forbids those practices acting at subliminal level without individuals being aware of them, inducing them to take potentially dangerous decisions that they would otherwise not have taken.

Neuroethical issues concerning autonomy are i) manipulation and ii) control, both made seemingly possible by the capacity of machines of penetrating the inner realm of the mind. Some even fear, and speak of, mind-reading: however, there is no clear consensus on what mind-reading is, while the extent of the capacity of machines in this sense is often exaggerated. On the contrary, mind- (or brain-)mining poses a far more real challenge both in ethical and legal terms. Even non-invasive (since they do not penetrate brain tissue) neurotechnologies such as EEG, functional near-infrared spectroscopy (fNIRS) and functional magnetic resonance imaging (fMRI) allow unprecedented access to our mental processes, laying them bare and exacerbating our vulnerability. Another issue is iii) opacity, shared by all AIs and AI-powered machines and devices.

1) Manipulation: Manipulation can be defined from its outcomes (i.e., the damages produced on the manipulated person), the process it employs (i.e., a hidden influence or by bypassing rationality), or the norms it violates (i.e., manipulators do not abide by the ideal norms of respectful and honest

communication). Not all manipulation stems from malicious intent: for instance, I could produce a deepfake video representing a historic event for educational purposes. Still, such epistemic manipulation challenges the observer's implicit trust bound between representation and reality. Overall, manipulation is a cognitive distortion that damages a person's capacity of shaping justified believes. In this respect, manipulation damages one's autonomy, either epistemically or also morally, in the case that it is enacted for malicious purposes and/or by exploiting the vulnerabilities of a person or a group. While in ordinary forms of manipulation victims may be or not partly or fully aware of such vulnerabilities, neurotechnology, with its capacity to operate at a far deeper, partially or fully unconscious level of the mind, makes a much more difficult case for breaching human autonomy, both from the ethical and legal perspective. The risk of mental manipulation is often associated with intrusive activities of "mining the mind (or at least informationally rich structural aspects of the mind)" ([7], p. 4): such activities are usually carried on in order to infer mental preferences, but they could also prime, imprint or trigger them, leading to a loss of autonomy. Since the emergence of cognitive economics—a branch of economic theory founded by the 2002 Nobel prize winners Vernon Smith and Daniel Kahneman, which rejects the standard neo-classical economic theory and its Cartesian model of perfect rationality—we know that decisions-making processes affecting economic choices emotions and states of mind are just as important as rationality. A new data industry emerged, as unaware consumers' choices could be detected (for instance, by tracking emotional responses) and profiled, while neuromarketing could use such neurodata to artfully divert them. These tasks can be easily done using non-invasive neurophysiological or neuroimaging methods (EEG and fMRI), while EEG signal may predict and simulate consumers' behaviour. Finally, citizens' and customers' autonomy could also be limited by the bigger autonomy enjoyed by AIs as such [8], thus worsening the problem of neurotechnology-driven manipulation.

2) Control: Neurotechnology has also transformed the way social surveillance and control are enacted. It is employed i) in workplaces, ii) in prisons and iii) for enforcement purposes.

i) In workplaces, these applications are designed to reach a maximum level of integration and interoperability. The massive acquisition of different sorts of neurodata (including emotional states, levels of attention or engagement at work, health data that may be useful to determine the level and extent of individual productivity) helps measuring productivity and cognitive performance, allowing staff and executives alike in managing better work schedules. By redirecting workers to activities more coherent with their current level of attention or cognitive performance, it can be used in order to maximise profits, or creating more sustainable work conditions, or both. However, such intrusive hyperconnectivity, enacted both at neurophysiological and neuropsychological level, could easily exacerbate its already well-known downsides, spanning from burnout to schizophrenia and epileptic strokes.

ii) and iii) Possible uses of neurotechnology in prisons and for enforcement purposes may include image recording and applying a BCI in order to control dangerous subjects. Again, a specific economic niche known as correction industry is gradually setting foot. Societies market neurotechnology with AI embedded in them to keep prisoners in check (e.g., *Five ways AI could be utilised in jails* by Equivalent Corrections https://equivant-corrections.com/five-ways-ai-could-be-utilized-in-jails/), offering an automatic (therefore verifiable) management of their behaviour by monitoring their mental states. Correction industry, of course, has much to gain from AI-powered predictive justice, but it could also easily replicate its potentially discriminatory shortcomings: just as some individuals from certain social environments are far likely to be mistakenly classified as more crime-prone (see section II-B1 above), they could also be unjustly deemed worthier of being controlled.

3) Opacity: Sub-symbolic languages may be extrinsically as well as intrinsically opaque. Examples of extrinsic, or upstream, opacity are: data may not be revealed if they are sensitive, or protected by industrial copyright, or used by institutions or professionals that need secrecy, such as the police, or privacy, such as hospitals. Intrinsic opacity means that the ways systems use to produce their decisional output is fragmented and opaque (indeed they are known as black boxes), therefore it is not possible to understand how and why they have reached their conclusions, not even if we were granted access to the source code. Furthermore, such conclusions may be elaborated in ways that neither necessarily mirror, nor match, human reasoning. Intrinsic opacity makes therefore extremely difficult (or systematically impossible) the attribution of responsibility in case of incorrect or plainly wrong output, and strips individuals of their right to appeal against any decision they think unfair or damaging to them, thus dramatically reducing their autonomy.

This has dramatic effects over the very notion of medicine as we know it. In clinical practice, all decisions are taken "using science and conscience": the good clinician acts knowing what to do and having good reasons to do it. On the contrary, AIs do not really know what they do, even less why they are doing it: they only churn out extremely sophisticated statistical inductive inferences on a scale unimaginable and unattainable to the human mind. As a consequence, i) such intrinsic opacity forbids patients to exert their right to appeal against machine-powered clinical decisions; ii) it ushers in an entirely different sort of medicine, grounded not in clinical observation and explanation but in statistical classification and forecast [9]; iii) this new sort of medicine conflicts with clinicians' duty to act always "using science and conscience". The usage of AI in medicine also raises two big questions, concerning disclosure (should patients be informed that clinicians are employing a tool, whose way of operating cannot be explained?) and responsibility (who or what is responsible when a black box machine goes wrong?). Against opacity, the principle of explicability states that humans should always be allowed to

object to machine's conclusions. Machines, in turn, should be designed and built in a way that makes possible to understand where and when they went wrong because this would strengthen humans' case in an appeal against machine-powered decisions if they feel their autonomy has been hampered, their dignity violated, and their interests damaged. A proposed solution to this problem is to build XAIs, which will be discussed in another section of this paper (see the section IV-B below). By helping fighting biases and building trust in machines and devices, XAI could indeed prove decisive to provide trustworthy neurotechnology, AI-powered or not.

D. Neurorights and the debate on neuroexceptionalism

The complexity of this picture and the challenges brought about by neurotechnology have led some authors to doubt whether existing moral and legal conceptual frameworks are actually capable to successfully address the challenges brought about by neurotechnology. Neuroexceptionalism, or the need to introduce a set of distinctively different ethical and legal categories, or neurorights, seems justified by the peculiarly intimate nature of the brain, as well as the danger of manipulation at an unprecedented scale. Those who favour neuroexceptionalism, such as Ienca and Andorno [7], draw a comparison between the exceptionality of neurotechnology and that of biotechnological revolution, which led to the recognition of new rights embedded in the Universal Declaration on the Human Genome and Human Rights (1997) and the International Declaration on Human Genetic Data (2003)—a case that the ongoing marriage of neurotechnology and AI, another technology often hinted at as potentially disastrous for humankind, would only strengthen. The authors list four main ethical issues at stake, each of them corresponding to a specific category of neurorights that are summarized in Table I.

TABLE I: Proposed neurorights: definitions, associated characteristics, and practical implications.

Right	Characteristics	What the right entails / What to do
Cognitive freedom, or mental self-determination.	The right to use neurotechnology Protection of individuals from any coercive or non-consensual usage of them.	Ensures an equal access to them, preventing any discrimination based on racial, sexual or social bias, or them to be available only to a privileged elite, but to all who need them.
Mental privacy.	Upside: neurotechnology provides people with the capacity to monitor and control their brain activity. Downside: it also makes an enormous amount of data available to third parties, but i) there are no legal protections from having your mind involuntarily read and ii) privacy is not just a right, but also an ability (for instance, a person's ability to consent to the collection of brain data).	People should be safeguarded from any illegal access to cerebral information, as well as be able to prevent its diffusion in the infosphere. (The authors even encourage the adoption of a "filter" option allowing people to decide what chunks and bits of information they want to go public.) People's privacy should always be protected, even in the case they consented to such access.
Mental integrity.	Our capacity of neural computation should not be damaged by "malicious brain-hacking" [10], for instance by adding "noise or overriding the signal sent to the device with the purpose of diminishing or expunging the control of the user over the application" ([7], p. 19). Other forging techniques include the capacity "to engineer (e.g. boost or selectively erase) memories from a person's mind" ([7], p. 19).	Mental integrity is protected by Article 3 of the EU's Charter of fundamental rights, but without any reference to neurotechnology.

Continued on next page

Right	Characteristics	What the right entails / What to do
Psychological continuity.	It somewhat summarises the previous rights. It is aimed at preserving individual identity in terms of thoughts, preferences and habitual choices, by protecting the neural substrate underlying each of these acts, given that changes in brain function caused by brain stimulation may also cause unintended alterations in mental states critical to personality, and can thereby affect an individual's personal identity.	

However, while some authors go as far as to consider neurorights universal, as they are rooted in the human brain [11], neuroexceptionalism is fiercely debated. To begin with, the proposed comparison between neurotechnology's breakthroughs and biotechnology's being questionable, therefore neurotechnology would not differ from other well-known technologies implying deeply personal data collection, such as gene sequencing tools [12]. However, only retrospectively the exceptionalism of gene-related technology looks outdated (only today we can consider it to be "well-known") while neurotechnology seems to possess dangers of its own: for instance, whereas hacking is actually common to biotechnology and neurotechnology, cracking (and mind-manipulation) is not. Johnson [13], while recognising that "questions of whether new (neuro)rights are needed and what norms should be included in an instrument remain pertinent", highlights that "issues of who would build and administer such an instrument—and how—often remain underexplored in scholarly and policy conversations". Hallinan et al. [14] distinguish between the claim of exceptionalism of neurodata from the claim that such data need a new ethical-legal framework. Something can be considered exceptional only against a pre-existing set of norms: the authors use fMRI to show how it allows "representations of aspects of individuals' mental and emotional states to be made, which are arguably not—or at least not the same extent—possible via the processing of other forms". Therefore, neurodata obtained by fMRI are fundamentally exceptional, a notion reinforced by the impossibility to anonymise them using techniques such as de-facing. However, it is not clear why the exceptional nature of neurodata should logically imply the necessity of exceptional ethical consideration and legislation. As a consequence, any such regulation should follow boundary conditions limiting, rather than enhancing, a neuroexceptionalist approach. Though the authors' conclusions look sensible, it is worth noting that they overlook the peculiar capacity of neurotechnology to exert its influence also

at subliminal and unconscious level: while acknowledging that neuroexceptionalism' claims may be somewhat overreaching, it is true that neurotechnology, showing the limits of the Cartesian model of rationality on which our notions of autonomy are based, invites to maintain a precautionary approach.

III. The Neurorights and regulatory canvas

A. Some significative attempts at positive legislation on neurotechnology, legal imits and juridical perspectives

As already described in the section on ethics (with some juridical-legal elements), in the legal sphere the analysis studies on neurotechnologies prompted some scholars [15] to coin the term "neurorights" with the proposal to extend the catalogue of human rights to four new rights: the right to cognitive freedom, the right to mental privacy, the right to mental integrity and the right to psychological continuity. According to the Neurorights Foundation (Columbia University), the catalogue of neurorights is extended to five "new" rights: mental privacy, personal identity, free will, fair access to mental augmentation, protection from bias [16]. Under the impetus of such initiatives, the risks of neurotechnology are also beginning to be "law" by some legal order (https://www.vox.com/future-perfect/24078512/brain-tech-privacy-rights-neurorights-colorado-yuste; accessed: 2025-03-10). Chile is considered the first country to have included a reference to this effect in its Constitution. In fact, Art. 19 of the Chilean Constitution (which safeguards mental integrity among the rights of the individual) has been supplemented with the following provision: "Scientific and technological development shall be at the service of individuals and shall be carried out with respect for life and physical and mental integrity. The law shall regulate the requirements, conditions and restrictions for its use on individuals, and shall especially protect brain activity, as well as the information derived therefrom". Consequently, in August 2023, the Supreme Court of Chile ruled that the commercialization of a wearable EEG device and the provision, by the same manufacturer, of an application for managing the data obtained from its use, are incompatible with the right to the protection of the user's mental integrity, unless subject to prior oversight by the relevant national health authorities. It is very interesting how the Chilean decision was taken precisely on the basis of the reformed Art. 19 of the Constitution evaluated together with sources of international law. In fact, we read in the Judgment (pp. 9-10): "The International Covenant on Economic, Social and Cultural Rights establishes in its Article 15 the right of every individual to enjoy the benefits of scientific progress and its applications, as well as their dissemination, preservation, and development". In a similar vein, the UNESCO Declaration on Science and the Use of Scientific Knowledge and the Science Agenda–Framework for Action, outlines in its programme: "that scientific research and the use of scientific knowledge should respect human rights and the dignity of human beings, in accordance with the Universal Declaration of Human Rights and in light of the Universal Declaration on the Human

Genome and Human Rights; that some applications of science may be harmful to individuals and society, the environment, and human health, and may even endanger the survival of the human species; that the contribution of science is indispensable to the cause of peace, development, and the protection of global security; and that scientists, alongside other key stakeholders, bear a special responsibility to prevent the misuse of science in ways that are ethically wrong or have negative consequences, as well as the need to practice and apply science in accordance with appropriate ethical standards, grounded in broad public debate" (Supreme Court of Chile, Rol 105.065–2023, https://www.doe.cl/alerta/11082023/20230811001).

Similar initiatives are beginning to take place also in the US: Colorado is the first state to have a "charter" of neural data, classified as "extremely sensitive". So, just as fingerprints and facial images are under the Colorado Privacy Act, the whisperings of the brain are, too. Signed into law by Gov. Jared Polis, the bill had impressive bipartisan support, passing by a 34–to–0 vote in the state Senate and 61–to–1 in the House. Also in California the Consumer Privacy Act (CCPA) includes now neural data protection under the new California NeuroRights Act. This Act, passed in August 2024, offers landmark protections, regulating how neural data from wearable neurotech devices can be used and shared. Notably, Mexico incorporated NeuroRights into its Digital Rights Charter in December 2023, embedding protections for brain data. Brazil is also taking steps, with its higher education institutions launching NeuroRights degree programs and initiating discussions to incorporate NeuroRights in civil code. These steps emphasize the need for protections as neurotechnologies advance across markets and cultures.

This type of legal strategy may no longer suffice in the present day. While an ethical framework is undoubtedly useful in the context of neurotechnologies, a legal analysis is even more imperative. With companies such as Meta and Snapchat exploring neurotechnology and Apple having patented a future version of AirPods capable of scanning brain activity through the ears, we may soon find ourselves living in a world where companies collect our neural data, irrespective of individuals' consent. These companies could create databases containing tens of millions of brain scans, which could be used to determine whether a person suffers from a condition such as epilepsy, even if they do not wish this information to be disclosed. Furthermore, such data could one day be used to identify individuals against their will (and to undermine their autonomy). From a legal perspective, this issue is even more pressing than its ethical dimension, as it disrupts the action–responsibility relationship, with immediate consequences in terms of civil damages and criminal liability.

B. Limits of theoretical proposals for renewed positive regulation: proposal of a case by case approach

From a regulatory perspective, some scholars have proposed the adoption of an international treaty on "neurorights", accompanied by the establishment of an international agency to ensure compliance by

participating nations [17]. However, it is likely that the time is not yet ripe for the approval of such a global treaty, and this legal solution might emerge as already "outdated" or "obsolete". Consequently, it is necessary to explore alternative juridical solutions that go beyond a purely normative approach and address the well-known challenges of implementing ethical principles in practice.

One concrete first solution, for instance, would be to train future legal professionals within academic institutions by expanding the concept of "privacy" (understood in a flexible manner, to be assessed on a case-by-case basis within jurisprudence) to encompass mental privacy in all its relevant dimensions. From a legal standpoint, this would provide a useful and adaptable tool for safeguarding human dignity and, above all, individual autonomy. Another potential legal solution, returning to the realm of "normative" approaches, could involve classifying all neurotechnological devices as medical devices. This would require them to undergo approval by relevant health authorities prior to their commercialization.

If the aforementioned considerations hold true, discussing "neurorights" in the legal context, rather than referring to entirely new "rights", essentially entails advocating for a more robust protection of rights that already exist. In this context, some law-philosophers [15] believe that the introduction of new human rights would be preferable to develop research in the hermeneutic sphere so as to place these "new" rights in classical legal categories [18]. The investigation is very extensive and, limiting it, for now, to the sphere of the individual person, could focus on the philosophical-legal categories of "privacy" (with elements of constitutional law, administrative law and private law), of "action–omission" (with elements of moral philosophy and constitutional law) and of "responsibility" (with elements of criminal law and criminal procedural law, without addressing the vast topic of "will" in the general theory of civil law with the consequences in the area of "form" and "validity" of contracts or in the area of imputability for "aquilian–extracontractual" responsibility).

Within the limits of the methods and techniques presented in this contribution in relation to neuroscience and neurotechnology, we can confirm that from a legal point of view the most prudent and most effective approach remains that characterised by case-by-case judgement.

Consequently, regulating the matter through purely normative choices (as we have already seen above in section III-A) is insufficient because the matter to be regulated is inevitably connected to (and conditioned by) subjective elements that differ from person to person: autonomy, capacity, awareness, technological understanding, information asymmetry. Each case should have its own specific regulation, as taught by the tradition of classical and pre-classical Roman law.

For reasons of economy, the general framework of guarantees and protections for the "body", which already lends itself to new forms of protection from neurotechnological interference, could be hypothesised as a solution. On these basic rules could then rest the prospect of a case-by-case solution. In this perspective,

the juridical solution that many stakeholders require to go beyond the simple ethical level, we do not have a clash of paradigms, but rather we could be faced with two compatible and coherent points of view.

If these assumptions are accepted, the legal solution that many stakeholders require to go beyond the mere ethical level may already be in our possession. It only remains to be seen in general terms how the evolution of AI needs a general defining-interpretative framework for applications in individual cases. In these terms, an initial analysis of AI law is useful.

C. The EU case: AI Act as an attempt to "update" regulation with reference to standards and prohibition

To confine the legal analysis to EU regulation (without, for now, delving into a comparative analysis with the AI regulations adopted thus far in China and the United Kingdom), it must be noted that the AI Act subjects AI systems to a classification based on three risk levels: unacceptable (and thus prohibited), high, or minimal [19]. Among the various prohibitions is the manipulative use of an AI system, which we may refer to as "algorithmic manipulation", and which, pursuant to Art. 5, pertains to the use of subliminal techniques (letter a) and the exploitation of individuals' vulnerabilities (letter b). From a legal perspective, however, the issue remains the absence of a clear definition of the term "manipulation", as used in the AI Act.

Although not explicitly mentioned in the AI Act, neurotechnologies that rely on AI systems could certainly fall under this prohibition, as confirmed by the reading of Recital 29, which explain by machine-brain interfaces or virtual reality as they allow for a higher degree of control of what stimuli are presented to persons, insofar as they may materially distort their behaviour in a significantly harmful manner. However, as emerges from the text of Art. 5 of the AI Act, manipulative conduct must be potentially harmful to the recipient or others, suggesting that strictly therapeutic applications, as well as the use of devices aimed exclusively at benefiting users without posing risks even to third parties, would appear to be permitted. It remains unclear how this complete absence of "danger" can be qualified, except by entrusting the responsibility to judges in the adjudication of specific cases.

Moreover, "algorithmic manipulation" is penalized under the AI Act even when the harmful or dangerous effects impact individuals (or groups of individuals) other than the direct recipient, and even when the consequences do not pertain to the mental integrity (or moral freedom) of those individuals. For instance, this applies when the manipulated individual is induced to make a decision that is unjustly discriminatory towards another person (or group of persons), in which case the infringement of the moral freedom of the decision-maker is merely a means to harm other interests (or interests belonging to other individuals). Nonetheless, the causal link between the "subliminal" use of technology, behavioural distortion, and harm

(or risk of harm) to oneself or others is subject to a legal-procedural nexus that is not easily established and is therefore of immediate relevance to the community.

D. The fundamental legal core: the person's freedom of conduct–BCI and AI

From a legal-procedural perspective, the use of a BCI device would raise complex questions regarding the identification of conduct as the cornerstone of the *actus reus*. In this context, the example has been formulated of a user with an implanted BCI system who commits the crime of disseminating intimate images of others without the owner's consent (so-called revenge porn), uploading the images online without any physical movement (as would occur when moving a mouse or typing on a keyboard), but rather through a signal transmitted from the brain to the computer [20]. More precisely, in the hypothetical case, the user imagines moving a cursor connected to the BCI device, through which the upload of the prohibited images occurs: this mental act corresponds to an impulse read by a computer, on whose screen the cursor is activated, selecting the upload icon. This represents a clear example of the urgent legal relevance of such scenarios. It would be necessary to ascertain whether this decision stems from the user's free mental activity or from a mental state conditioned by their interaction with the device. In the latter case, it would further need to be determined whether the user could have acted differently, becoming aware of the conditioning and activating their inhibitory mechanisms. We know that neuroscientists teach that the agent becomes aware of their decision to act a fraction of a second after the brain impulse. When the decision results in a bodily movement, we can ignore (or pretend to ignore) this temporal gap. However, when it is not externalised in a physically perceptible act, verifying the possibility of acting differently requires a reconstruction of the dynamics that led to the decision. Even adopting a compatibilist model of freedom (according to which action is never free from conditioning, yet remains free to the extent that the agent recognises the motives behind it as "their own"), we would need to assess to what extent the agent was aware of acting based on their own motives: a verification that varies depending on the type of device and its level of invasiveness. This verification is further complicated by the fact that the functioning of the BCI typically relies on an AI system, capable of learning and performing its operations autonomously, independent of the human agent's perception [21].

E. Perspectives of juridical evaluation or/and evolution

Based on these legal arguments, some directions of philosophical-legal research interlinked with each other are described below.

i) If for some scholars the legal universe would already be ready for the protection of mental integrity resulting from the development of neurotechnology, to complete the protection, the introduction of a case

of direct mental manipulation should be considered, the formulation of which, however, would require a legal definition of the threshold of tolerable risk in the use of neurotechnologies.

ii) Privacy legislation is, in principle, also adequate to preserve the so-called "mental privacy", provided that the exact classification of neural data is clarified. In the (desirable) hypothesis that these data would be accorded more intense protection, on account of their sensitivity, consequential extensions of the relevant apparatus of criminal and/or administrative protection would follow. Moreover, given the possibilities of ascertaining psychic phenomena based on neural data, "mental privacy" should constitute a specific source of procedural guarantees in both civil and criminal law.

iii) The use of neurotechnologies could have effects on the concept of legal "conduct/responsibility", going so far as to modify the notion of "acting subject". Should one agree with the idea that an offence can be committed through the emanation of nerve impulses (case law has already expressed itself on this matter), without bodily movement, there would still remain the problem of ascertaining the actual controllability of such impulses by the user and of providing an adequate sanctioning response.

iv) The research direction that encapsulates the three preceding approaches, as it constitutes their philosophical-legal foundation, now emerges as arguably the most significant. Acknowledging the inadequacy of a national normative approach and the prematurity (and perhaps equal inadequacy) of an international normative approach, it is deemed essential to focus on the hermeneutic-legal dimension to determine, on a case-by-case basis, what degree of "autonomy" can be recognised for the individual human subject "conditioned" by interaction with AI-driven neurotechnologies. This task is not straightforward and will likely need to be limited to preliminary statistical investigations. However, if it aims to yield meaningful results for the coexistence of humans and neurotechnologies, it must be carried out in collaboration with experts in neuroscience, AI, physics, and technology.

BOX. EEG-based BCIs–A methodological perspective on AI integration

EEG-based BCIs open a direct channel from brain activity to external devices, providing a versatile and non-invasive technology widely accessible in many disciplines (medicine, rehabilitation, communication, education, neuroergonomics, gaming, etc.). EEG signals capture the brain's electrical activity with excellent temporal resolution but suffer from low signal-to-noise ratio, non-stationarity, and high inter-subject variability (user-dependent). The integration of AI models, when informed by neurophysiological knowledge, can provide a solution to overcome these limitations.

Signal acquisition and preprocessing. Raw EEG data require a preprocessing step, that includes artifact removal (e.g., ocular, muscular) via independent component analysis, filtering within relevant frequency bands (e.g., alpha 8–12 Hz, beta 13–30 Hz), and normalization. The choice of preprocessing

steps deeply influences the following AI performance.

Feature extraction and representation. Traditional machine learning approaches depend on carefully engineered features extracted from time, frequency, and connectivity domains using methods such as spectral power densities, wavelet, event-related potentials, coherence, phase-locking value, and graph-based connectivity metrics. However, recently AI introduced end-to-end learning, allowing convolutional and recurrent neural networks (RNNs) to extract hierarchical features directly from raw or minimally processed signals.

AI-model architectures. Different AI architectures can be employed for EEG decoding, either applied on the extracted features or on the EEG signals. Convolutional neural networks (CNNs) capture spatial patterns across electrode arrays, while RNNs and their variants (e.g., long short-term memory networks [LSTMs] and gated recurrent units [GRUs]) model temporal dependencies in brain signals. More recently, Transformer-based models, introduced originally for natural language processing, are adapted to EEG time-series for their ability to learn long-range temporal dependencies and contextual relationships. Graph neural networks (GNNs) are especially suited to exploit brain connectivity structures by treating EEG electrodes as nodes and functional/effective connectivity measures as edges to capture the brain's network dynamics and topological organization. In AI model architectures, the classification step translates extracted features or learned representations into mental states or commands. This process critically influences BCI performance metrics such as sensitivity, specificity, and response speed.

Training and validation. The limited availability of large, labeled EEG datasets complicates model generalization, but the increasing use of transfer learning, domain adaptation, and data augmentation techniques is helping to solve these limitations. However, the demand for personalized calibration or adaptive learning strategies to address cross-subject variability is important because the robustness across users must be maintained in real time.

Interpretability and clinical translation. In clinical or real-world applications, AI-driven BCIs require interpretability and trustworthiness. Post-hoc techniques such as Saliency mapping, layer-wise relevance propagation (LRP), and attention-weight visualization help model explain decisions. However, the development of intrinsically explainable models or hybrid architectures that reflect the neurophysiological knowledge is still an open research direction.

The integration of neuroscience principles in BCIs, combined with AI models adapted for EEG data, have a significant potential to improve the sensitivity, specificity, usability, and bit rate of BCI technologies.

Open research questions

- Are the data representative of the task?
- Which features better represent the problem?
- How many features are needed?
- Can the model generalize across users and sessions?
- Is the output interpretable?
- Can the system adapt over time?

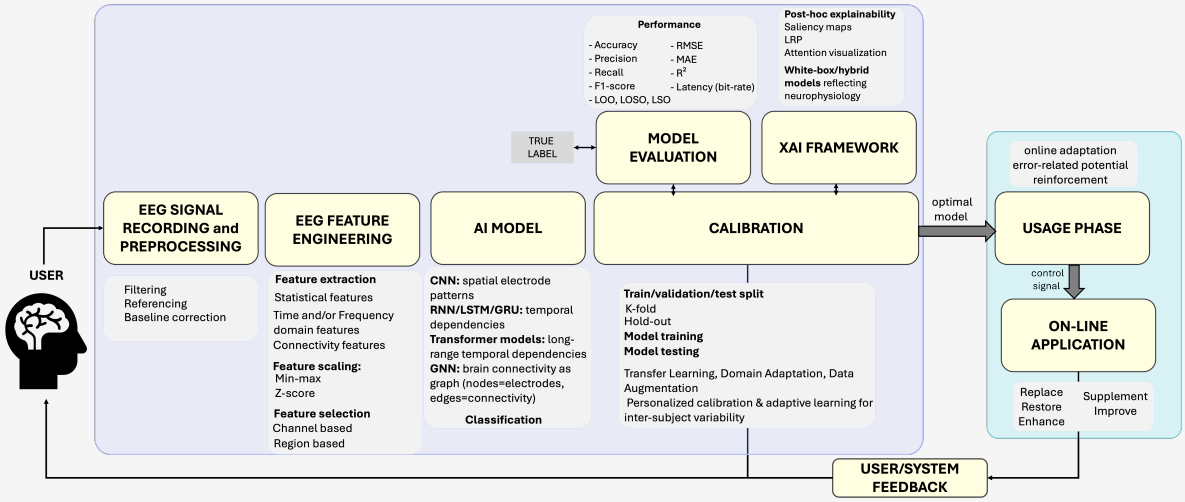

Fig. 2: Overview of the AI-based BCI workflow. Calibration phase (gray-shaded box) and on-line phase (blue-shaded box) integrate AI and are enclosed in the user–system closed-loop at the basis of the BCI framework. The calibration phase includes EEG signal acquisition, preprocessing, and feature extraction, either via engineered metrics or through end-to-end AI. AI models are trained in an XAI framework to optimize decoding performance and perform model selection. The on-line phase includes the "optimal model" calibration in real time to decode the brain function. Users receive real-time feedback, reinforcing motor relearning.
Abbreviations: LOO = leave-one-out, LOSO = leave-one-subject-out, LSO = leave-subject-out, RMSE = root mean square error, MAE = mean absolute error.

IV. The Technical Canvas: Promises and Challenges of Neurothecnology

A. A glimpse into neurotechnology beyond the clinic

As illustrated so far, advanced non-invasive neurotechnologies to image, record and monitor brain activity, such as EEG, fMRI and fNIRS, make it increasingly possible to access individual's mental processes, nowadays often referred to as "brain (or mind) reading". Although this is a fascinating yet controversial topic [7], in the current literature the term mind-reading itself is ill-defined. Several studies focusing on these technologies use this term to simply refer to the possibility to detect and decode the neural signals, such as identifying brain areas associated with memory, visual or motor tasks, as done in

many fMRI protocols. Others, instead, have suggested the ability of neurotechnologies to make a step forward, reading hidden thoughts and consciousness, with all the ethical issues connected with these aspects. The increasing combination of neurotechnologies with AI has further exacerbated these concerns and raised unprecedented questions, mainly concerning whether AI can truly decrypt hidden aspects of the mind, unveiling sensitive features about one's mental activity [22]. Reviews on the topic have recently been published, such as the one by Gilbert & Russo (2024) [23]. In their work, the authors conducted an interpretive content analysis of more than 1000 papers focusing on brain/mind-reading combined with neurotechnologies to investigate prevalent claims, trends, and methods associated with the AI ability to read the mind and identify whether possible evidence of hype is currently present. As already anticipated, their study highlighted inconsistent usage of the term mind-reading in the current literature, and often exaggerated claims about the capacity of current technologies to read the private cognitive states. In this scenario, it is important to mention a recent work by Tang and colleagues [24] demonstrating that the meaning of perceived and imagined stimuli can be decoded from the measured blood oxygenation level dependent fMRI signals into continuous language. In particular, the authors introduce a non-invasive decoder within a BCI language system that reconstructs continuous language from cortical semantic representations recorded using fMRI. Importantly, as recognized and tested by the authors, mental privacy is preserved in this framework because the subject cooperation is required both to train and to apply the decoder. Similarly, another study [25] introduced an AI-based model to decode self-supervised representations of perceived speech from non-invasive EEG or magnetoencephalography (MEG) recordings of a large cohort of healthy individuals. This holds potential applicability in a speech production context as well as for the diagnosis, prognosis and restoration of language processing in non-communicating or poorly communicating patients.

In this panorama, the increasing commercial availability of diverse neurotechnologies, often relatively low-cost and portable, opens new perspectives and reshapes the way humans interact with machines. At the same time it contributes to raise several concerns regarding privacy risks and security of end users, the possible commercialization of brain data, and the unintended consequences on behavior. Such devices can be broadly divided into three main groups: i) BCIs used for self-monitoring, cognitive and performance enhancement/device control; ii) invasive/non-invasive neurostimulation devices; and iii) neurotechnologies applied in neuromarketing. In particular, closed-loop BCIs and other neurotechnologies establish a high degree of integration and mutual influence between the cognitive and affective processes of the human mind and AI systems (please, refer to the Box *EEG-based BCIs–A methodological perspective on AI integration* for an example), and cognitive/performance monitoring and enhancement is one of their most recent applications beyond the clinical environment. Companies are increasingly marketing non-invasive BCIs,

such as devices aimed at improving focus, relaxation, or even sleep quality via neurofeedback training (e.g., MindMaze, Emotiv, and Neurosky). Moreover, consumer-grade portable EEGs, such as the Emotiv Epoc+ and Neurosky Mindwave, can monitor users brain activity by accessing raw EEG data. These devices can be used for tracking attention levels or controlling virtual objects. For example, the P300 event-related potential linked to decision-making processes can be exploited to extract personal and financial information from users without their awareness or consent. In one experiment, users were exposed to specific visual stimuli designed to capture sensitive data, such as PIN codes or home addresses, showing a significant risk of information leakage [1]. Similarly, a BCI game developed at the University of Washington called Flappy Whale showed that EEG and MEG responses to subliminal stimuli could be used to extract private information, such as financial data or even personal beliefs [1]. Looking ahead, one particularly frontier in neurotechnology is the brain-to-brain interfaces (BBIs) which aim to establish a direct communication bridge between two or more human brains. Preliminary experiments have demonstrated the feasibility of sharing sensorimotor information, i.e. motor imagery of feet or hands transformed in a binary code for visual stimuli with transcranial magnetic stimulation, where communication between the BCI and BBI components is mediated by the internet [26] or simple decisions across multiple individuals, opening discussions about the concept of memory sharing. In these scenarios, external implanted memories or quasi-memories could become indistinguishable from a recipient's authentic recollections, raising questions about autonomy, self-identity, and consent. If realized on a broader scale, BBIs could open to new modes of collaboration and empathy but also carry a high risk of coercive manipulation or malicious tampering with personal identity.

While all these paradigms hold great potential in healthcare and especially off-clinics applications, they require a careful selection and definition of the most appropriate methodology to solve the different tasks. For example, the possibility to detect adversarial mental states or translate cognitive intentions into executable commands/actions requires modeling complex, nonlinear relationships in the brain activity, mostly measured with EEG. The non-stationary and high-dimensional nature of EEG signals makes deep learning (DL) particularly suitable for this task, allowing to extract hierarchical and task-relevant features directly from raw data. In particular, CNNs are widely adopted in motor imagery-based BCIs, where spatial filters are applied to EEG topographies to isolate discriminative patterns. For instance, EEGNet demonstrated strong performance in classifying motor imagery across subjects by using compact convolutional kernels adapted to the spatial organization of EEG electrodes [27].

CNNs have been widely used also for tasks such as EEG-based emotion recognition, exploiting their ability to extract local spatial features from multi-channel signals. A good example is CNNs for EEG-based emotion recognition, incorporating brain connectivity features. Instead of using only the activity recorded

at specific electrodes, these approaches integrate spatial and functional information reflecting synchronous activity among brain regions. In the work by Moon et al. [28], CNNs are trained on connectivity matrices derived from measures such as the Pearson correlation coefficient, phase-locking value, and phase lag index. The matrices encode pairwise relationships between electrodes and are displayed as topographic or adjacency matrices, allowing the model to exploit local signal properties and large-scale network interactions. This design is motivated by the need to mitigate issues related to EEG signals, such as low signal-to-noise ratio and inter-subject variability, by embedding neurophysiological priors directly in the learning architecture. The method was evaluated using a dataset, which includes EEG data recorded while viewing emotional video stimuli, annotated with valence, arousal, and dominance scores. Results show that connectivity-informed representations improve classification accuracy compared to models based on single-channel features.

DL methods as RNNs or their gated variants, such as LSTM networks, are often used on the EEG time series to capture the temporal dependencies and long term patterns in the data. Applications include drowsiness detection in driving scenarios [29] and cognitive workload estimation in multitasking activity [30]. Transformer-based models have also recently been introduced to EEG decoding tasks, proposing an alternative to sequential ones by exploiting self-attention mechanisms to shape long-range temporal interactions. The EEG–transformer architecture has shown promising results in decoding emotional states and recognizing user intentions from EEG recordings [31]. Concerns regarding data efficiency and interpretability remain, particularly in clinical or low-resource settings. For this reason, current research is exploring pretraining strategies, self-supervised learning, and task-specific data augmentation to improve generalization and transparency.

When considering some examples of possible medical-related scenarios, RNNs and LSTM approaches have shown success in EEG-based seizure detection and prediction by modeling temporal dependencies [32]. However, they can struggle with very long sequences due to vanishing gradient problems. On the other hand, transformer-based models process the full sequence in parallel, avoiding the sequential computation bottleneck of RNNs. This allows for more efficient processing of long-term EEG recordings and better captures bidirectional contexts. Hybrid models that combine CNN and LSTM have been designed to capture both spatial and temporal features of EEG data [33] but require careful architectural design to balance the two components. Conversely, the multi-head attention mechanism of transformers allows the model to attend to different spatial and temporal patterns simultaneously, making it well-suited for analyzing brain connectivity.

B. A brief look at explainability in extraclinical AI-powered neurotechnologies

Recently, large international consortia have been established including Trustworthy and Responsible AI Network (TRAIN; https://train4health.ai), Fairness Universality Traceability Usability Robustness Exaplainability (FUTURE)-AI (https://future-ai.eu), and Artificial Intelligence Safety Institute Consortium (AISIC; https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute-consortium-aisic) aiming to provide guidance on safe and trustworthy AI. They have identified several guiding principles that AI systems should fulfill. Among these, the highest consensus is on fairness, universality, traceability, usability, robustness, and explainability. Fairness indicates that the model output should not be dominated by the largest data sample so to be fair toward the minor data sample as well (e.g., sex, ethnicity and age). Universality calls for outcome of AI systems that are generalizable to other cohorts and centers which were not included in the stage of model development. Traceability imposes that the process of the AI system starting from development to the implementation and validation should be properly documented. Such document would help to audit the tool performance and support the detection of potential risks. Usability refers to AI systems that can be used safely and can achieve the intended outcome with effective user's advantage. Robustness measures to what extent the AI system is robust and able to perform well with unexpected variations in the used data. Explainability (XAI) refers to the tools and algorithms that are designed to make AI systems transparent by revealing the factors and input data mostly affecting the outcome of AI system. More in detail, XAI has the ability to reveal what group of pixels in an image or specific input features in tabular data mostly impacted the model outcome.

It has already been mentioned in section II-C the potential of XAI to make AI-powered neurotechnologies more ethical. When AIs are equipped with strategies able to highlight the working mechanisms which have conducted them to the result, humans have the means to preserve their autonomy, dignity, and interests. Indeed, the XAI methods are able to explain AIs by opening the black box in which they are framed by using strategies that span a wide range of complexity from the "simplest", like ranking the importance of each variable participating to the modeling of a problem when tabular data are available, to the "more complex", like visualizing which specific pixels of an image have contributed to the outcome of a DL neural network task. As an example if, while testing an image with an AI, the pixels enlighted by the XAI method (that are supposed to be those much contributing to the output) are mainly located in the background, this should trigger the user on the evidence that the AI employed has mainly grounded on noise (and not a plausible rationale) to perform a decision. By doing another example, it would be good practice to check for potential ethical biases an XAI used to deal with societal issues that comes out "ethnicity" as the most impactful variable on the output.

Many XAI methods exist, and differences among them can be relied on the properties that determine the so-called XAI taxonomy and that are here reported: XAI can be ante-hoc when it is built in the AI model or it can be post-hoc when it is applied to AI models after the training and test; in addition, it can be model-specific when it is built for a specific model or model-agnostic when it is applied to all AI models; XAI can be at global level to explain AI models for all samples or at local level where it explains a specific sample on the model [34].

Several methods were then proposed to assess the quality of XAI outcome that are: human and application-grounded evaluation, proxy-grounded evaluation, literature-grounded evaluation and guideline-grounded evaluation. Human and application-grounded evaluation indicates that the outcome of XAI should be assessed by the experts in the domain. Several measures were developed for this kind of evaluation including simplicity, completeness, plausibility and complexity. Proxy-grounded evaluation refers to evaluating the quality of XAI using some sort of proxies and statistical analysis including sensitivity, selectivity, correctness, normalized movement rate and computation time. Literature-grounded evaluation assesses the quality of XAI by comparing its outcome with what is already known in the literature. Finally, guideline-grounded evaluation indicates that the outcome of XAI is evaluated using a pipeline where the input is the outcome of XAI and there are several criteria to pass before adopting and deploying it in healthcare. One of the big limitations of the XAI is that there is not a standard or a formally agreed definition of what does explainability mean and what should cover. Moreover, XAI usage implications might include the risk of misuse when it is used to explain a case where the model was not designed for. In addition, XAI method might be used to explain the outcome of a model where the training and test data have different ethnicity characteristics and eventually violate the generalizability aims. One other limitation of the XAI is that it does not help to reveal the liability toward a wrong decision made by the model when the outcome is explained incorrectly. In addition, the designed XAI methods are mostly for research context and cannot be extended in practice because most of these methods were designed by technical people to explain a model that might not be appropriate to be adopted on a large scale. Finally, current XAI methods might work with current models using current neurological data acquisition, parameters, magnetic fields, and type of scanners. However, these technologies are evolving, questioning the longevity of current XAI methods [35].

In literature, when targeting the specific topic of applications where neurotechnologies and AI are combined to perform tasks such as “mind-reading” and the others mentioned until this point, the choice of the XAI method largely depends on the neurotechnology deployed [34]. We refer the reader to Table S1 in the Supplementary Material for an overview of the most common post-hoc XAI methods, with specific insights on their properties, data format they explain, and practical hints.

When considering XAI applications within the overall "mind-reading" task, there are a few studies that have been recently proposed. Thomas et al. [36], in particular, analyzed fMRI data from three different datasets in order to decode mental states relying on 3D-CNNs. Of note, they observed differences between the explanations provided by the many XAI attribution approaches used depending on their working mechanism. The observations led the authors to derive recommendations to the readers. The first was to use reference-based attribution methods like DeepLift SHAP, DeepLift or the Integrated Gradients whether the main interest in the XAI deployment is the understanding about the process which conducted the DL model to the resulting mental state. Indeed, they found that the explanations provided by these methods were the most faithful as they were able to identify the brain regions whose activity is the key for the accurate mental state decoding. The second recommendation was to use, instead, XAI approaches based on sensitivity (e.g., Gradient Analysis or, more commonly, Saliency) or backward decomposition (e.g., Guided Backpropagation, Guided GradCam) whether the intention is more strictly focused on finding the associations between mental states and brain activity. Such a preference should be imputed to the highest agreement reached by the explanations provided by these methods when compared with results from the standard general linear model or meta-analysis.

In addition to mental state decoding, XAI has been applied to developmental cognitive neuroscience. Andreu-Perez and colleagues [37] proposed an empowerment of the multivariate pattern analysis to explore fNIRS data that made it more explainable. The so-called xMVPA inference mechanism is, thus, an ante-hoc, model-specific XAI method expressly tailored for fNIRS data that returns textual explanations corresponding to the patterns (and relative dominance scores) found in the brain haemodynamics data. These explanations evidence configurations of inactive, active or very active channels in correspondence to a stimulus.

XAI has been also largely applied to EEG-based frameworks, e.g. to study emotion perception under many aspects from recognition to regulation, to investigate many facets of attention that can be related, for example, to autonomous driving (i.e., drowsiness detection, driver's hand-foot coordination, arousal and awareness) or to classify motor-imagery tasks (with a clear implication in BCI systems). The XAI methods employed in these cases included both visualization and feature ranking approaches. Among the formers there were Integrated Gradients, LRP, and LayerCAM, respectively to verify the faithfulness of the models proposed, visualize brain signals determining an awareness and arousal state, perform post-hoc connectivity analyses. The latters, instead, mainly reduced to SHAP, which was typically used to define the relevance of single-channel-derived signals or features extracted from them. As an example, Liew et al. [38] faced the emotion recognition problem by finding clusters of data samples representing more condensed memory templates that were subsequently used to train boosted decision trees then

investigated for feature ranking by means of SHAP. Considered features were, e.g., EEG band power, heart rate variability, and feature interactions.

XAI has been also used downstream the inference from a deep neural network of the relationship between cortical source estimation derived from MEG collected during directional reaching movements and along with kinematic parameters like acceleration, velocity, and position [39]. The Integrated Gradients applied post-hoc to such a deep neural network enabled the visualization of the cerebral cortices differently involved in the processing of the three kinematics parameters, potentially moving a step forward in the improvement of BCI applications.

An interesting point of view was found in the work of Kim et al. [40], who proposed an explanatory efficacy metric basically based on the feedback of users employing an XAI-based movie recommendation system, and to whom was requested to evaluate the efficacy of the explanations received. Such an investigation was conducted also by equipping the users with EEG, finding that hemispheric differences of the neural correlates were significant with 62.4% accuracy when dealing with systems performing perceptually driven iterative tasks with high explanatory efficacy. Such a finding revealed the potential of neurophysiology as a measure of explanatory efficacy by itself, with possible implications in human–AI relation.

C. Data, bias and privacy in neurotechnology

Neurotechnology applications rely on diverse forms of brain data. As previously introduced, handling these data raises challenges in collection, bias, and privacy. Data can be considered as lifeblood: advances in neurotechnology (from BCIs that help paralyzed patients communicate to digital psychiatry apps that monitor mood) depend critically on data. Signals from EEG headsets, MRI scans, and even smartphone usage patterns become the raw material for algorithms that decode brain activity or predict mental health status. However, acquiring high-quality neurodata is often difficult. Brain data tend to be noisy, high-dimensional, and limited in sample size, especially for conditions like rare neurological disorders. Data must be recorded in controlled settings (e.g., hospitals or research labs), making large-scale collection expensive and time-consuming. Moreover, neurotechnology data come from human subjects, so ethical and practical constraints limit how they can be gathered and shared. For instance, a single clinic may only see a few epilepsy patients with implanted devices, or a BCI study might involve only a dozen volunteers: not enough to capture the full variability of the population. To improve generalization, researchers need to pool data across sites, but this raises thorny issues of data sharing, standardization, and consent. The neuroimaging community has embraced data sharing mandates, spurring big repositories and consortia, yet “efficient and effective data sharing still faces several hurdles”, especially for sensitive patient data. Negotiating

data-use agreements, transferring massive brain scan files, and coordinating analysis protocols across institutions remain non-trivial challenges [41]. These data dilemmas must be solved for neurotechnology to reach its potential in real-world clinical and consumer applications.

1) Hidden biases in brain data and models: Bias is a central concern when building neurotechnology models. Who's in the dataset can deeply influence what a model learns (and whom it works for). Neurotech studies, like much of psychology and neuroscience, have historically suffered from W.E.I.R.D. sampling (participants that are Western, Educated, Industrialized, Rich, and Democratic). Such narrow samples can yield misleading conclusions. A University of California San Francisco study demonstrated that non-representative brain imaging datasets can "significantly distort" findings on brain development (https://www.ucsf.edu/news/2017/10/408651/brain-imaging-results-skewed-biased-study-samples). In that case, a pediatric MRI dataset skewed toward higher socioeconomic status led to incorrect estimates of how quickly children's brains mature, which could have misinformed our understanding of normal development and mental health conditions like autism. The lesson is clear: if our data lack diversity in age, gender, ethnicity, or health status, the models we train may not generalize. An AI that analyzes EEG patterns to detect depression, for example, might underperform on older adults if it was trained mostly on college students. Likewise, a BCI speller calibrated on healthy users might fail for some patients because of differences in neural signals (a phenomenon known as "BCI illiteracy", where 10–30% of users never achieve control of the BCI due to individual variability). Biases can also creep in through technical confounds. Different hospitals may use different EEG equipment or MRI scanners; if a model isn't carefully designed, it might latch onto these site-specific signatures rather than true neurological patterns. As discussed, XAI is emerging as a crucial tool to combat such biases. By peering into a model's decision-making, researchers can check if an algorithm is focusing on meaningful neural signals or just artifacts. Interpretable model designs can even enhance performance: one DL study added a "brain network" layer to make sense of fMRI data, and found that this architectural transparency improved accuracy in distinguishing patients with schizophrenia, autism, and dementia [42]. More generally, opening the black box builds trust: clinicians are more inclined to use a diagnostic algorithm if it can explain why it flags a certain patient. In neurotechnology, ensuring fairness may mean actively curating diverse datasets and using techniques like Saliency maps or Attention mechanisms to verify that brain models aren't inadvertently discriminating. As the European Data Protection Supervisor cautions, "if a tool is trained exclusively on a group defined by [some] characteristic, it [will] not correctly recognize patterns in individuals that do not share that characteristic" (https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2024-06-03-techdispatch-12024-neurodata_en). Mitigating bias is not just an academic exercise: it directly impacts patient outcomes and equity of care.

2) Privacy-preserving neurotech: from data silos to federated learning: Federated learning allows keeping the data at its source and moving the analysis to the data. Different methods have been proposed [41] among which a pioneering platform, COINSTAC (Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation), enables scientists to collaborate on brain data analyses without sharing raw data [43]. Each participating site (hospital, lab, or even a patient's device) retains its own dataset locally and runs computations on it; only the resulting learned parameters or summary stats are exchanged. Iterating in this way, COINSTAC can produce a result "as if the entire data were in hand"—but no single party ever sees all the data [43]. It's a bit like training a collective AI model where the knowledge travels, not the data. A federated learning platform for neuroimaging (COINSTAC) supports multiple modes of participation, including sites with local data and secure "vault" servers for sensitive datasets. By analyzing data in place (behind firewalls) and only sharing computed updates, such approaches protect privacy while enabling large-scale collaboration [43]. Privacy-preserving techniques go beyond just federation. Often, additional encryption or differential privacy noise is applied to the shared updates, ensuring that even the partial information exchanged can't be reverse-engineered to reveal individual records. In the realm of digital psychiatry, for instance, researchers are exploring federated learning for mood prediction models so that personal phone sensor data never leaves the user's device, alleviating fears of exposing one's daily routines or social interactions. An early study on depression detection found that a federated approach could train accurate classifiers on text data from patients, all while keeping the raw texts private. Beyond the technical benefits, these approaches build trust: patients and institutions become more willing to contribute data when they know it stays under their control. As a result, previously "closed" data silos (locked away due to privacy law or proprietary concern) can join the collective effort. Federated platforms have already demonstrated success in multi-site neuroimaging studies, effectively increasing sample sizes and statistical power for detecting brain patterns in disorders while maintaining compliance with data protection policies. There is still room to improve (standards are needed to ensure interoperability and rigorous privacy audits) but the trajectory is set. Privacy-by-design in neurotechnology is turning what once was a zero-sum trade-off (accuracy versus confidentiality) into a win-win, where we get robust models and respect patients' rights.

To ground these ideas, different scenarios that illustrate both pitfalls and solutions in handling neurodata responsibly can be envisioned. In a real-life context, bias in a mood prediction app could be developed using smartphone data (texts, voice tone, sleep patterns) to predict depressive episodes. Initial trials looked promising, but after launch the company finds the app misses warnings for older adults. The reason? The training data came mostly from young users with active phone habits. This sampling bias meant the AI learned patterns that didn't apply to seniors. Recognizing this, the team partners with a geriatric clinic to

collect a more representative dataset and retrains the model. The updated app performs much better across ages, a real-life lesson in why inclusive data collection matters for AI in mental health. It also spurred the team to add an XAI module that highlights which behavior changes trigger an alert, so clinicians and users can understand and trust the warnings (and ensure they're not based on irrelevant quirks).

In a clinical context, privacy issues could regard modern implanted devices for epilepsy (sometimes called responsive neurostimulators) that can detect seizures and even stimulate the brain to prevent them. These implants generate continuous streams of neural data. In a hypothetical extension, imagine an implant manufacturer that wants to improve its seizure detection algorithm by learning from all its devices in the field. A naive approach would upload every patient's EEG data to the cloud for analysis: a huge privacy risk and likely impossible under health data regulations. Instead, the company implements an on-device machine learning update. Each implant locally refines the detection model based on the patient's data and sends only the updated model parameters (not the raw EEG) back to the company's servers. The servers aggregate updates from hundreds of devices to yield a better global model, which is then sent back to each implant as a firmware update. In this way, patients benefit from the collective intelligence of the entire device network (more accurate seizure prediction) with minimal privacy exposure. This approach also protects against cyber intrusions: since there's no central trove of raw brain data, there's less incentive and opportunity for hackers to steal sensitive neural information.

D. Challenges with emerging technologies

While federated learning and XAI represent significant strides in addressing neurotechnology's data challenges, several emerging paradigms promise transformative potential, yet introduce novel complexities that demand proactive solutions.

i) Foundation models promise neurotechnology revolution but face dual challenges: their massive data needs exacerbate collection barriers and centralize development power, while inherited societal biases risk propagating harmful clinical stereotypes. The solution may lie in neuro-specific models with curated data and modular architectures enabling targeted bias auditing.

ii) Multimodal convergence of neural, behavioral, contextual, and genomic data enables holistic profiling but intensifies privacy risks through cross-modal re-identification (e.g., neural data linked to identity via gait patterns). Temporal misalignment between modalities (e.g., fMRI's slow hemodynamics versus millisecond-precise eye-tracking) creates additional hurdles. Solutions may involve differential privacy, homomorphic encryption, and improved multimodal approaches.

iii) Self-supervised learning (SSL) leverages inherent data structure (e.g., contrastive EEG coding) to reduce dependency on scarce labeled data-especially valuable for rare disorders. However, SSL risks

amplifying artifacts (e.g., misinterpreting MRI noise as pathology) and encoding sensitive attributes (e.g., demographics) as covert biases. Mitigation requires artifact-invariant augmentation and adversarial filtering. Alternatively, synthetic data offers promise but faces simulator development challenges.

iv) Neuromorphic chips could enable ultra-efficient neural signal processing directly on wearables/implants. Combined with federated learning, they facilitate real-time personalization without cloud dependency (critical for closed-loop neurostimulation). However, hardware constraints introduce bias: limited memory prioritizes common neural patterns over rare anomalies, while energy tradeoffs between model complexity and battery life disproportionately affect users needing high-precision decoding. Standardized benchmarking for resource-constrained neuro-AI is thus urgent.

Though future neurotechnologies will inevitably emerge, they must all navigate the persistent triad of data scarcity, bias amplification, and privacy erosion (not as isolated challenges, but as dynamically interacting constraints). This interdependence necessitates holistic governance frameworks that evolve alongside technological innovation.

V. Conclusions

Neurotechnology sits at the intersection of cutting-edge data science and the most sensitive personal information: our brain signals and mental health. Data, bias, and privacy are thus not side issues: they determine whether neurotech innovations can be translated safely and effectively into practice. Ensuring we have enough high-quality data (and the means to share it), guarding against biases that could deepen health disparities, and protecting patients' privacy are all essential for ethical and reliable neurotechnology. The community is rising to these challenges with creative solutions: from frameworks like COINSTAC that enable "analysis without data sharing", to interpretability techniques that turn black box models into glass boxes for scrutiny, to emerging data standards and privacy laws tailored for neural data (https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2024-06-03-techdispatch-12024-neurodata_en). The practical payoff is already visible in applications like BCIs and digital psychiatry, which are becoming more robust and trustworthy. Moving forward, interdisciplinary collaboration will be key: engineers, neuroscientists, ethicists, and clinicians must continue to work together to handle neurodata responsibly. With thoughtful design and governance, we can harness the power of brain data to improve lives without sacrificing individual rights or scientific integrity. The brain may indeed be the "last frontier of privacy", but with the right tools and mindset, neurotechnology can advance that frontier in a way that benefits everyone. In conclusion, it is essential to guarantee full compliance with existing regulatory frameworks and full disclosure concerning how neurodata are collected and stored, and where. The full integration of neurotechnologies with AI is a frontier likely to be reached

soon. In order not to fall prey of a "surprise effect", we can act at different levels. First, imagining new "neurorights" and implementing them. Even though the case for exceptional ethical regulation of neurodata is not compelling, giving it is not directly inferred from the exceptional nature of at least some kinds of neurodata, a prudential attitude would not rule out entirely imagining them and how to implement them. This is especially true, given that neurotechnology marks the emergence of new disciplinary social forms. Then, a biopolitical reflection is necessary. We should think of these new technologies as new disciplinary forms, calling for a new field of research, that is "neurobiopolitics". As such, a public approach (both in reflection and legislation) should be favoured over an individualistic one. Third, legal materiality: any attempt at regulation should follow clear boundary conditions, while at the same time considering neural activity at all levels, including the subliminal one, as an essential characteristic of human decision process. Finally, given that integrated AI-neurotechnological systems (e.g. BCIs) are based on inductive inference, they could be de facto considered (and reduced to) AI systems; accordingly, ethical reflection on these systems could be considered a special branch of AI ethics.

ACKNOWLEDGMENT

C.C. and A.F. acknowledge support by the European Union—NextGenerationEU, in the framework of the "Future Artificial Intelligence Research (FAIR)" programme for the project AIPRAH—AI-powered Robotic Assistant for Healthcare FAIR PE00000013–CUP C63C22000770006. A.F., S.F.S., and G.M. acknowledge support by Programma Regionale Veneto del Fondo Europeo di Sviluppo Regionale (PR Veneto FESR) 2021–2027 (INTELLI-MI, project–reference code 24729_002214). S.F.S, L.B., I.B.G., and G.M. acknowledge support by the Ministry of Education, Universities and Research [Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) D.M. 737/2021, "AI4Health: empowering neurosciences with eXplainable AI methods" project]. L.B., I.B.G., and G.M. acknowledge support by the Ministry of University and Research (MUR) (Call for Projects of Significant National Interest, PRIN 2022, project–reference code 202292PHR2). A.M.S. acknowledges support by the Leicester City Football Club (LCFC). I.B.G. acknowledges support by the European Union—NextGenerationEU, in the framework of the iNEST—Interconnected Nord–Est Innovation Ecosystem (iNEST ECS00000043–CUP B43C22000450006). S.P. acknowledges support by the National Institutes of Health (R01MH129047) and National Science Foundation (2112455).

REFERENCES

[1] Marcello Ienca, Pim Haselager, and Ezekiel Emanuel. "Brain leaks and consumer neurotechnology". In: Nature Biotechnology 36 (2018), pp. 805–810. DOI: 10.1038/nbt.4240.

[2] L. Floridi et al. “Ai4people – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations”. In: Minds and Machines 28.4 (2018), pp. 689–707. DOI: 10.1007/s11023-018-9482-5.

[3] Oliver Müller and Stefan Rotter. “Neurotechnology: Current Developments and Ethical Issues”. In: Frontiers in Systems Neuroscience Volume 11 - 2017 (2017). DOI: 10.3389/fnsys.2017.00093.

[4] Françoise Baylis. ““I Am Who I Am”: On the Perceived Threats to Personal Identity from Deep Brain Stimulation”. In: Neuroethics 6.3 (2013), pp. 513–526. DOI: 10.1007/s12152-011-9137-1.

[5] Yukari Yamada et al. “Monitoring reaction time to digital device in the very-old to detect early cognitive decline”. In: npj Aging 10 (2024), p. 40. DOI: 10.1038/s41514-024-00167-z.

[6] A. Clark and D. Chalmers. “The extended mind”. In: Analysis 58.1 (1998), pp. 7–19. DOI: 10.1111/1467-8284.00096.

[7] Marcello Ienca and Roberto Andorno. “Towards New Human Rights in the Age of Neuroscience and Neuropathology”. In: Life Sciences, Society and Policy 13.5 (2017), pp. 1–27. DOI: 10.1186/s40504-017-0050-1.

[8] Christopher Burr, Nello Cristianini, and James Ladyman. “An Analysis of the Interaction Between Intelligent Software Agents and Human Users”. In: Minds and Machines 28.4 (2018), pp. 735–774. DOI: 10.1007/s11023-018-9479-0.

[9] Jack Wilkinson et al. “Time to reality check the promises of machine learning-powered precision medicine”. In: The Lancet Digital Health 2.12 (2020), e677–e680. DOI: 10.1016/S2589-7500(20)30200-4.

[10] M. Ienca and P. Haselager. “Hacking the Brain: Brain–Computer Interfacing Technology and the Ethics of Neurosecurity”. In: Ethics and Information Technology 18.2 (2016), pp. 117–129. DOI: 10.1007/s10676-016-9398-9.

[11] Tara L. White and Meghan A. Gonsalves. “Dignity neuroscience: universal rights are rooted in human brain science”. In: Annals of the New York Academy of Sciences 1505.1 (2021), pp. 40–54. DOI: 10.1111/nyas.14670.

[12] Daniel Susser and Laura Y. Cabrera. “Brain Data in Context: Are New Rights the Way to Mental and Brain Privacy?” In: AJOB Neuroscience 15.2 (2024), pp. 122–133. DOI: 10.1080/21507740.2023.2188275.

[13] W.J. Johnson. “Beyond Substance. Structural and Political Questions for Neurotechnologies and Human Rights”. In: AJOB Neuroscience 15.2 (2024), pp. 134–136. DOI: 10.1080/21507740.2024.2326915.

[14] Dara Hallinan et al. “Neuroexceptionalism: Framing an emergent debate”. In: Available at SSRN: https://ssrn.com/abstract=3909816 (2021).

[15] Christoph Bublitz, Jennifer Chandler, and Marcello Ienca. “Human–Machine Symbiosis and the Hybrid Mind: Implications for Ethics, Law and Human Rights”. In: The Cambridge Handbook of Information Technology, Life Sciences and Human Rights. Ed. by Marcello Ienca et al. Cambridge Law Handbooks. Cambridge University Press, 2022, 286–303.

[16] Nora Hertz. “Neurorights – Do we Need New Human Rights? A Reconsideration of the Right to Freedom of Thought”. In: Neuroethics 16 (2022), p. 5. DOI: 10.1007/s12152-022-09511-0.

[17] Thibault Moulin. “‘I Will Control Your Mind’: The International Regulation of Brain-Hacking”. In: San Diego International Law Journal 24 (2022), p. 65.

[18] Pablo López-Silva. “The Concept of Mind in the Neuroprotection Debate”. In: Protecting the Mind: Challenges in Law, Neuroprotection, and Neurorights. Ed. by Pablo López-Silva and Luca Valera. Cham: Springer International Publishing, 2022, pp. 9–18. DOI: 10.1007/978-3-030-94032-4_2.

[19] Stephen Rainey et al. “Is the European Data Protection Regulation sufficient to deal with emerging data concerns relating to neurotechnology?” In: Journal of Law and the Biosciences 7.1 (2020), lsaa051. DOI: 10.1093/jlb/lsaa051.

[20] Pim Haselager and Giulio Mecacci. “Is Brain Reading Mind Reading?” In: Neurolaw and Responsibility for Action: Concepts, Crimes, and Courts. Ed. by BebhinnEditor Donnelly-Lazarov. Cambridge University Press, 2018, 182–192.

[21] Nicole A Vincent, Thomas Nadelhoffer, and Allan McCay. Neurointerventions and the Law: Regulating Human Mental Capacity. Oxford University Press, Mar. 2020. ISBN: 9780190651145. DOI: 10.1093/oso/9780190651145.001.0001.

[22] Cohen Marcus Lionel Brown. “Neurorights, Mental Privacy, and Mind Reading”. In: Neuroethics 17.2 (2024), p. 34. DOI: 10.1007/s12152-024-09568-z.

[23] Frederic Gilbert and Ingrid Russo. “Mind-reading in AI and neurotechnology: Evaluating claims, hype, and ethical implications for neurorights”. In: AI and Ethics 4.3 (2024), pp. 855–872. DOI: 10.1007/s43681-024-00514-6.

[24] Jerry Tang et al. “Semantic reconstruction of continuous language from non-invasive brain recordings”. In: Nature Neuroscience 26.5 (2023), pp. 858–866. DOI: 10.1038/s41593-023-01304-9.

[25] Alexandre Défossez et al. “Decoding speech perception from non-invasive brain recordings”. In: Nature Machine Intelligence 5.10 (2023), pp. 1097–1107. DOI: 10.1038/s42256-023-00714-5.

[26] Carles Grau et al. “Conscious brain-to-brain communication in humans using non-invasive technologies”. In: PloS one 9.8 (2014), e105225. DOI: 10.1371/journal.pone.0105225.

[27] Vernon J Lawhern et al. “EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces”. In: Journal of neural engineering 15.5 (2018), p. 056013. DOI: 10.1088/1741-2552/aace8c.

[28] Seong-Eun Moon, Soobeom Jang, and Jong-Seok Lee. “Convolutional neural network approach for EEG-based emotion recognition using brain connectivity and its spatial information”. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018, pp. 2556–2560. DOI: 10.1109/ICASSP.2018.8461315.

[29] Sadegh Arefnezhad et al. “Applying deep neural networks for multi-level classification of driver drowsiness using Vehicle-based measures”. In: Expert Systems with Applications 162 (2020), p. 113778. DOI: 10.1016/j.eswa.2020.113778.

[30] Debashis Das Chakladar et al. “EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm”. In: Biomedical Signal Processing and Control 60 (2020), p. 101989. DOI: 10.1016/j.bspc.2020.101989.

[31] Young-Eun Lee and Seo-Hyun Lee. “EEG-transformer: Self-attention from transformer architecture for decoding EEG of imagined speech”. In: 2022 10th International winter conference on brain-computer interface (BCI). 2022, pp. 1–4. DOI: 10.1109/BCI53720.2022.9735124.

[32] Xiang Lu et al. “An epileptic seizure prediction method based on CBAM-3D CNN-LSTM model”. In: IEEE Journal of Translational Engineering in Health and Medicine 11 (2023), pp. 417–423. DOI: 10.1109/JTEHM.2023.3290036.

[33] Jian Liang et al. “Predicting seizures from electroencephalography recordings: a knowledge transfer strategy”. In: 2016 IEEE International Conference on Healthcare Informatics (ICHI). 2016, pp. 184–191. DOI: 10.1109/ICHI.2016.27.

[34] Melkamu Mersha et al. “Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction”. In: Neurocomputing 599 (2024), p. 128111. DOI: 10.1016/j.neucom.2024.128111.

[35] Ahmed M Salih et al. “A perspective on explainable artificial intelligence methods: SHAP and LIME”. In: Advanced Intelligent Systems 7.1 (2025), p. 2400304. DOI: 10.1002/aisy.202400304.

[36] Armin W. Thomas, Christopher Ré, and Russell A. Poldrack. “Benchmarking explanation methods for mental state decoding with deep learning models”. In: NeuroImage 273 (2023), p. 120109. DOI: 10.1016/j.neuroimage.2023.120109.

[37] Javier Andreu-Perez et al. “Explainable artificial intelligence based analysis for interpreting infant fNIRS data in developmental cognitive neuroscience”. In: Communications Biology 4 (2021), p. 1077. DOI: 10.1038/s42003-021-02534-y.

[38] Wei Shiung Liew, Chu Kiong Loo, and Stefan Wermter. “Emotion Recognition Using Explainable Genetically Optimized Fuzzy ART Ensembles”. In: IEEE Access 9 (2021), pp. 61513–61531. DOI: 10.1109/ACCESS.2021.3072120.

[39] HongJune Kim, June Sic Kim, and Chun Kee Chung. “Identification of cerebral cortices processing acceleration, velocity, and position during directional reaching movement with deep neural network and explainable AI”. In: NeuroImage 266 (2023), p. 119783. DOI: 10.1016/j.neuroimage.2022.119783.

[40] Byung Hyung Kim et al. “Improved Explanatory Efficacy on Human Affect and Workload Through Interactive Process in Artificial Intelligence”. In: IEEE Access 8 (2020), pp. 189013–189024. DOI: 10.1109/ACCESS.2020.3032056.

[41] Walter Riviera, Ilaria Boscolo Galazzo, and Gloria Menegaz. “FeLebrities: A User-Centric Assessment of Federated Learning Frameworks”. In: IEEE Access 11 (2023), pp. 96865–96878. DOI: 10.1109/ACCESS.2023.3312579.

[42] Usman Mahmood et al. “Through the looking glass: Deep interpretable dynamic directed connectivity in resting fMRI”. In: NeuroImage 264 (2022), p. 119737. DOI: 10.1016/j.neuroimage.2022.119737.

[43] Sergey M. Plis et al. “COINSTAC: A Privacy Enabled Model and Prototype for Leveraging and Processing Decentralized Brain Imaging Data”. In: Frontiers in Neuroscience 10 (2016). DOI: 10.3389/fnins.2016.00365.

BIOGRAPHIES

Carlo Chiurco (carlo.chiurco@univr.it) is Associated Professor of Ethics at the University of Verona (Department of Engineering for Innovation Medicine), where he leads the research center ERMES—Ethics, Right, Medicine and Science on applied ethics, with a focus on bioethics/medical ethics and the ethics of AI. He has overseen several projects on medical ethics and medical humanities, financed by local, national and international institutions (University of Verona, Italian Ministry of Research, European Commission). His research focuses on the ways technological advancement affects the meaning of the human at ethical, anthropological, and social level.

Andrea Favaro (andrea.favaro@univr.it), Associate Professor of Philosophy of Law and Legal Informatics at the University of Verona (Department of Engineering for Innovation Medicine), initially focused on the theoretical study of the paradigm of autonomy (authoring two monographs) and the history of the relationship between the human person and legal institutions (authoring two additional monographs). In recent years, he has dedicated his research to examining issues of autonomy and responsibility of human subjects in relation to AI, both from the perspective of general theoretical foundations and applied aspects.

Silvia Francesca Storti (silviafrancesca.storti@univr.it) is Associate Professor in Bioengineering (BraiNavLab) at the University of Verona (Department of Engineering for Innovation Medicine). She received the M.Sc. degree in Electronic Engineering from the University of Padova and the Ph.D. degree in Neuroscience from the University of Verona. Her research lies in neuroengineering, BCIs, and biomedical signal processing. She develops intelligent systems to analyze dynamic multivariate data, forecast brain states, and estimate brain connectivity from wearable technologies. She has co-authored 68 publications on international peer-reviewed journals and 100+ conference contributions. She is Ass. Ed. for IEEE J-BHI and BioMedical Engineering OnLine.

Lorenza Brusini (lorenza.brusini@univr.it) is temporary assistant professor (RTDa) in Bioengineering at the University of Verona (Department of Engineering for Innovation Medicine). She received a MSc in Bioinformatics and Medical Biotechnology (2014) and a PhD in Computer Science (2018) from the University of Verona, Italy. She is Associate Editor for Pattern Recognition. She is co-author of 15 publications on international peer-reviewed journal and more than 20 contributions to international conferences. Her main research interests focus on brain microstructure and structural connectivity, explainable artificial intelligence for neuroimaging, and imaging-genetics.

Ahmed M. Salih (a.salih@leicester.ac.uk) received his BSc degree in Statistics and Informatics from Mosul University and the MSc in Bioinformatics from Leicester University, UK. In 2019 he was awarded a grant under the H2020 INVITE COFUND initiative to pursue a PhD in Computer Science at the University of Verona, Italy and graduated in 2022. He worked at William Harvey Research Institute, Queen Mary University of London and then at the Department of Population Health Sciences, University of Leicester. His research activity focuses on explainable artificial intelligence, imaging-genetics, brain and cardiac age estimation relying on machine and deep learning.

Ilaria Boscolo Galazzo (ilaria.boscologalazzo@univr.it) is tenure-track assistant professor (RTDb) in Bioengineering at the University of Verona (Department of Engineering for Innovation Medicine). She graduated cum laude in Biomedical Engineering at the University of Padova (2010) and received the Ph.D. degree in Neuroscience from the University of Verona (May 2014). She took a position as Research Associate at the Institute of Nuclear Medicine, UCL, London (2014-2016) and at the Dept. of Computer Science, University of Verona (2016-2023). Her main research activities include modelling of functional MRI data,

brain connectivity, imaging genetics, and multimodal data integration with AI-based methods.

Sergey Plis (splis@gsu.edu) received the Ph.D. in Computer Science from the University of New Mexico in 2007. He is Professor of Computer Science at Georgia State University and Director of the Machine Learning Core at the TReNDS Center. His research develops and applies machine-learning and data-science methods for large-scale datasets, emphasizing multimodal brain imaging. A central theme is modeling and explaining mechanisms in brain function. His long-term goal is to infer hidden structure noninvasively, including transient cognition-related networks, by fusing fast and slow modalities. Current efforts span deep learning for feature estimation and reliable causal modeling of network dynamics.

Gloria Menegaz (gloria.menegaz@univr.it) is Professor of Bioengineering at the University of Verona (Department of Engineering for Innovation Medicine), where she leads the BraiNavLab. She holds a PhD in Applied Sciences (EPFL 2000) and a MSc in Electronic Engineering (1993) and Information Technology (1995) (Politecnico di Milano). She was awarded the Rita Levi Montalcini grant by the Italian Ministry of University and Research (DM20/03/2003 n501). She is Senior Member of the IEEE, member of the IEEE-WIE, IEEE-SPS, IEEE-EMBS, SPS Liason representative and Chair of the IEEE TMI Steering Committee. Her research activity focuses on AI-based neuroimaging, brain connectivity and imaging genetics.