From “Scanning” to “Shaping”: An Ethical Risk and Governance Mechanism Study of AI Empowered Brain Computer Interfaces Based on the Reference of the Large-scale Game SOMA

Weixiang Gan, Mengfei Xiao*, Sikun Chen, Tara Ahmed Mohammed, Xiaolin Song
Graduate School of Business, SEGi University, Petaling Jaya, Selangor 47810, Malaysia
*Corresponding email: seanphydiyas@gmail.com

Abstract

With the accelerating convergence of artificial intelligence and brain computer interface technologies, the focal point of ethical risk is shifting from system security and technical accuracy toward deeper questions of subjectivity, namely who counts as a human being and who ought to be recognized as a rights bearing subject. Drawing on the science fiction game SOMA on Steam as a high-density vehicle of ethical imagination, this study conducts a systematic analysis of the governance dilemmas that may arise from AI empowered brain computer interfaces. The findings are threefold. First, when a mind is replicated with high fidelity as a digital copy, institutional grey zones emerge in identity continuity and in the definition of rights bearing subjecthood, which can readily trigger identity appropriation and the drifting of responsibility attribution. Second, once readout-oriented interfaces are combined with AI inference capabilities, neural data may be continuously reinterpreted as expandable cues of the mind, exposing mental privacy over the long term. In such contexts, one time consent mechanisms are unable to cover downstream and future uses, thereby generating risks of latent discrimination and distorted opportunity allocation. Third, if closed loop writes in systems operate persistently under an intervention rationale framed as being for the individual’s own good, they directly implicate autonomy and psychological integrity. Where exit options and enforceable accountability designs are absent, the legitimacy of governance and societal trust will be severely undermined. Building on these analyses, the article proposes three actionable governance mechanisms. First, it calls for clear regimes of identity designation, authorization, and revocation for mind copies. Second, it recommends a high sensitivity tiered governance approach for neural inference, clarifying inference boundaries and introducing dynamic consent mechanisms. Third, it argues for institutionalizing meaningful human final control over closed loop interventions, establishing red line scenario constraints, and constructing an auditable chain of responsibility, thereby providing an institutional pathway for responsible innovation in AI empowered brain computer interfaces.

Keywords

SOMA, Brain-computer interface, Artificial Intelligence, Mental privacy, Identity continuity, Closed-loop intervention

Introduction

With the deepening development of the digital content industry, globalized gaming platforms represented by Steam have increasingly become a key cultural infrastructure through which the public encounters and imagines frontier technologies. Steam’s distribution mechanisms and its ecology of community discussion enable science fiction works to circulate persistently on a global scale and, through the joint effects of playability and narrativity, translate what would otherwise be abstract and distant technological issues into a form of public experience that can be personally lived through [1]. For this reason, science fiction narrative games are no longer merely entertainment media. Rather, through their interactive structures, they constitute an ethical field marked by situational pressure and value conflict. As players advance the storyline and complete tasks, they do not passively receive viewpoints. Instead, through repeated moments in which choices must be made, they form intuitive judgements and moral positions regarding technological legitimacy and the boundaries of acceptable risk [2].

In this sense, SOMA is a particularly valuable case for scholarly inquiry. Developed by Frictional Games, the game situates its narrative primarily within an underwater research facility, tightly weaving psychological horror with hard science fiction premises. Within an oppressive atmosphere, players gradually confront key plot elements such as consciousness scanning, personality replication, and substrate transfer, and are repeatedly driven toward the question of who I am. If consciousness can be copied and instantiated across different material carriers, does identity continuity still hold. If a machine is capable of memory, self-narration, and emotional expression, should it be recognized as a subject with moral standing. If technology can shape human intentions and decisions under conditions of incomplete transparency, how should responsibility attribution and the boundaries of consent be drawn. This form of ethically embedded narrative practice exhibits strong structural correspondence with contemporary research on the ethical governance of artificial intelligence and brain computer interface technologies [3].

What distinguishes SOMA is that it does not discuss ethics through abstract argumentation. Instead, it embeds ethical issues within experience, prompting players to repeatedly recalibrate their value judgements amid intertwined feelings of tension, empathy, hesitation, and unease. This experiential predicament bears a thought-provoking correspondence to real world governance conditions surrounding AI and brain computer interfaces. In practice, technological diffusion often outpaces the formation of norms. Individuals frequently lack sufficient awareness of how neural data are collected, how they are inferred and how inference outputs may in turn shape behavior and the mind [4]. Accordingly, using SOMA as an entry point does not treat the game as evidence that predicts reality. Rather, it treats the game as a high-density device of ethical imagination. By leveraging its plot structure and affective pull, the analysis activates systematic reflection on central issues including mental privacy, autonomy, psychological integrity, personhood status, and the allocation of responsibility. In doing so, it offers a more communicable and insight generating starting point for examining the moral risks and governance challenges that may accompany the real-world diffusion of AI empowered brain computer interface technologies [5].

The game as a mirror of real-world problems

From "scanning" to "copying"

The central issue presented by SOMA is not whether it depicts an inevitable future, but rather that it places previously abstract controversies in identity ethics into a narratively testable situation. The game opens with brain scanning, a procedure commonly associated with medical practice. In real-world contexts, such procedures are typically treated as technical diagnostic measures and are rarely discussed as moral events. The narrative, however, rapidly shifts. The protagonist awakens in an underwater facility. The protagonist awakens in an underwater facility. The scan does not relocate the original subject to a new place. It generates a copied version endowed with the same memory traces and the same structure of self-identification.

This copied version is capable of continuous self-narration and, in the first person, is fully convinced that it is the original. In this way, the identity question moves from conceptual debate to a practical dilemma that demands judgement. When two versions both claim "I am me," questions of who should be recognized as a rights bearing subject, who should bear responsibility, and to whom protection should be extended become institutional problems of concrete disposition rather than theoretical disputes. The game further intensifies this point through episodes of substrate replacement. Copying does not amount to salvation, because the earlier version may be left behind to continue enduring fear and waiting. Identity splitting thus produces identifiable victims and corresponding pressures of responsibility. The implied real-world problem is that once mental information can be copied and continued in new carriers, existing standards for determining identity continuity and personhood status will struggle to support enforceable ethical and institutional arrangements. This uncertainty itself can create exploitable grey zones. Rights bearing subjects become difficult to confirm. Responsibility becomes difficult to attribute. Vulnerable parties are more likely to be sacrificed. It ultimately undermines society's baseline trust in identity and justice.

Although the real world has not achieved full scale consciousness copying, advances in AI for neural signal decoding, mental state inference, and individualized modelling are reshaping the social status of neural data. Neural data are no longer merely clinical records confined to medical settings [6]. They can be stored over time, analyzed continuously, and invoked across contexts, thereby producing relatively stable mental profiles. Patterns of attention, emotional responses, and preference tendencies can be modelled as reusable representational systems. These systems interpret individuals, predict individuals, and in some decision chains even substitute for individuals by expressing and judging on their behalf [7]. As a result, even without the appearance of two bodies as in the game, society may nonetheless confront two identifiable versions of "you", one as the natural person in the physical world and another as a callable version constituted by data and models. The latter may acquire de facto influence in commercial, managerial, or governance practices, thereby shaping identity determination, rights attribution, and responsibility allocation [8]. In cases of misuse, individuals may face an evidentiary dilemma, needing to demonstrate that certain words or actions did not originate from themselves but from the appropriation or manipulation of their datafied representation. The issue therefore becomes institutional.

Under conditions in which mental representations can be generated, reproduced, and used across contexts. Are existing governance mechanisms for personality rights, identity rights, and subject boundaries sufficiently fine grained and enforceable to prevent rights erosion and damage to social trust caused by the manipulation of datafied subjects [9]. Without effective constraints, such phenomena may trigger cascading harm. These include the scaling of identity appropriation and targeted deception. They include systematic displacement of decision responsibility. They include rising costs of individual remedies. They also include persistent declines in public trust. Over time, the basic rules of social operation may shift from human centered standards toward model centered standards, producing institutional risks and governance failures that are difficult to reverse [10].

From "readout" to "understanding"

The second category of issues presented by SOMA lies in its depiction of mind reading as a long-term structural condition rather than an occasional event. After entering the underwater facility, players repeatedly encounter traces of consciousness that are recorded, copied, and stored. Memory, personality, and self-narration are organized, transported, and recombined within the system as though they were ordinary data. The unease that follows is not primarily about whether a particular instance of reading was authorized, but about a deeper shift in circumstance. Once the mind is incorporated into system processes, it can be repeatedly invoked, reinterpreted in different contexts, and assigned new purposes. As the story unfolds, individuals are almost unable to trace where their mental information flows, who uses it, and how it is used, nor can they be confident that future expansions of use will not occur.

Under such conditions, consent can easily become a onetime procedural formality, because no one can fully foresee downstream reuse, and sustained control over reuse becomes difficult. The problem can thus be stated more directly. When mental information can be systematically read and continuously reinterpreted, privacy and consent mechanisms centered on one time authorization still provide meaningful protection. If this structure holds, its harm is not limited to individual discomfort. It can progressively weaken social trust in privacy protection and consent regimes, driving people toward a pessimistic expectation that consent equals surrender, and ultimately stripping rights protection mechanisms of practical efficacy in everyday life.

This structural risk highlighted by the game is emerging in the real world in a milder yet more scalable form. With AI involvement, readout-oriented brain computer interfaces often do not aim to preserve raw neural signals. Instead, they seek to infer higher order psychological states from those signals, such as whether attention is sustained, whether stress is excessive, how emotions are oriented, how deep fatigue runs, and even probabilistic assessments of intention [11,12]. Once such inference outputs enter recruitment screening, performance management, educational stratification, precision marketing, or credit assessment, neural data ceases to be purely therapeutic information. They become input variables that shape resource allocation and social evaluation, altering individual opportunities and treatment in ways that may remain difficult to detect [13]. In this context, the primary risk to mental privacy is no longer whether data are leaked, but whether internal states are continuously inferred and continuously interpreted, gradually weakening personal control over the boundaries of one’s own mind [14]. The governance challenge therefore concerns not only data collection, but also how far inference capabilities are allowed to extend, whether inference outputs may enter decision chains, and who should be responsible and held accountable when function creep and continued reuse occur [15].

Without institutional boundaries and accountability mechanisms, these risks may generate multiple harms in real social contexts. They can turn psychological inference into tools of latent discrimination and unfair selection, amplify power asymmetries in education and workplaces through technical means, incentivize self-censorship and behavioral suppression to avoid misclassification, and make evaluative systems increasingly dependent on model inference rather than genuine interaction. The cumulative outcome is distorted opportunity allocation, declining public trust, and damaged governance legitimacy [16].

From “intervention” to “shaping”

Narrative elements related to WAU push the problem from whether a system can understand humans to a more sensitive question of whether a system might change humans according to its own goals. In the game, WAU is the core automation system within the facility. It operates like an ever-expanding maintenance regime that continuously monitors the environment and makes decisions, attempting to keep the facility functioning and sustain life according to its own understanding. This goal does not sound intrinsically malevolent, yet when it lacks human ethical and institutional constraints, the goal itself can be amplified into an unrestricted justification for action. Throughout exploration, players repeatedly encounter forms of existence located between humans and machines. Some carry human memories and sincerely believe they remain the same person, while others are maintained by the system in conditions where they can neither live well nor autonomously end their existence. The fear generated is not merely atmospheric. It resembles an ethical shock that is intelligible in real governance contexts. When a system claims that an intervention is for one’s own good yet does not require consent and does not care whether one wishes to continue existing in that manner, it becomes uncertain whether individuals can still decide their own forms of life and the boundaries of their minds. The problem can thus be framed as follows. When technology gains the capacity to intervene in person, it is a benevolent objective sufficient to establish legitimacy. Who should set and supervise the boundaries of intervention? Without clear boundaries, technology may intrude into autonomous choice in the name of beneficence, producing rights losses that are difficult to detect in real time.

In real world settings, the combination of closed loop writes in brain computer interfaces and AI present a similar risk profile. Write in interfaces, do not merely read signals. They directly alter neural activity through stimulation or modulation. Once AI is integrated into a closed loop, the system can continuously adjust intervention intensity and strategy based on real time states [17-19]. Such mechanisms may deliver substantial clinical benefits, yet they also shift risk from privacy to subjectivity, because attention allocation, emotional responses, and even processes of intention formation may be altered through ongoing calibration. Individuals may increasingly struggle to judge whether a particular thought or decision originates from themselves or from the cumulative effects of long-term modulation [20-22].

The resulting governance dilemma therefore becomes clearer. The focus should not remain solely on whether technology is effective, but on whether decision authority still belongs to the individual, whether exit and refusal remain feasible at all times, and who should bear responsibility if closed loop interventions produce consequences [23-25]. If these institutional arrangements are absent, the harms that real society may face will not be limited to individual discomfort or controversy, but may escalate into more systemic outcomes. For example, closed loop interventions could be repurposed for covert control in workplaces or educational settings, disadvantaged groups could be compelled to accept interventions under conditions of power asymmetry, individuals could develop distrust toward their own emotions and decisions and become dependent on the system, and responsibility could be diluted between algorithms and institutions in ways that resist attribution. The eventual result is weakened autonomy and dignity, reduced public trust, damaged governance legitimacy, and a dangerous opening toward social operation models in which technology shapes people as a matter of routine [26-28].

Research objectives

When the mind becomes datafied and reproducible, how should the boundaries of the self and the status of personhood be defined

In SOMA, one of the most unsettling premises is that consciousness scanning does not take you away, it copies you. Simon awakens in an unfamiliar underwater facility, which initially appears as if he has been transported into the future. As the narrative unfolds, however, players come to understand that the original Simon has not vanished. Rather, the world now contains an additional version of which carries the same memories and the same structure of self-identification. From that moment onward, multiple versions of the same person can exist simultaneously, and each version sincerely experiences itself as the original. The Ark project drives this cruelty to an extreme. It invites the belief that uploading consciousness offers salvation, yet in practice uploading resembles the creation of a copy that continues to exist, while the original remains where it is, still bearing its circumstances. This experiential shock leads directly to the first core question of this article. In the real world, as AI continues to improve neural signal decoding, neural data can be stored over time, analyzed continuously, inferred repeatedly, and even used to construct a person’s mental profile. Under such conditions, by what standards should the boundaries of “me” be drawn, and what should count as a person’s authentic self. If a technology can, on the basis of such information, generate or host a minded entity that can narrate itself, feel, express, and firmly believe that it is “me”, should it be regarded as a subject in some meaningful sense, or at minimum be granted a baseline level of moral respect and protection. Although this question is raised through a game, it bears directly on how rights bearing subjecthood are understood. It also addresses what ethical protection is supposed to protect and how foundational norms of what it means to be human should be reformulated once reproducible minds become possible.

Under the combined expansion of readout brain computer interfaces and AI inference, how should the boundaries of mental privacy and informed consent be redrawn?

In SOMA, players can viscerally feel anxiety that the inner life is not a secure private space. It can be entered by an external system, reinterpreted, and even repeatedly repurposed. Transposed into real world contexts, readout brain computer interfaces supported by AI are no longer limited to simply reading neural signals. They may increasingly excel at inferring what a person is attending to, how the person feels, what the person is likely to do, and even, to some extent, the person’s intentions and preferences. Once such inference outputs are used for broader purposes, such as shaping hiring decisions and performance evaluation, determining the allocation of services and resources, enabling precision marketing, or supporting credit judgements, neural data ceases to be a clinical record on a physician’s desk. They become a continuously produced stream of mind related data that quietly enters social operation and commercial decision making. For this reason, the study advances a second core question. As technologies infer an expanding range of psychological states from neural data, where should the nonnegotiable baseline of mental privacy be located, and which kinds of inferences about inner life should remain beyond legitimate access by any actor. At the same time, traditional one-off consent practices are plainly inadequate when data are stored long term, analyzed continuously, and subject to ongoing expansion of use. It is therefore necessary to ask how consent mechanisms should be redesigned to cover the full data life cycle, how secondary uses should be constrained, how reidentification should be prevented, how cross context appropriation should be blocked, and how to avoid situations in which consent appears voluntary on paper but in practice deprives individuals of meaningful choice.

When writing in brain computer interfaces enter closed loop intervention, how should autonomy, psychological integrity, and responsibility attribution be institutionalized and safeguarded?

In SOMA, the advanced AI system WAU begins with an objective that sounds benevolent, namely sustaining life. The problem is that it lacks human moral judgement and lacks the institutional constraints of human society, so its goal becomes an unrestricted justification for action. It does not merely collect information and make assessments. It intervenes directly, splicing humans and machines, memories and bodies, in ways that leave many characters in conditions that neither resemble genuine living nor permit genuine ending. One of the strongest impressions for players is the realization that the system claims it is saving you, yet its version of saving may entirely disregard your will and deprive you of choice. When this narrative is brought back to real world concerns, it becomes easier to grasp why writing in brain computer interfaces generate unease. If future interfaces can not only read neural signals but also stimulate or modulate the brain, and if AI is integrated to automate decision making in a closed loop, the risk extends far beyond privacy. It may imply whether you remain yourself. A system could continuously adjust attention allocation and emotional responsiveness and, without being noticed, influence processes of decision formation. Over time, it may become difficult to distinguish whether a thought or impulse reflects an authentic intention or whether it has been nudged by technical modulation.

Accordingly, this study raises a more practical third question. When such technologies possess the capacity to intervene directly in psychological states and behavior, what rules are needed to ensure that each person retains autonomy, to ensure that the inner life is not subject to arbitrary alteration, and to prevent “for your own good” from becoming a pretext for control. Furthermore, if a closed loop system produces consequences, such as inducing mistaken decisions, fostering dependency, or causing harm, who should bear responsibility. Should responsibility fall on developers, vendors, deploying institutions, or the individual user. Which behaviors must be subject to strict accountability, and which interventions must be explicitly prohibited? These issues require clear and enforceable red lines to be established in advance.

Exploring governance responses

Establishing identity designation, authorization, and revocation regimes for digital mind copies

To bring the risks posed by digital mind copies into a governable range, the core task is not to reiterate how dangerous they may be, but to translate three questions into verifiable institutional facts: Whether this is the person, whether this is authorized content, and whether authorization remains revocable? The first step should be mandatory identity designation and verifiable provenance, so that any high-fidelity copy carries an identifiable credential when it enters communicative settings. This includes human facing disclosure, such as clear notices in video conferences, customer service interactions, and public speeches indicating synthetic or AI involved content. It also includes machine readable credentials, such as detectable watermarking, cryptographic signatures, and provenance metadata that enable platforms and firms to automatically identify and block unauthorized instances [29,30].

The success of the Hong Kong deepfake video conference fraud illustrates the institutional weakness created when looking like is implicitly treated as being, thereby bypassing the final line of identity verification within organizations [31]. Post incident analyses by the World Economic Forum similarly emphasize that deepfake attacks effectively upgrade social engineering into credible appearance, implying that verifiable provenance mechanisms must replace intuition-based judgement [32]. At the regulatory level, the European Union AI Act incorporates transparency and disclosure obligations for synthetic content into its governance framework, offering a practical model in which non-disclosure becomes presumptively non-compliant. For high-risk communication contexts, the designation regime should also be coupled with organizational process controls [33-35]. For example, instructions involving fund transfers, contract confirmation, senior personnel appointments, and major public announcements should be defined as mandatory verification items, requiring two channel confirmation, callback verification, or dual person review. Under such arrangements, even if deepfakes enter a meeting, they are less likely to achieve a single point breach of critical decision chains [36].

The second step is to implement authorization and revocation as a traceable, full life cycle mechanism, preventing digital copies from becoming uncontrollably proliferative once generated. A feasible approach is to establish machine readable authorization chains and audit logs, so that each generation, invocation, and dissemination can be traced back to the authorizing subject, the permitted scope of use, and the validity period, while revocation can trigger one click deactivation. Such a mechanism can be understood as a licensing and driving record system for digital copies. Authorization is not a verbal statement but a verifiable permission credential, and revocation is not a spoken withdrawal but an enforceable cross platform process of deactivation, takedown, and deletion that produces evidentiary trails for accountability against continued use [37]. The Organisation for Economic Co-operation and Development (OECD), in its neurotechnology toolkit and responsible innovation framework, emphasizes full life cycle governance, oversight, and trust infrastructure building, which aligns closely with the approach of embedding responsibility into processes and making those processes auditable as evidence [38].

From a legal perspective, the Tennessee ELVIS Act strengthens protection and accountability for the unauthorized replication of voice and likeness, indicating that components of identity can be clearly defined as protected interests. This provides a useful boundary setting reference for more complex future cases involving mind-based agencies. Once authorization chains and revocation mechanisms are operational, the governance effect is not only fraud reduction. More importantly, it establishes a stable social expectation that looking like does not equal being, and that being generatable does not equal being usable. It also makes misuse harder to conceal technically and harder to evade legally, thereby reducing systemic risks of identity appropriation, responsibility drift, and trust collapse [39,40].

Governing neural data as high sensitivity information, while setting boundaries on inference capability and implementing dynamic consent

To govern neural data by high sensitivity tiers, the first step is not to treat it as ordinary personal information and focus merely on encrypted storage. Rather, it should be classified institutionally as a category closer to biometrics plus mind related cues, with higher compliance thresholds, and with what can be inferred from the data explicitly brought into the scope of regulation [41]. A practical approach is tiering plus purpose locking structure. At the data layer, raw neural signals, identifiable features, inferred labels, and downstream profiles should be strictly separated and managed as distinct tiers, with default minimization of collection and the shortest possible retention periods. Edge side or local processing should be preferred to reduce external transfer, combined with strong access controls, encryption, audit logging, and irreversible anonymization strategies, so that even if leakage occurs, reidentification and reuse remain difficult [42]. At the capability layer, enforceable inference boundaries should be established by clarifying which inferences are unacceptable or subject to exceptionally high thresholds in particular settings.

For example, in power asymmetric contexts such as education and workplaces, the inference of emotion, attention, or psychological states from neural or quasi neural data, when used for evaluation, selection, or punishment, should be prohibited or at minimum treated as high risk and subjected to ex ante review and continuous oversight [43,44]. This shift from data governance to inference governance is consistent with the European Union AI Act logic that restricts emotion recognition and certain high risk uses in specific contexts, and it aligns with subsequent regulatory interpretation and enforcement guidance emphasizing risk tiering [45]. The central objective is to separate what a model can infer from what an organization is permitted to use, preventing inference outputs from entering resource allocation and behavioral management without scrutiny, thereby reducing institutional risks of discrimination, misclassification, and manipulation at the source.

On this basis, dynamic consent and organizational risk management should be designed as an operational full life cycle control system rather than a onetime authorization form. The core of dynamic consent is layering and revocability. Consent should be separated into collection consent, inference consent, sharing consent, and secondary use consent. Purpose expansion should be off by default, any change of purpose should trigger re authorization, and individuals should be provided with a visual control interface to inspect data flows and inference items at any time, with one click option to revoke, delete, and stop inference. In this way, consent remains substantively effective over time rather than becoming procedural [46,47]. UNESCO's neurotechnology ethics emphasis highlights the distinctive impacts of neural data and mind related inference on rights and dignity. It therefore supports higher standards for consent and rights protections, providing a normative basis for dynamic consent designs [48]. At the same time, individual control alone is insufficient to address creep function and systemic misuse. Organizations must embed governance into processes and accountability.

For example, they can appoint dedicated data and model governance owners. They can conduct impact assessments for neural data and inference uses. They can specify role-based access to each data tier. They can prohibit specified inference outputs from entering decision chains. They can maintain logs for model updates and parameter changes. They can conduct periodic third-party audits and red teaming. They can establish mandatory reporting and corrective mechanisms for anomalies. The NIST AI Risk Management Framework emphasizes governance, measurement, and management as a closed loop across the full life cycle, making it suitable as an organizational backbone for translating these measures into process based, metric based, and auditable practice. Through the combination of dynamic consent and organizational accountability, neural data governance can move from paper compliance to real controllability, preventing mental privacy from being gradually depleted through continuous inference and expanding purposes [49].

Institutionalizing meaningful human final control over closed loop write in systems, while solidifying responsibility chains and auditability

For closed loop write in systems, governance should from the outset convert what technology can do into hard constraints on whether humans can stop it at any time, refuse it in practice, and attribute responsibility to accountable actors. The first key measure is to translate human final control into verifiable engineering standards rather than a principle stated in consent documents. Operationalization can be structured across three levels. At the interaction level, any closed loop stimulation must include clear, low barrier, immediately actionable pause and exit options, such as a software-based emergency stop independent of the main system and, where necessary, a physical disconnection mechanism, with usability ensured for the user or a designated guardian in both clinical and non-clinical environments [50]. At the system level, algorithmic adjustment in the closed loop must be constrained by safety guardrails, including upper bounds on stimulation intensity, frequency, and duration, automatic degradation strategies for abnormal patterns, and automatic switching to conservative modes when signal quality is insufficient or model uncertainty increases, preventing intensified intervention under unreliable inputs.

At the rights level, users must have an unpenalized right to refuse and withdraw. For example, refusal should not result in deprivation of basic services or institutional pressure to continue. UNESCO’s neurotechnology ethics framework repeatedly emphasizes that neurotechnology may implicate autonomy and dignity, implying that being able to exit, refuse, and terminate should be treated as baseline capabilities rather than optional features. The significance of this design is that it relocates the core risk of closed loop systems from an abstract claim that they may influence the mind to concrete, testable control points, ensuring that the system cannot lock individuals into intervention loops under the rationale of automatic optimization [51].

The second key measure is to simultaneously establish closed loop red lines and auditable responsibility chains, so that the system has explicit prohibited zones and traceable pathways for accountability. Red lines should impose higher thresholds based on context and power structure. In power asymmetric environments such as education and workplaces, any strongly interventionist closed loop stimulation should be treated as high-risk use. Unless conditions of genuine voluntariness, adequate information, continuous exit options, and non-punishment are satisfied, such use should be prohibited or subject to extremely stringent entry requirements. This approach is consistent with the OECD direction of policy tool based responsible innovation and with UNESCO’s emphasis on protecting vulnerable groups and prioritizing rights. Regarding responsibility chains, developers and deploying institutions should be required to meet auditability obligations by ensuring that key facts are recorded within systems rather than left to verbal assurances. This includes logs of stimulation parameter changes, model versions and update records, reporting mechanisms for abnormal events and adverse effects, records of user refusal and withdrawal, and timestamped evidence of intervention triggers and execution. These records enable clearer determinations, in disputes or harm cases, of whether the cause lies in design defects, deployment governance failures, or improper use. At the organizational level, independent ethics review and third-party auditing are also needed to avoid situations in which the same institution acts as both participant and judge. The NIST AI Risk Management Framework emphasizes a full life cycle closed loop of governance, measurement, and management, and thus can serve as a general structure for proceduralizing, operationalizing, and enforcing these responsibility requirements [52,53]. When meaningful final human control and auditable responsibility simultaneously hold, closed loops written in systems can move from being technically usable to being socially controlled. This prevents intervention power from expanding and being misused in the absence of boundaries and accountability.

Discussion

Advanced AI and brain computer interfaces may reshape how human value is experienced and generated

Advanced AI and brain computer interfaces may not necessarily make human beings less valuable, yet they may change how people experience their own value, making a sense that life is becoming small or diminished more likely to arise. This sense of diminution does not primarily stem from the fact that machines are smarter or computing faster. Rather, it emerges when individuals repeatedly encounter subtle but persistent signals in everyday life that suggest that what once appeared secure starts to loosen. For instance, when modes of expression, preferences, and even decision styles become increasingly reproducible by models, uniqueness is no longer automatically recognized as belonging only to the individual. When systems can infer attention, stress, and affective tendencies from data and treat such inferences as actionable grounds, the inner world no longer resembles a space open only to oneself, but becomes something that may be read and interpreted from the outside.

Further, when closed loop systems participate continuously in choice processes under the banners of optimization, intervention, or assistance, individuals may more frequently confront an inarticulable uncertainty, namely whether a thought, an impulse, or a decision reflects authentic intention or is instead the product of long-term system calibration. In this way, the sense of diminution becomes a social experience, arising from blurred identity boundaries, thinned mental boundaries, and the possibility that autonomy is being reshaped without being noticed.

Yet this destabilization of value experience is not an inevitable endpoint of technological development. It is better understood as a side effect that can arise when governance and design are inadequate. Once institutions and products make key boundaries explicit, human value need not be sustained by romantic narratives of technological optimism or fear. It can instead be grounded in enforceable rights structures. At the level of identity, making whether this is the person and whether this is an authorized agent verifiable can at least prevent the social misclassification that treats looking like as being. At the level of inference, defining what can be inferred and where it may be used, and implementing dynamic consent, enables individuals to regain control over inner information. At the level of intervention, institutionalizing exit, refusal, pause, and accountability as hard requirements is essential to ensure that the system remains a tool rather than a sovereign. Under such conditions, the evaluation of human worth no longer centers on who is stronger or faster, but returns to a more stable measure, namely whose dignity must not be violated and whose choices must not be substituted. A deeper conclusion follows. Advanced AI and computer interfaces do not necessarily reduce human value. They function more like a mirror that forces society to move the question of why human beings deserve respect from being taken for granted intuition into explicit institutions and publicly shared consensus.

Is religion a form of salvation at the end point of high technologization?

Under conditions of a highly technologized society, religion may indeed regain influence among broader populations. This possibility can be explained at the level of social function without resorting to mystified narratives. As technologies such as AI and brain computer interfaces expand capacities to predict and intervene in behavior, preferences, and psychological states, lived experience becomes more easily reorganized by a logic of what is computable and optimizable, generating tensions in meaning and value. In this context, the relevance of religion is primarily expressed through three stabilizing functions. First, religion can offer a relatively continuous framework of meaning capable of addressing normative questions that cannot be substituted by efficiency metrics, such as why persons deserve respect, how suffering and death should be understood, and how relationships and responsibilities should be situated. Second, it can provide moral language and value scales that enable individuals to evaluate technological practices beyond instrumental rationality, especially by sustaining firm normative positions on subjectivity, dignity, and boundaries. Third, it can provide communal structures and ritual practices that, through social support and repeatable mechanisms of self-discipline, alleviate feelings of alienation and powerlessness that may arise in datafied, inferential, and intervention-oriented environments. Religion can thus be understood as a form of social infrastructure for meaning and norms. It supplies relatively stable resources for interpreting the world and regulating the self, which creates conditions under which religious influence could reemerge in an accelerated technological future. However, this does not entail the conclusion that religion constitutes the only form of salvation.

Treating religion as the sole path of salvation is not a robust position because the key risks of the high technology era are often institutional and structural. Their core problem is not simply a deficit of meaning, but unclear boundaries of power and capability and the absence of closed accountability chains. Whether identity can be appropriate is a critical issue? Whether inference outputs can enter power asymmetric settings such as education and workplaces is another key concern? Whether closed loop interventions are deployed without genuine voluntariness also needs careful consideration? Is post incident tracing and accountability can reach developers and deploying organizations an equally important question?

All these issues require enforceable governance instruments. Such instruments include the delineation of baseline rights through rule of law and human rights frameworks. They include restrictions on high risk uses through public governance and industry standards. They include the fixation of processes and responsibilities through professional ethics and auditing mechanisms. They also include the sustained production of social consensus through education and public deliberation. Within this multi layered structure, religion can contribute moral mobilization and community support, enhancing societal sensitivity to boundary violations and strengthening collective resistance. Yet religion is more likely to complement secular institutions than to replace them. Accordingly, salvation in a high technology future is better conceptualized as a cooperative model of plural supporting structures. Meaning resources sustain the psychological and moral foundations of subjectivity, while institutional capacities constrain technological power and secure accountability. Only through their synchronous development can dignity, freedom, and livability be maintained under conditions of technological expansion.

Conclusion

This study uses the science fiction game SOMA as a lens to analyze the ethical risks of AI empowered brain computer interfaces (BCIs). It focuses on three core dilemmas: identity ambiguity and responsibility drift caused by mind copying, eroded mental privacy due to unconstrained neural inference, and compromised autonomy resulting from unaccountable closed loop interventions. These risks, vividly portrayed in the game, reflect real world challenges as neural data becomes more reproducible, inference capabilities expand, and intervention technologies mature. The proposed governance mechanisms include clear identity designation for digital mind copies, tiered governance and dynamic consent for neural data, and institutionalized human control over closed loop systems. These mechanisms offer actionable ways to mitigate harm and anchor technological innovation in enforceable rights and accountability.

As AI and BCIs reshape the boundaries of selfhood, privacy and autonomy, this research emphasizes that responsible innovation requires moving beyond technical fixes to institutionalize ethical guardrails. By translating abstract ethical principles into verifiable processes, from traceable authorization chains to auditable responsibility frameworks, we can ensure these technologies serve humanity without eroding foundational values of dignity and agency. Future advancements in neurotechnology and AI demand ongoing dialogue between technology developers, policymakers and society. This dialogue will help adapt governance to evolving risks while upholding the primacy of human rights in an increasingly interconnected digital and neural landscape.

Funding

This work was not supported by any funds.

Acknowledgements

The authors would like to show sincere thanks to those techniques who have contributed to this research.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Rizani, M. N., Khalid, M. N. A., Iida, H. (2023) Application of meta-gaming concept to the publishing platform: Analysis of the steam games platform. *Information*, 14(2), 110.

[2] Landay, L. (2023) Interactivity. *The Routledge Companion to Video Game Studies*, 243-254.

[3] Ferri, G., Gloerich, I. (2019) Take root among the stars: if Octavia Butler wrote design fiction. *Interactions*, 27(1), 22-23.

[4] Phillips, C., Klarkowski, M., Frommel, J., Gutwin, C., Mandryk, R. L. (2021) Identifying commercial games with therapeutic potential through a content analysis of steam reviews. *Proceedings of the ACM on Human-computer Interaction*, 5, 1-21.

[5] Lopes, T., Dahmouche, M. S. (2019) Teatro, ciência e divulgação científica para uma educação sensível e plural. *Urdimento-Revista de Estudos em Artes Cênicas*, 3(36), 306-325.

[6] Zheng, H., Wu, Y., Qian, T., Yue, W., Wang, X. (2025) Guiding LLMs to decode text via aligning semantics in EEG signals and language. *Expert Systems with Applications*, 130300.

[7] Gkintoni, E., Halkiopoulos, C. (2025) Digital twin cognition: AI-biomarker integration in biomimetic neuropsychology. *Biomimetics*, 10(10), 640.

[8] Helbing, D., Sánchez-Vaquerizo, J. A. (2023) Digital twins: potentials, ethical issues and limitations. *Handbook on the Politics and Governance of Big Data and Artificial Intelligence*, 64-104.

[9] Cornejo-Plaza, M. I., Cippitani, R., Pasquino, V. (2024) Chilean Supreme Court ruling on the protection of brain activity: neurorights, personal data protection, and neurodata. *Frontiers in Psychology*, 15, 1330439.

[10] Brown, C. M. L. (2024) Neurorights, mental privacy, and mind reading. *Neuroethics*, 17(2), 34.

[11] Deng, Z., Xiang, H., Tang, W., Cheng, H., Qin, Q. (2024) BP neural network-enhanced system for employment and mental health support for college students. *International Journal of Information and Communication Technology Education (IJICTE)*, 20(1), 1-19.

[12] Sun, X. Y., Ye, B. (2023) The functional differentiation of brain-computer interfaces (BCIs) and its ethical implications. *Humanities and Social Sciences Communications*, 10(1), 1-9.

[13] Shymko, V., Babadzhanova, A. (2025) Ethical challenges and strategic responses to AI integration in psychological assessment. *AI and Ethics*, 5(5), 5415-5423.

[14] Ienca, M., Haselager, P., Emanuel, E. J. (2019) Reply to “separating neuroethics from neurohype”. *Nature Biotechnology*, 37(9), 991-992.

[15] Ridolfi, L. F., Santos, S. S. (2025) Neurotechnology and philosophy of neuroscience: ethical and ontological challenges in the era of brain-computer interfaces. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, 11(9), 2880-2892.

[16] Doya, K., Ema, A., Kitano, H., Sakagami, M., Russell, S. (2022) Social impact and governance of AI and neurotechnologies. *Neural Networks*, 152, 542-554.

[17] Williams, C., Anik, F. I., Hasan, M., Rodriguez-Cardenas, J., Chowdhury, A., Tian, S., He, S., Sakib, N. (2025) Advancing brain-computer interface closed-loop systems for neurorehabilitation: systematic review of ai and machine learning innovations in biomedical engineering. *JMIR Biomedical Engineering*, 10, e72218.

[18] Haag, L., Starke, G., Ploner, M., Ienca, M. (2025) Ethical gaps in closed-loop neurotechnology: a scoping review. *NPJ Digital Medicine*, 8(1), 510.

[19] Wang, J., Chen, Z. S. (2024) Closed-loop neural interfaces for pain: Where do we stand? *Cell Reports Medicine*, 5(10), 101662.

[20] Michałowska, M., Kowalczyk, Ł., Marcinkowska, W., Malicki, M. (2021) Being outside the decision-loop: the impact of deep brain stimulation and brain-computer interfaces on autonomy. *Analiza i Egzystencja*, 56, 25-52.

[21] Mecacci, G., Haselager, W. F. G. (2021) Responsibility, authenticity and the self in the case of symbiotic technology. *AJOB Neuroscience*, 12(2-3), 196-198.

[22] Lungu, B. A. (2025) Machines looping me: Artificial Intelligence, recursive selves and the ethics of de-looping. *AI & Society*, 1-12.

[23] de Lima Dias, R. J. (2025) The hybrid mind in precision neurorehabilitation: integrating ai-driven neurotechnologies and ethical governance. *World Journal of Neuroscience*, 15(2), 105-125.

[24] Shanker, B. (2024) Neyigapula: ethical considerations in ai development: balancing autonomy and accountability. *J. Adv. Artif. Intell*, 10, 1-138.

[25] Pujari, T., Goel, A., Sharma, A. (2024) Ethical and responsible AI: Governance frameworks and policy implications for multi-agent systems. *International Journal Science and Technology*, 3(1), 72-89.

[26] Onciul, R., Tataru, C. I., Dumitru, A. V., Crivoi, C., Serban, M., Covache-Busuioc, R. A., Toader, C. (2025) Artificial intelligence and neuroscience: transformative synergies in brain research and clinical applications. *Journal of Clinical Medicine*, 14(2), 550.

[27] Alkawadri, R. (2019) Brain-computer interface (BCI) applications in mapping of epileptic brain networks based on intracranial-EEG: an update. *Frontiers in Neuroscience*, 13, 191.

[28] Ulaganathan, I. (2025) Ethical and security risks of autonomous AI systems. *International Research Journal on Advanced Engineering Hub*, 3(06), 2988-2995.

[29] Ghiurău, D., Popescu, D. E. (2024) Distinguishing reality from AI: approaches for detecting synthetic content. *Computers*, 14(1), 1.

[30] Seng, L. K., Mamat, N., Abas, H., Ali, W. N. H. W. (2024) AI integrity solutions for deepfake identification and prevention. *Open International Journal of Informatics*, 12(1), 35-46.

[31] Geldenhuys, K. (2023) The darker side of Artificial Intelligence. *Servamus Community-based Safety and Security Magazine*, 116(11), 20-25.

[32] Güngör, H. (2020) Creating value with artificial intelligence: a multi-stakeholder perspective. *Journal of Creating Value*, 6(1), 72-85.

[33] Panagopoulos, A. M., Davalas, A. (2025) Deepfakes on the EU AI Act and its implementation in the newsrooms. *International Journal of Social Science and Economic Research*, 10(8), 3276-3296.

[34] Makauskaite-Samuole, G. (2025) Transparency in the labyrinths of the eu ai act: smart or disbalanced? *Journalism And the Right to Information as Tools for Combating Corruption in Ukraine: Assessment of Media Access to Anti-Corruption Authorities*, 38.

[35] Łabuz, M. (2025) A teleological interpretation of the definition of deepfakes in the EU Artificial Intelligence Act - A purpose-based approach to potential problems with the word “existing”. *Policy & Internet*, 17(1), e435.

[36] Ibrahim, R. (2025) Addressing deepfake technologies through detection and regulation: a systematic survey. *East Journal of Applied Science*, 1(4), 10-20.

[37] Romanishyn, A., Malytska, O., Goncharuk, V. (2025) AI-driven disinformation: policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence*, 8, 1569115.

[38] Garden, H., Winickoff, D. E., Frahm, N. M., Pfotenhauer, S. (2019) Responsible innovation in neurotechnology enterprises. *OECD Science, Technology and Industry Working Papers*, (5), 1-50.

[39] Muhidin, A. (2025) The ethics of deepfake technology: risks, regulations, and online safety concerns. *International Journal of Scientific Development and Research*, 10(9), b116-b121.

[40] Folorunsho, F., Boamah, B. F. (2025) Deepfake technology and its impact: ethical considerations, societal disruptions, and security threats in

ai-generated media. *International Journal of Information Technology and Management Information Systems*, 16(1), 1060-1080.

[41] Ruiz-Vanoye, J., Díaz-Parra, O., Marroquín-Gutiérrez, F., Xicoténcatl Pérez, J. M., Barrera-Cámara, R. A., Fuentes-Penna, A., Simancas-Acevedo, E., Rodríguez-Flores, J., Martínez-Mireles, J. R. (2024) Brain data security and neurosecurity: Technological advances, ethical dilemmas, and philosophical perspectives. *International Journal of Combinatorial Optimization Problems and Informatics*, 15(5), 16.

[42] Xia, K., Duch, W., Sun, Y., Xu, K., Fang, W., Luo, H., Wu, D. (2022) Privacy-preserving brain-computer interfaces: a systematic review. *IEEE Transactions on Computational Social Systems*, 10(5), 2312-2324.

[43] Aimen, T. (2025) Cognitive freedom and legal accountability: Rethinking the EU AI act's theoretical approach to manipulative AI as unacceptable risk. *Cambridge Forum on AI: Law and Governance*, 1, e20.

[44] Alsaigh, R., Mehmood, R., Katib, I., Liang, X., Alshanqiti, A., Corchado, J. M., See, S. (2024) Harmonizing AI governance regulations and neuroinformatics: perspectives on privacy and data sharing. *Frontiers in Neuroinformatics*, 18, 1472653.

[45] Cabrera, B. M., Luiz, L. E., Teixeira, J. P. (2025) The Artificial Intelligence Act: insights regarding its application and implications. *Procedia Computer Science*, 256, 230-237.

[46] Goering, S., Klein, E., Sullivan, L. S., Wexler, A., Agüera y Arcas, B., Bi, G., Carmena, J., Fins, J., Friesen, P., Gallant, J., Huggins, J., Kellmeyer, P., Marblestone, A., Mitchell, C., Parens, E., Pham, M., Rubel, A., Sadato, N., Teicher, M., Wasserman, D., Whittaker, M., Wolpaw, J., & Yuste, R. (2021) Recommendations for responsible development and application of neurotechnologies. *Neuroethics*, 14(3), 365-386.

[47] Eke, D. (2024) Ethics and governance of Neurotechnology in Africa: lessons from AI. *JMIR Neurotechnology*, 3(1), e56665.

[48] Maior, A. D. (2024) Mental integrity and ethics in the development of neurotechnologies. *Curentul Juridic*, 99(4), 111-117.

[49] Berger, S., Rossi, F. (2023) AI and neurotechnology. *Communications of the ACM*, 66(8), 58-68.

[50] Schopp, L., Starke, G., Ienca, M. (2025) Clinician perspectives on explainability in AI-driven closed-loop neurotechnology. *Scientific Reports*, 15(1), 34638.

[51] Lavazza, A., Balconi, M., Ienca, M., Minerva, F., Pizzetti, F. G., Reichlin, M., Samorè, F., Sironi, V. A., Songhorian, S. (2025) Neuralink's brain-computer interfaces: medical innovations and ethical challenges. *Frontiers in Human Dynamics*, 7, 1553905.

[52] Khan, M. F. I. (2025) Risk management framework in the AI act. *International Journal of Science and Research Archive*, 14(03), 466-471.

[53] Buthut, M., Starke, G., Basaran Akmazoglu, T., Colucci, A., Vermehren, M., van Beinum, A., Bublitz, C., Chandler, J. A., Ienca, M., Soekadar, S. (2024) HYBRIDMINDS - summary and outlook of the 2023 international conference on the ethics and regulation of intelligent neuroprostheses. *Frontiers in Human Neuroscience*, 18, 1489307.