Ethics and philosophy articles to act as reference of cognition and consciousness research (D2.8 – SGA3)

Figure 1: Literature reference

Ethics and philosophy publications as a conceptual reference for scientific research

Project Number:	945539	Project Title:	HBP SGA3

Document Title:	Ethics and philosophy articles to act as reference of cognition and consciousness research (D2.8 – SGA3)
Document Filename:	D2.8 (D19) SGA3 M42 SUBMITTED 230911.docx
Deliverable Number:	SGA3 D2.8 (D19)
Deliverable Type:	Report
Dissemination Level:	PU = Public
Planned Delivery Date:	SGA3 M42 / 30 SEP 2023
Actual Delivery Date:	SGA3 M42 / 11 SEP 2023
Author(s):	Kathinka EVERS, UU (P83) and Michele FARISCO, UU (P83)
Compiled by:	Michele FARISCO, UU (P83)
Contributor(s):	Kathinka EVERS, UU (P83) and Michele FARISCO, UU (P83)
WP QC Review:	Eurídice ÁLVARO, IDIBAPS (P93); Angelica DA SILVA LANTYER, UvA (P98)
WP Leader / Deputy Leader Sign Off:	Maria V. SÁNCHEZ-VIVES, IDIBAPS (P93)
T7.4 QC Review:	N/A
Description in GA:	These articles propose criteria to analyse broadly relevant issues on the relationships between the complex networks and emergence of consciousness as well as between human and artificial intelligence/cognition. They will provide concrete empirical, theoretical, and behavioural criteria for ascribing consciousness to humans also in clinical contexts (e.g. patients who suffer from disorders of consciousness), animals, and machines. They will suggest strategies for dealing with key ethical and societal implications of the research results and broadly communicate concrete measures to connect new communities of potential users of EBRAINS from social science and the humanities.
Abstract:	D2.8 presents a collection of articles (peer-reviewed papers, book chapters, and blog posts) that resulted from the ethics and philosophy research conducted by Task 2.7. This scientific production aimed to act as a reference for consciousness and cognition research. D2.8 recapitulates the research conducted in about the last three years, covering different but reciprocally connected topics: criteria for a theoretical and ethical analysis of the relationship between complex natural and artificial networks and consciousness; criteria for ascribing consciousness in clinical, animal, and machine contexts; broad ethical and philosophical implications of research on consciousness and cognition, including a focus on illustrative case studies; concrete measures and strategies for bridging different communities in order to maximize the exploitation of EBRAINS services.
Keywords:	Consciousness, Cognition, Complexity, EBRAINS, Community Building, Ethics of Consciousness
Target Users/Readers:	Clinicians, computational neuroscience community, consortium members, experts in ethics, funders, public, neuroscientific community, clinical neuroscientists, EBRAINS users.

Table of Contents

1. Introduction 4
2. Criteria for analysing relationships between complex networks and consciousness 4
3. Ascribing consciousness: clinical, animal and machine contexts..... 6
4. Ethical and philosophical implications of cognition and consciousness research 9
5. Connecting communities: bridging social sciences, humanities, and cognition research 10
6. Conclusions and looking forward 11
7. References.....12

Table of Figures

Figure 1: Literature reference..... 1

1. Introduction

Deliverable D2.8 presents a collection of articles (peer-reviewed papers, book chapters, and blog posts) that resulted from the ethics and philosophy research conducted by Task 2.7. This scientific production aimed to act as a reference for consciousness and cognition research. D2.8 recapitulates the research conducted in about the last three years, covering different but reciprocally connected topics: criteria for a theoretical and ethical analysis of the relationship between complex natural and artificial networks and consciousness; criteria for ascribing consciousness in clinical, animal, and machine contexts; broad ethical and philosophical implications of research on consciousness and cognition, including a focus on illustrative case studies; concrete measures and strategies for bridging different communities in order to maximize the exploitation of EBRAINS services.

The scientific area D2.8 contributes to Ethics, but the reflection here presented relies on a multi- and inter-disciplinary methodology that makes it potentially relevant also for other scientific areas, particularly network complexity, cognitive functions, neuromorphic computing, robotics. Therefore, different communities should be interested in the work that D2.8 describes, including scholars in social science and humanities, researchers from within and outside the HBP, clinicians working with cognitive and consciousness disorders, patients' associations, regulatory bodies, policy makers, and general public.

The work here described makes an explicit reference to EBRAINS, particularly to the possibilities that it offers to foster an embedded and scientifically informed ethical and philosophical reflection, but it lacks a direct, structured connection with the actual EBRAINS Research Infrastructure. For this reason, D2.8 calls for the inclusion of high-level services dedicated to ethics and philosophy. The work described in D2.8 may help in identifying good practices and priority areas for future research, and specifically the challenges that the services provided by EBRAINS may help to mitigate. The final goal of D2.8 is maximizing the societal benefits deriving from EBRAINS.

D2.8 directly contributes to neuroethical reflection on neuroscientific research and emerging applications, providing both a method for identifying arising issues and illustrations of its application. Specifically, D2.8 provides further contribution to the development of fundamental neuroethics, which is the specific neuroethical methodology and approach refined within the HBP.

Each section reports the main results concerning the topics at issue, including a list of reference work resulting from T2.7 research.

2. Criteria for analysing relationships between complex networks and consciousness

Developing a common theoretical framework is crucial to advance in understanding consciousness and assessing related issues, particularly emerging ethical issues. The challenge is to find a common ground among the several experimental and theoretical approaches characterizing contemporary research on consciousness. A strong candidate that is achieving increasing consensus is the notion of complexity. The basic idea is that consciousness can be explained as a particular kind of neural information processing. The idea of associating consciousness with complexity was originally introduced by Giulio Tononi and Gerald Edelman in a 1998 paper titled Consciousness and Complexity (Tononi & Edelman, 1998).

Despite the increasing popularity of the notion, there are some theoretical challenges that need to be faced, particularly concerning the supposed explanatory role of complexity. These challenges are philosophically relevant. They might also affect the scientific reliability of complexity and the legitimacy of invoking this concept in the interpretation of emerging data and in the elaboration of scientific explanations. In addition, the theoretical challenges have a direct ethical impact, because an unreliable conceptual assumption may lead to misplaced ethical choices. For example, we might wrongly assume that a patient with low complexity is not conscious, or vice-versa, eventually making medical decisions that are inappropriate to the actual clinical condition.

The claimed explanatory power of complexity is challenged in two main ways: semantically and logically.

Semantic challenges arise from the fact that complexity is such a general and open-ended concept. In fact, complexity is so general to seem meaning different things in different contexts and in different disciplines (Adami, 2002). One is simply that we qualify complex "an organization or a system that we don't understand and master" (Thom, 1990). For this reason, the concept of complexity is applicable to life, in general, rather than to consciousness only.

This open-ended generality and lack of shared definition can be a barrier to a common scientific use of the term, which may impact its explanatory value in relation to consciousness. In the landmark paper by Tononi and Edelman, complexity is defined as the combination of integration (conscious experience is unified) and differentiation (we can experience different states).

To properly evaluate the explanatory value of complexity for consciousness, it is crucial to clarify the meaning of consciousness it refers to. In fact, consciousness may assume different specific meanings. For instance, in the clinical context, there is a distinction between the level (or state) of consciousness (i.e., wakefulness/sleep) and its content (e.g., images or words) (Laureys, 2005). Another popular distinction, even though not unanimously accepted (Naccache, 2018), has been introduced by Ned Block, who differentiated phenomenal consciousness (i.e., subjective experience or "what it is like to be") from access consciousness (i.e., availability of information for use in reasoning and rationally guiding speech and action) (Block, 1995). Importantly, the notion of complexity is usually referred to the "level/state" of consciousness rather than to its "contents." This means that complexity-related measures can provide relevant information about the "level/state" of consciousness, while they are silent about the corresponding object as well as its phenomenology. This is an ethically salient point, since the dimensions of consciousness that appear most relevant to making ethical decisions are those related to subjective positive and/or negative experiences. For instance, while it is generally considered as ethically neutral how we treat a machine, it is considered ethically wrong to cause negative experiences to other humans or to animals.

Logical challenges arise if one wants to refer to complexity not only as a requirement for consciousness but as its explanation. The second option needs a sound justification. This justification usually takes one of two alternative forms. The justification is either bottom-up (from data to theory) or top-down (from phenomenology to physical structure). Both raise specific issues.

Bottom-up: Starting from empirical data indicating that brain structures or functions correlate to conscious states, relevant theoretical conclusions are inferred. More specifically, since the brains of subjects that are manifestly conscious exhibit complex patterns (integrated and differentiated patterns), we are supposed to be justified to infer that complexity indexes consciousness. This conclusion is a sound inference to the best explanation, but the fact that a conscious state correlates with a complex brain pattern in healthy subjects does not justify the inference that the same level of complexity is needed for all possible conditions (for example, disorders of consciousness) to be conscious, and even if empirically true it does not logically imply that complexity is a necessary and/or sufficient condition for consciousness.

Top-down: Starting from certain characteristics of personal experience, we are supposed to be justified to infer corresponding characteristics of the underlying physical brain structure. More specifically, if some conscious experience is complex in the technical sense of being both integrated and differentiated, we are supposed to be justified to infer that the correlated brain structures must be complex in the same technical sense. This conclusion does not seem logically justified unless we start from the assumption that consciousness and corresponding physical brain structures must be similarly structured. Otherwise, it is logically possible that conscious experience is complex while the corresponding brain structure is not, and vice versa. In other words, it does not appear justified to infer that since our conscious experience is integrated and differentiated, the corresponding brain functional state must be integrated and differentiated. This is a possibility, but not a necessity.

The abovementioned theoretical challenges do not deny the practical utility of complexity as a relevant measure in specific clinical contexts, for example, to quantify residual consciousness in patients with disorders of consciousness. In fact, relevant complexity-related clinical tools, like the Perturbational Complexity Index, are increasingly used as reliable tools for measuring the conscious

state in these people (Sarasso et al., 2021). What is at stake is the explanatory status of the notion. Even if we question complexity as a key factor in explaining consciousness, we can still acknowledge that complexity is practically relevant and useful, for example, in the clinic. In other words, while complexity as an explanatory category raises serious conceptual challenges that remain to be faced, complexity represents at the practical level one of the most promising tools that we have to date for improving the detection of consciousness and for implementing effective therapeutic strategies.

3. Ascribing consciousness: clinical, animal and machine contexts

A reliable detection of conscious activity is among the main neuroscientific and technological enterprises today.

An immediate and intuitive approach consists in attributing consciousness based on the externally observable behavior. Yet cases of disconnection between conscious state and behavior (e.g., in patients with cognitive-motor dissociation (Schiff, 2015) shows that we cannot rely on behavior alone in order to attribute consciousness.

As mentioned above, the case of patients with devastating neurological impairments, like disorders of consciousness (unresponsive wakefulness syndrome, minimally conscious state, and cognitive-motor dissociation) is highly illustrative. In fact, some of these patients might retain residual conscious abilities although they are unable to show them behaviourally. Similarly, subjects with locked-in syndrome have a fully conscious mind even if they do not exhibit any behaviours other than blinking.

We can conclude that absence of behavioural evidence for consciousness is not evidence for the absence of consciousness. Therefore, it is urgent to look for reliable indicators of consciousness independent from behavior.

The identification of indicators of consciousness is necessarily a conceptual and an empirical task: we need a clear idea of what to look for to define appropriate empirical strategies. Accordingly, we have introduced a list of six indicators of consciousness (C. M. A. Pennartz, M. Farisco, & K. Evers, 2019) that do not rely only on behavior, but can be assessed also through technological and clinical approaches:

1) **Goal directed behaviour (GDB) and model-based learning.** In GDB I am driven by expected consequences of my action, and I know that my action is causal for obtaining a desirable outcome. Model-based learning depends on my ability to have an explicit model of myself and the world surrounding me.
2) **Brain anatomy and physiology.** Since the consciousness of mammals depends on the integrity of cerebral systems (i.e., thalamocortical systems), it is reasonable to think that preserved structures in brain-injured patients and similar structures in non-human animals indicate the presence of consciousness.
3) **Psychometrics and meta-cognitive judgement.** If I can detect and discriminate stimuli and can make some meta-cognitive judgements about perceived stimuli, I am probably conscious.
4) **Episodic memory.** If I can remember events (“what”) I experienced at a particular place (“where”) and time (“when”), I am probably conscious.
5) **Acting out one’s subjective, situational survey: illusion and multistable perception.** If I am susceptible to illusions and perceptual ambiguity, I am probably conscious.
6) **Acting out one’s subjective, situational survey: visuospatial behaviour.** If I can perceive objects as stably positioned, even when I move in my environment and scan it with my eyes, I am probably conscious.

This list is conceived to be provisional and heuristic but also operational: it is not a definitive answer to the problem, but it is sufficiently concrete to help identify consciousness in others.

To recapitulate, indicators of consciousness are conceived as particular capacities that can be deduced from the behavior or cognitive performance of a subject or from relevant neuronal/physiological correlates, and that serve as a basis for a reasonable inference about the level of consciousness of the subject in question. This implies the relevance of the indicators to patients with DoCs, who are often unable to behave or to communicate overtly, as well as to non-human animals and even to machines.

To illustrate, we investigated how the different indicators of consciousness might be applied to patients with DoCs with the final goal of contributing to improve the assessment of their residual conscious activity (M. Farisco, Pennartz, Annen, Cecconi, & Evers, 2022b). In fact, a still astonishing rate of misdiagnosis affects this clinical population. It is estimated that up to 40 % of patients with DoCs are wrongly diagnosed as being in Vegetative State/Unresponsive Wakefulness Syndrome, while they are in a Minimally Conscious State. The difference of these diagnoses is not minimal, since they have importantly different prognostic implications, which raises a huge ethical problem.

We also argue for the need to recognize and explore the specific quality of the consciousness possibly retained by patients with DoCs. Because of the devastating damages of their brain, it is likely that their residual consciousness is very different from that of healthy subjects, usually assumed as a reference standard in diagnostic classification. To illustrate, while consciousness in healthy subjects is characterized by several distinct sensory modalities (for example, seeing, hearing, and smelling), it is possible that in patients with DoCs, conscious contents (if any) are very limited in sensory modalities. These limitations may be evaluated based on the extent of the brain damage and on the patients' residual behaviours (for instance, sniffing for smelling). Also, consciousness in healthy subjects is characterized by both dynamics and stability: it includes both dynamic changes and short-term stabilization of contents. Again, in the case of patients with DoCs, it is likely that their residual consciousness is very unstable and flickering, without any capacity for stabilization. If we approach patients with DoCs without acknowledging that consciousness is like a spectrum that accommodates different possible shapes and grades, we exclude a priori the possibility of recognizing the peculiarity of consciousness possibly retained by these patients.

The indicators of consciousness we introduced offer a potential help to identify the specific conscious abilities of these patients. In the paper mentioned above (M. Farisco et al., 2022b) we argue for the rationale behind the clinical use of these indicators, and for their relevance to patients with DoCs, while we also acknowledge that they open up new lines of research with concrete application to patients with DoCs. This more applied work is in progress and almost ready for submission as a live paper.

Another potential field of application for the indicators of consciousness is AI research. This has generated a linguistic and conceptual process of re-conceptualization of traditional human features, stretching their meaning or even reinventing their semantics to attribute these traits also to machines. The attribution of concepts like learning, experience, training, prediction, to name just a few, to AI is illustrative in this sense. Even if they have a specific technical meaning among AI specialists, lay people tend to interpret them within an anthropomorphic view of AI.

One human feature is considered the Holy Grail when AI is interpreted within an anthropomorphic framework: consciousness. Yet can AI really be conscious? To answer this question, it is necessary to analyse a few preliminary issues.

First, we should clarify what we mean by consciousness. As mentioned above, in philosophy and in cognitive science there is a useful distinction, originally introduced by Ned Block, between access consciousness and phenomenal consciousness. The first refers to the interaction between different mental states, particularly the availability of one state's content for use in reasoning and rationally guiding speech and action. In other words, access consciousness refers to the possibility of using what I am conscious of. Phenomenal consciousness refers to the subjective feeling of a particular experience, "what it is like to be" in a particular state, according to the definition by Thomas Nagel. So, what is the sense of the word "consciousness" presupposed in the question if AI can be conscious?

A paper by Dehaene et al. (Dehaene, Lau, & Kouider, 2017) is illustrative of how the sense in which we choose to talk about consciousness makes a difference in the assessment of the possibility of conscious AI. The authors frame the question of AI consciousness within the Global Neuronal Workspace Theory, one of the leading contemporary theories of consciousness. As the authors write,

according to this theory, conscious access corresponds to the selection, amplification, and global broadcasting of information, selected for its salience or relevance to current goals, to many distant areas. More specifically, Dehaene and colleagues explore the question of conscious AI along two lines within an overall computational framework:

1) Global availability of information (the ability to select, access, and report information)

2) Metacognition (the capacity for self-monitoring and confidence estimation).

Their conclusion is that AI might implement the first meaning of consciousness, while it currently lacks the necessary architecture for the second one.

As mentioned, the premise of their analysis is a computational view of consciousness: they choose to reduce consciousness to specific types of information-processing computations. We can legitimately ask whether such a choice covers the richness of consciousness, particularly whether a computational view can account for the experiential dimension of consciousness.

This shows how the main obstacle in assessing the question whether AI can be conscious is a lack of agreement about a theory of consciousness in the first place. For this reason, rather than asking whether AI can be conscious, maybe it is better to ask what might indicate that AI is conscious according to some more or less objective indicators.

Another important preliminary issue to consider, if we want to seriously address the possibility of conscious AI, is whether we can use the same term, "consciousness," to refer to a different kind of entity: a machine instead of a living being. Should we expand our definition to include machines, or should we rather create a new term to denote it? The term "consciousness" is arguably too charged, from several different perspectives, including ethical, social, and legal perspectives, to be extended to machines. Using the term to qualify AI risks extending it so far that it eventually becomes meaningless.

If we create AI that manifests abilities that are similar to those that we see as expressions of consciousness in humans, we probably need a new language to denote and think about it. Otherwise, important preliminary philosophical questions risk being dismissed or lost sight of behind a conceptual veil of possibly superficial linguistic analogies.

A last point to consider is that the word "consciousness" covers a number of abilities and has several dimensions of components that may be expressed separately. The notion of consciousness dimensions has been widely analysed in a 2016 paper by Tim Bayne, Jakob Hohwy, and Adrian Owen(Bayne, Hohwy, & Owen, 2016). Focusing on global states of consciousness (i.e., states of consciousness characterizing the overall conscious condition of a subject) as distinguished from local states of consciousness (i.e., specific conscious contents or experiences), they argue that consciousness is manifested in multiple ways in different global states, and that the notion of levels should be replaced by that of dimensions of consciousness to properly describe the multifaceted nature of consciousness. The central thesis is that global states of consciousness are not gradable along one dimension, but rather distinguished along different dimensions. The authors introduce two main families of consciousness dimensions: content-related and functional dimensions. The first family includes, for instance, gating of conscious content (e.g., low-level features vs high-level features of an object). The second family includes, for instance, cognitive and behavioural control (i.e., the availability of conscious contents for control of thought and action). More specifically, Walter has recently proposed the following content-related dimensions: sensory richness, high-order object representation, semantic comprehension, and the following functional dimensions: executive functioning, memory consolidation, intentional agency, reasoning, attention control, vigilance, meta-awareness (Walter, 2021).

Relevant reflections about the dimensions of consciousness come from Birch et al., who focusing on animal consciousness specifically introduce the following dimensions (Birch, Schnell, & Clayton, 2020):

- Perceptual-Richness: any measure is specific to a sense modality, so there is no overall level of perceptual richness. Also, within a particular sense modality, perceptual richness can be resolved into different components.

- Evaluative-Richness: affectively based positive or negative valence which grounds decision-making. Also, evaluative richness can be resolved into different components.
- Integration at a time (unity): conscious experience is (usually) highly unified.
- Integration across time (temporality): conscious experience takes the form of a continuous stream.
- Self-consciousness (Selfhood): awareness of oneself as distinct from the world outside.

Also Chalmers has recently referred to different consciousness dimensions in a reflection about the relationship between Large Language Models and consciousness (Chalmers, 2023). Referring to conscious experience, he identifies the following dimensions:

- Sensory experience: e.g., seeing red.
- Affective experience: e.g., feeling pain.
- Cognitive experience: e.g., thinking hard.
- Agentive experience: e.g., deciding to act.
- Self-consciousness: awareness of oneself

This multidimensional view of consciousness is particularly relevant for a balanced reflection about the possibility of conscious AI, which may arguably manifest one or selected dimensions without necessarily cover all the richness of human and non-human animal consciousness.

4. Ethical and philosophical implications of cognition and consciousness research

Notwithstanding significant progresses (e.g., the identification of brain mechanisms/structures critical for conscious processing and the development of technologies for a better detection of residual conscious abilities, including more reliable diagnosis and prognosis of its disorders and disfunctions), research on consciousness still appears affected by several difficulties and shortcomings. This resulting stalemate is both a matter of theory (e.g., ill-defined explanandum) and of interpretation of empirical data: one of the main challenges is the lack of a shared theoretical framework, so that the interpretation of data is not uniform, and it eventually results in different conceptual models. How to handle and possibly reconcile these differences is still an open question, despite some recent attempts to identify shared theoretical features (Francken et al., 2022; Michel et al., 2019; Northoff & Lamme, 2020; Seth & Bayne, 2022; Wiese, 2020) and even to implement an "adversarial collaboration" between concurrent theories (Yaron, Melloni, Pitts, & Mudrik, 2022). This theoretical diversity has also ethical implications, namely for the assessment of several emerging practical issues. Consider, for instance, the very challenging questions concerning residual conscious activity in patients with Disorders of Consciousness (M. Farisco, Pennartz, Annen, Cecconi, & Evers, 2022a), or the possibility of creating artificial consciousness (C. Pennartz, M. Farisco, & K. Evers, 2019), or even to create potentially conscious organoids in the lab (Reardon, 2020). We are unlikely to reach a consensus on how to approach and manage these ethical issues without a shared theoretical framework on consciousness to build on, or at least without a minimal consensus on what are the necessary conditions for an agent/entity (either biological or artificial) to be considered capable of conscious processing. In fact, we cannot wait for elaborating a strong unifying theoretical model of consciousness before thinking about effective strategies for dealing with the abovementioned emerging issues. Ethical needs exceed theoretical desires, and we urgently must reach a consensus on a shared ethical approach.

Among the ethical and philosophical implications of consciousness and cognition research, the following can be considered as particularly urgent (Michele Farisco, 2023):

1) Ethics of Disorders of Consciousness. Ethical reflection can play an important role in optimizing the management of patients with Disorders of Consciousness, for instance revealing both good and bad practices. In fact, several issues arise, both from research and from clinics, and they

deserve specific attention. It is urgent to move beyond a defensive type of ethics focused on safeguarding from potential risks towards an ethics that is also pro-active, and to think about a methodology and a model for translating ethical thinking into the actual clinical treatment of affected patients. Accordingly, I have proposed the distinction between fundamental and practical ethical issues as a methodology for the ethics in the management of patients with Disorders of Consciousness (M. Farisco, In Press), introducing the Distributed Responsibility Model for the clinical operationalization of ethics (M. Farisco & Salles, 2022).

2) Consciousness and moral status. In moral philosophy and in ethics, consciousness has traditionally played a crucial role in attributing moral status to both human and non-human agents, to the extent to be considered paradigmatic for ascribing both moral relevance and moral saliency (Levy, 2014). Yet the necessity of consciousness for attributing moral status, particularly to non-humans, is not uncontroversial (Shepherd, 2022).

3) Ethics of technologically extended consciousness. Technology is increasingly impacting our brains and our minds: the humanistic paradigm of human identity as autonomous and self-centred is challenged by the increasing hybridization between human and technology that according to some scholars results in a post-humanist subjectivity (Braidotti, 2013). Besides the theoretical implications of this trend, it appears urgent to investigate the related ethical issues, including the impact on the notion of privacy and confidentiality (Palermos, 2022).

4) Ethics of potential artificial consciousness. We are witnessing astonishing advances in Artificial Intelligence (AI) as well as in associated robotic applications. An increasing number of typically human features are emulated and simulated by AI, including scenes and objects recognition, game playing, and action planning, among others. Discussion is open regarding the possibility of creating Artificial Consciousness (Carter et al., 2018; Dehaene et al., 2017). Even if we do not have definitive answers from a technical and a theoretical point of view, we already must face the ethical dilemma concerning possible artificial forms of consciousness and their potential moral status. It is crucial to explore both foundational and practical, including regulatory aspects of this possibility (Hildt, 2022).

5) Ethics of potential conscious cerebral organoids. Human cerebral organoids are three-dimensional in vitro cell cultures mimicking the developmental process and organization of the human brain. They are particularly useful in research about diseases of the nervous systems and their pharmacological treatments. Human cerebral organoids significantly manifest electrical activity and connectivity similar to human subjects (e.g., preterm infants). This originated a lively debate about the possibility that human cerebral organoids will develop a form of consciousness. This scenario raises important ethical issues deserving a dedicated reflection, about its conceptual plausibility', its technical feasibility, and its possible practical implications (Reardon, 2020).

6) Ethics of altered states of consciousness: the case of psychedelics. In recent years there has been a renaissance of psychedelic research which has shed light on the neurophysiology of altered states of consciousness induced by classical psychedelics, such as psilocybin and LSD. Among other things, this kind of research has led to a new theory of consciousness (Carhart-Harris, 2018; Carhart-Harris et al., 2014), that promises to have clinical implications in a number of contexts, and to experimental applications in the treatment of Disorders of Consciousness (A. Peterson, Tagliazucchi, & Weijer, 2019) and Alzheimer's Disease, among others (Andrew Peterson, Largent, Lynch, Karlawish, & Sisti, 2022).

5. Connecting communities: bridging social sciences, humanities, and cognition research

EBRAINS has been conceived as a digital research infrastructure aimed, among other things, to make available the most up-to-date data and tools resulting from the Human Brain Project (HBP). The final goal of these freely available resources is to speed up scientific research to maximize the resulting societal benefits, including, among other things, the clinical translation of emerging results and the maximization of the social exploitation of the most up-to-date scientific knowledge and

technological applications. For this reason, it is crucially important to involve people from fields different than neuroscience and Information and Communication Technology, like communities of social science and humanities scholars as well as patients' associations and other stakeholders from the public. Since its very beginning, the HBP has been conceived as a multi- and inter-disciplinary research project, involving researchers from different disciplines, including both sciences and humanities (Salles et al., 2019).

In this way it has been possible to incorporate public needs into the planning and implementation of scientific research activities and to timely identify social and ethical issues emerging from research and related technologies. This connection with the public has been possible through tailored communication and engagement activities as well as through research activities involving scholars in social sciences, philosophy, and ethics.

In his attempt to engage social science and humanities communities, T2.7 has considered two needs, which have specific characteristics deserving tailored strategies even if they are not mutually exclusive: the need for a multi- and inter-disciplinary academic dialogue and collaboration; the need for a strong and scientifically informed communication and engagement with the public. On the basis of these two main needs T2.7 has planned and organized a number of activities, including online Capacity Development Courses, Public Engagement activities, multidisciplinary conferences, the coordination of a multidisciplinary scientific paper, the coordination of a journal special issue devoted to the ethical questions emerging from research on consciousness and related technology, and the collaboration with the International Brain Injury Association and with a COST Action devoted to the connection between kidney diseases and brain function.

In this way T2.7 has offered a spectrum of activities that have been tailored to address the needs and interests of different stakeholders, including people from social sciences and humanities. These strategies and activities have been aimed to fostering multi- and inter-disciplinary collaborations between social science, humanities, and other scientific domains within the EBRAINS framework. Also, they aim to facilitate the translation of HBP outputs and EBRAINS platform in easy-to-access resources for the public, particularly for lay people and patients' associations.

The collaborative scientific publications resulting from T2.7 research, particularly the live paper on DoCs (M. Farisco et al., Under review), the collaboration with the International Brain Injury Association (IBIA) (M. Farisco et al., In Press), and the Capacity Development Courses (CDC) are three illustrations of complementary and successful strategies to enlarge the number of people from different disciplines involved in the discussion and the potential exploitation of HBP and EBRAINS resources. People from humanities and social sciences may use EBRAINS to increase their knowledge of the most advanced research tools on brain and related diseases, as well as of how to translate them in socially beneficial activities, including clinical applications. Also, EBRAINS offers a model of embedded ethical reflection on neuroscience and related technologies. To increase the interest by people from humanities and social sciences to use it, EBRAINS may include more relevant high-level services (e.g., in the forms of guidelines, decision making tools for ethically sensible scenarios, question and answer tools specific for the significance of EBRAINS for humanities and social sciences, etc.).

6. Conclusions and looking forward

Task T2.7 has produced several scientific papers, blog posts, and books chapters covering different philosophical and ethical issues arising from research on consciousness and cognition. What emerged as a productive strategic choice is the inter-disciplinary collaboration, notably between philosophy, ethics, medicine, biology, and neuroscience connecting practical/clinical and theoretical perspectives. This kind of collaboration has been key to advance in the identification and assessment of emerging issues in the field. HBP has been a forerunner in this quest and on inspiration for other large research initiatives to follow suit. Therefore, the plan is to keep working in this direction, further refining the methodology and strengthening the collaboration with neuroscience, particularly with clinical and computational neuroscience, in both theoretical and ethical reflections. EBRAINS may serve as a good avenue in this respect, functioning as a platform for making possible the interaction between different disciplines and stakeholders.

7. References

Adami, C. (2002). What is complexity? *Bioessays, 24*(12), 1085-1094. doi:10.1002/bies.10192

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are There Levels of Consciousness? *Trends Cogn Sci, 20*(6), 405-413. doi:10.1016/j.tics.2016.03.009

Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of Animal Consciousness. *Trends Cogn Sci, 24*(10), 789-801. doi:10.1016/j.tics.2020.07.007

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227-287.

Braidotti, R. (2013). *The posthuman*. Cambridge: Polity.

Carhart-Harris, R. L. (2018). The entropic brain - revisited. *Neuropharmacology, 142*, 167-178. doi:10.1016/j.neuropharm.2018.03.010

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., . . . Nutt, D. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Front Hum Neurosci, 8*, 20. doi:10.3389/fnhum.2014.00020

Carter, O., Hohwy, J., van Boxtel, J., Lamme, V., Block, N., Koch, C., & Tsuchiya, N. (2018). Conscious machines: Defining questions. *Science, 359*(6374), 400. doi:10.1126/science.aar4163

Chalmers, D. (2023). Could a Large Language Model be Conscious? Retrieved from arXiv:2303.07103 website:

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science, 358*(6362), 486-492. doi:10.1126/science.aan8871

Farisco, M. (2023). The Ethical Spectrum of Consciousness. *AJOB Neuroscience, 14*(2), 55-57. doi:10.1080/21507740.2023.2188312 (PLUS ID: P3935)

Farisco, M. (In Press). The Ethics in the Management of Patients with Disorders of Consciousness. In S. Laureys & C. Schnakers (Eds.), *Coma and Disorders of Consciousness*: Springer Nature.

Farisco, M., Evers, K., Annen, J., Blandin, V., Camassa, A., Cecconi, B., . . . Zamora-Lopez, G. (Under review). Advancing the science of consciousness: from ethics to clinical care. *Brain*.

Farisco, M., Formisano, R., Gosseries, O., Kato, Y., Koboyashi, S., Laureys, S., . . . Estraneo, A. (In Press). International survey on the implementation of the European and American guidelines on disorders of consciousness. *J Neurol*.

Farisco, M., Pennartz, C., Annen, J., Cecconi, B., & Evers, K. (2022a). Indicators and criteria of consciousness: ethical implications for the care of behaviourally unresponsive patients. *BMC Med Ethics, 23*(1), 30. doi:10.1186/s12910-022-00770-3

Farisco, M., Evers, K. & Salles, A. On the Contribution of Neuroethics to the Ethics and Regulation of Artificial Intelligence. Neuroethics 15, 4 (2022). https://doi.org/10.1007/s12152-022-09484-0 (PLUS ID: P3225)

Farisco, M., Changeux, J.P. (2023) About the compatibility between the perturbational complexity index and the global neuronal workspace theory of consciousness, Neuroscience of Consciousness, Volume 2023, Issue 1, 2023, niad016, https://doi.org/10.1093/nc/niad016 (PLUS ID: P4038)

Farisco, M., & Salles, A. (2022). American and European Guidelines on Disorders of Consciousness: Ethical Challenges of Implementation. *J Head Trauma Rehabil, 37*(4), 258-262. doi:10.1097/HTR.0000000000000776 (PLUS ID: P3311)

Farisco, M., Evers, K. & Salles, A. Towards Establishing Criteria for the Ethical Analysis of Artificial Intelligence. Sci Eng Ethics 26, 2413–2425 (2020). https://doi.org/10.1007/s11948-020-00238-w (PLUS ID: P2577)

Farisco, M. et al. (2023). The need for a multi-disciplinary reflection about frailty and cognitive impairment in chronic kidney disease, Nephrology Dialysis Transplantation, Volume 38, Issue 5, May 2023, Pages 1064–1066, https://doi.org/10.1093/ndt/gfac334 (PLUS ID: P3789)

Francken, J. C., Beerendonk, L., Molenaar, D., Fahrenfort, J. J., Kiverstein, J. D., Seth, A. K., & van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neurosci Conscious, 2022*(1), niac011. doi:10.1093/nc/niac011

Hildt, E. (2022). The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns. *AJOB Neuroscience*, 1-14. doi:10.1080/21507740.2022.2148773

Laureys, S. (2005). The neural correlate of (un)awareness: lessons from the vegetative state. *Trends Cogn Sci, 9*(12), 556-559. doi:10.1016/j.tics.2005.10.010

Levy, N. (2014). The Value of Consciousness. *J Conscious Stud, 21*(1-2), 127-138.

Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., . . . Yoshida, M. (2019). Opportunities and challenges for a maturing science of consciousness. *Nat Hum Behav, 3*(2), 104-107. doi:10.1038/s41562-019-0531-8

Naccache, L. (2018). Minimally conscious state or cortically mediated state? *Brain, 141*(4), 949-960. doi:10.1093/brain/awx324

Northoff, G., & Lamme, V. (2020). Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? *Neurosci Biobehav Rev, 118*, 568-587. doi:10.1016/j.neubiorev.2020.07.019

Palermos, S. O. (2022). Data, Metadata, Mental Data? Privacy and the Extended Mind. *AJOB Neuroscience*, 1-13. doi:10.1080/21507740.2022.2148772

Pennartz, C., Farisco, M., & Evers, K. (2019). Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach. *Front Syst Neurosci, 13*, 25. doi:10.3389/fnsys.2019.00025 (PLUS ID: P2013)

Peterson, A., Largent, E. A., Lynch, H. F., Karlawish, J., & Sisti, D. (2022). Journeying to Ixtlan: Ethics of Psychedelic Medicine and Research for Alzheimer's Disease and Related Dementias. *AJOB Neuroscience*, 1-17. doi:10.1080/21507740.2022.2148771

Peterson, A., Tagliazucchi, E., & Weijer, C. (2019). The ethics of psychedelic research in disorders of consciousness. *Neurosci Conscious, 2019*(1), niz013. doi:10.1093/nc/niz013

Reardon, S. (2020). Can lab-grown brains become conscious? *Nature, 586*(7831), 658-661. doi:10.1038/d41586-020-02986-y

Salles, A., Evers, K. & Farisco, M. (2020) Anthropomorphism in AI, AJOB Neuroscience, 11:2, 88-95, DOI: 10.1080/21507740.2020.1740350. https://arxiv.org/abs/2305.10938 (PLUS ID: P2506)

Salles, A., Bjaalie, J. G., Evers, K., Farisco, M., Fothergill, B. T., Guerrero, M., . . . Amunts, K. (2019). The Human Brain Project: Responsible Brain Research for the Benefit of Society. *Neuron, 101*(3), 380-384. doi:10.1016/j.neuron.2019.01.005 (PLUS ID: P1665)

Salles, A. & Farisco, M. (2020). Of Ethical Frameworks and Neuroethics in Big Neuroscience Projects: A View from the HBP, AJOB Neuroscience, 11:3, 167-175, DOI: 10.1080/21507740.2020.1778116 (PLUS ID: P2578)

Sarasso, Casali, A. G., Casarotto, S., Rosanova, M., Sinigaglia, C., & Massimini, M. (2021). Consciousness and complexity: a consilience of evidence. *Neuroscience of Consciousness*. doi:10.1093/nc/niab023

Schiff, N. D. (2015). Cognitive Motor Dissociation Following Severe Brain Injuries. *JAMA Neurol, 72*(12), 1413-1415. doi:10.1001/jamaneurol.2015.2899

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nat Rev Neurosci, 23*(7), 439-452. doi:10.1038/s41583-022-00587-4

Shepherd, J. (2022). Non-Human Moral Status: Problems with Phenomenal Consciousness. *AJOB Neuroscience*, 1-10. doi:10.1080/21507740.2022.2148770

Thom, R. (1990). *Semio physics : a sketch*. Redwood City, Calif.: Addison-Wesley Pub. Co., Advanced Book Program.

Tononi, & Edelman, G. M. (1998). Consciousness and complexity. *Science, 282*(5395), 1846-1851. doi:10.1126/science.282.5395.1846

Walter, J. (2021). Consciousness as a multidimensional phenomenon: implications for the assessment of disorders of consciousness. *Neurosci Conscious, 2021*(2), niab047. doi:10.1093/nc/niab047

Wiese, W. (2020). The science of consciousness does not need another theory, it needs a minimal unifying model. *Neurosci Conscious, 2020*(1), niaa013. doi:10.1093/nc/niaa013

Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nat Hum Behav, 6*(4), 593-604. doi:10.1038/s41562-021-01284-5

Farisco, M. (2021) Consciousness and complexity: theoretical challenges for a practically useful idea. THE ETHICS BLOG. A blog from the Centre for Research Ethics & Bioethics (CRB). (Plus ID: E3432) https://plus.humanbrainproject.eu/disseminations/3432/

Farisco, M. (2021). Can AI be conscious? Let's think about the question. THE ETHICS BLOG. A blog from the Centre for Research Ethics & Bioethics (CRB). Available at: https://ethicsblog.crb.uu.se/2021/05/04/can-ai-be-conscious-let-us-think-about-the-question/ (Plus ID: 3411)

Farisco, M. (2022). How can we detect consciousness in brain damaged patients? THE ETHICS BLOG. A blog from the Centre for Research Ethics & Bioethics (CRB). Available at:https://ethicsblog.crb.uu.se/2022/04/19/how-can-we-detect-consciousness-in-brain-damaged-patients/?utm_source=rss&utm_medium=rss&utm_campaign=how-can-we-detect-consciousness-in-brain-damaged-patients (Plus ID: E5033)

Farisco, M. (2020). Are you conscious? Looking for reliable indicators. THE ETHICS BLOG. A blog from the Centre for Research Ethics & Bioethics (CRB). Available at: https://ethicsblog.crb.uu.se/2020/12/01/are-you-conscious-looking-for-reliable-indicators/(Plus ID:E3181)

Farisco, M. (2022) An ethical strategy for improving the healthcare of brain-damaged patients. THE ETHICS BLOG. A blog from the Centre for Research Ethics & Bioethics (CRB). Available at: https://ethicsblog.crb.uu.se/2022/05/31/an-ethical-strategy-for-improving-the-healthcare-of-brain-damaged-patients/ (Plus ID: E5154).