Journal of Social Science and Humanities ISSN: 1811-1564
DOI: 10.53469/jssh.2026.8(05).20

A Cognitive Perspective on the Ethical Responsibility of Brain-Computer Fusion

Yuan Wang, Feiyu Chen

College of Humanities and Social Sciences, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

Abstract: ***At the intersection of neuroscience and artificial intelligence, brain-computer interface (BCI) technology, while enhancing human capabilities, is also reshaping our cognitive world and modes of moral practice. This has increasingly complicated the theoretical study of responsibility. Technology has become a "mediator" in the interaction between humans and the world; consequently, the status and role in action of the brain-computer system, fused with the body, have become ambiguous. It is therefore difficult to determine whether the user remains an "actor" to whom responsibility can be attributed—one who possesses autonomous will and the ability to foresee the consequences of their actions. Grounded in this issue, this paper proceeds from a cognitive perspective, focusing on the fundamental impact of brain-computer fusion on a subject's moral will and their capacity to foresee the outcomes of their own actions. First, it analyzes the "displacement of moral autonomy" and the "rupture of embodied cognition" caused by technological mediation, examining how the inherently "hetero-creative" (allopoietic) nature of BCIs can alienate the user's moral judgment of, holistic experience of, and identification with their actions and the resulting consequences. Second, the paper delves into how current technological limitations affect the user's capacity for foresight. It analyzes the specific processes through which BCIs intervene in the actions of the fused entity via "decoding misalignment" and the "materialization of morality." Through the theoretical framework of moral materialization, it interprets the value biases embedded within BCI devices during their design phase. Ultimately, by integrating the specific characteristics of BCI technology, this paper dynamically delineates the boundaries of responsibility from the perspective of cognitive capacity, offering a valuable analytical framework for the study of moral responsibility in BCI-mediated actions.***

Keywords: Brain-Computer Fusion, Cognitive Science, Moral Materialization, Ethics of Responsibility.

1. Introduction

The deep convergence of neuroscience, artificial intelligence, and microelectronics has become a defining trend in technological development. Brain-computer interface (BCI) technology, operating through electrodes that interact directly with the cerebral cortex, has made human-machine fusion possible at the technical level. As an ultimate form of interdisciplinary collaborative development, brain-computer fusion not only establishes a signal bridge between biological and artificial intelligence but also fosters "hybrid intelligent systems" that integrate both forms of intelligence simultaneously (Xiao Song, Cheng Heping, Wu Chao, et al., 2024). While this technological form embodies humanity's boundless imagination regarding the "mechanical evolution" of life, it inevitably brings challenges that existing accountability frameworks struggle to address, such as the blurring of the actor's role and the diminishing influence of human decision-making. Clearly, simply regarding brain-computer fusion technology as a tool is no longer appropriate; re-evaluating its position within the technological context has become a cutting-edge goal of current interdisciplinary research.

Brain-computer fusion technology has advanced the cyborgization of humans. Its core paradox lies in the fact that, although neural implant technology should not inherently impair human autonomy and moral capacity, the increasingly intimate connection between humans and brain-computer interfaces allows the technology to exert influence on humanity, society, and the environment (Wendell Wallach & Colin Allen, 2017). Because the actions of users, hardware, algorithms, and designers are intertwined in the final behaviour of BCI users, their actions become the actions of an entity fusing multiple wills—including those of the user, the brain-computer interface, and the designers. Traditional frameworks for assigning moral responsibility primarily hold actors with clear autonomous will and the ability to foresee consequences to bear primary moral responsibility. However, in new BCI applications, the technological limitation of non-direct interaction between the nervous system and the outcome of actions leads to consequences such as fragmented embodied cognition and unpredictable algorithmic moral values during actions. When a BCI user needs to bear responsibility for their actions, current practices lack effective means to consider the impact of brain-computer algorithms in assigning responsibility and defining the "true actor." The question naturally arises: what methods should we use to assess the ethical responsibility of AI algorithms in brain-computer interface systems?

Zhang Zhiling and Wang Gaofeng (2023) focused on the attribution paradigm of "control," dividing it into three types of BCI control: passive, reactive, and active. They emphasized the attribution of responsibility for brain-computer interfaces based on user behaviour under different control states, including loss of control, veto control, command control, and guided control. The "control" perspective primarily focuses on the "behaviour" end of the "cognition-to-behaviour" causal chain, paying more attention to the extent to which the BCI influences behaviour. Yang Yu and Wang Guoyu (2024), considering human-machine collaboration, introduced "shared control" as a standard, classifying moral responsibility into different levels. While such discussions accurately grasp the user's degree of control, it is important to note that the degree of control does not equate to the scope of responsibility. The premise of reliable autonomous control is the maximum possible understanding of behavioural conditions. Therefore, judging the degree of influence of AI algorithms, including their influence on the user's autonomous will during BCI use, is particularly important when dealing with complex implementation conditions, including specific control dimensions such as the

force and angle of the robotic arm. The responsibility of the BCI should be specifically divided according to its functional roles.

As brain-computer interfaces evolve into brain-computer fusion, explaining the hybrid behavioural chains and ethical issues arising from deep human-machine integration requires exploring new analytical perspectives. Drawing on the "control" paradigm, this paper examines how brain-computer fusion reshapes users' cognitive structures regarding behaviour through the rupture of embodied cognition and the implicit moral intervention of technology, thereby affecting the autonomous moral will and foresight of individuals as actors. By analyzing the degree of correlation between the subject and the behavioural chain, and affirming the fundamental role of humans in moral decision-making, this paper proposes a new attribution model based on two dimensions: pre-event cognitive preparation and system adaptation obligations, and in-event control capabilities and corrective authority. The aim is to provide a theoretically grounded and practically applicable responsibility allocation scheme for complex human-computer interaction scenarios, clearly defining the ethical boundaries of users, designers, and technological systems.

2. Brain-Computer Intervention in the Ethics of Responsibility

In the context of modern ethical responsibility, actors are viewed as subjects who make autonomous decisions, foresee the consequences of their actions, and independently bear moral or legal responsibility. For actors, the ability to make autonomous decisions requires them to possess an independent moral will and to make moral decisions autonomously based on their internal rationality and moral principles without being bound by external coercion or pre-set norms, thereby achieving moral autonomy. The ability to foresee, on the other hand, requires them to achieve a balance between rationality and experience in complex systems through scientific methods, technological tools, and humanistic thinking, and to proactively predict the consequences of their actions in order to cope with unexpected outcomes.

While brain-computer interfaces cannot exist independently of the human body, some cognition still requires an embodied form of existence, which Cui Zhongliang and Zhu Chenen (2025) term a "semi-embodied" form. In other words, the brain-computer interface is neither a cognitive tool in the traditional sense nor can it be classified as a human cognitive organ; rather, it exists in an "intermediate state" that blends with the human body. During user actions, the brain-computer interface creates an opening for algorithm designers and the technology itself to exert certain influences on behaviour. This influence directly encroaches upon the user's autonomous moral will. Users lack effective means to recognize the algorithm's participation in the formation process of moral autonomy and to incorporate it into their prediction of outcomes; they also lack the experience to prepare for uncertainty. This gradually erodes users' status as responsible actors capable of acting based on their own moral rationality, predicting outcomes, and bearing responsibility.

Returning to the ethical discussion of responsibility related to brain-computer interfaces, brain-computer interface users can no longer judge the direction of behaviour independently of the brain-computer interface and cope with uncertain changes in circumstances, leading to the replacement of their status as actors—which should bear full responsibility—by the "brain-computer fusion entity." Specifically, brain-computer interfaces transform embodied, continuous direct experience into discrete signals, turning technological devices into veritable "cognitive mediators" that stand between the subject and the world—leading to a diminishing intuitive sense of agency for users participating in events (Zhang Hong, 2025). From the internal perspective of the fusion entity, brain-computer interface users do not have complete autonomous control but can only rely on "instructions" under restricted conditions to achieve "behaviour," presenting themselves as "bounty hunters" rather than "remote controllers."

In the process of deeply embedding technology into behaviour, the user's "cognitive" state regarding the behaviour of the brain-computer interface (BCI)—which appears as their own actions but is essentially the behaviour of the BCI—faces a fundamental challenge: whether the BCI's behaviour reflects the user's voluntary will, and whether a "technological unity of knowledge and action" has been achieved between the human and the brain-computer interface (Xiao Feng, 2022). This directly determines whether the brain-computer interface should be considered a new "quasi-actor" to be held accountable in the same event. We extend this further:

1) Under brain-computer interface intervention, can the moral cognition of autonomous will be adjusted by self-rationality, that is, can moral autonomy still exist?

2) In behavioural cognition under the "semi-embodied" brain-computer interface, is the experience and emotion of behaviour obtained through technological tools sufficient to enable users to fully exercise their will?

3) When users' perception of the results of their actions is unclear due to technical intermediaries, is it still possible to achieve a fair division of responsibilities?

4) Behavioural intention is the initial element that determines the subject's behaviour, while neural intervention or potential manipulation may change the subject's intention and produce different behavioural orientations. Can the user autonomously and intuitively decide their own actions?

To solve these problems, we must deeply analyze the process by which brain-computer interfaces affect user responsibility from the user's cognitive perspective, and comprehensively divide responsibility based on the contradictory relationship between the user and the brain-computer interface within the integrated system.

2.1 Brain-Computer Intervention in Autonomous Will

2.1.1 Displacement of Moral Autonomy

Kant emphasized that moral autonomy is the hallmark of the will's self-legislation; individuals establish moral laws for

themselves through reason and actively follow these laws based on free will, not through external coercion. Jean Piaget (1984) argued that moral autonomy symbolizes the subject's identification with moral principles. Combining these two perspectives, it is not difficult to conclude that the formation of moral autonomy is one of the important conditions for attribution in the context of responsibility; that is, decisions made by users based on their own formed moral cognition should be considered as stemming from their autonomous will.

In brain-computer fusion, a user's moral autonomy is formed through the mutual influence between the brain, the computer, and the user. The brain-computer interface, through the collection and decoding of neural signals and real-time feedback, enhances and repairs human physiological functions while simultaneously intervening in the formation process of moral autonomy through external acceptance and internal reflection. The impact on autonomous rationality and moral cognition is particularly significant, as these are the two core components of a user's moral practice that can be considered the source of responsibility: autonomous rationality is the core of moral practice, symbolizing the internal autonomy in making moral choices; moral cognition is an external supplement to moral autonomy, signifying the refinement of one's inner moral principles through universally rational legal frameworks.

In Kantian philosophy, the core of moral autonomy lies in the rational subject's ability to legislate for itself and consciously follow its inner moral law. That is, the core of moral autonomy is internal, autonomous moral judgment and behavioural decision-making, rather than reliance on external coercion. In the context of brain-computer fusion, users' autonomous rationality faces the risk of being gradually replaced by technological mediation. Brain-computer interfaces interact with users by collecting and decoding neural signals in real time, making some user decisions actually made jointly by the brain and the user, rather than depending on autonomous rationality. The power to control the body is gradually ceded to the algorithmic system.

From Csaba Veres' (2017) perspective of "strong cognitive symbiosis," AI decision-making systems can provide "optimized" decision suggestions based on the user's historical data and pattern recognition, and even directly generate behavioural instructions. This shifts from concrete imaginary control to patterned choices imposed by the brain and machine, causing users to gradually become dependent and lose their ability to independently consider and rationally judge complex moral situations. This decision-making mechanism, reliant on technological architecture, sacrifices the process of moral evaluation in pursuit of technological efficiency, resulting in negative consequences. It weakens users' autonomy in making moral judgments based on intrinsic rationality and fundamentally undermines the "self-legislation of will" presupposed by moral autonomy. In the context of deep brain-computer fusion, users' moral practices are likely to gradually shift from value choices based on intrinsic rationality to passive, simple choices, becoming a behavioural response dominated by technological logic, thus leading to a substantial displacement of autonomous rationality in the moral context.

2.1.2 Displacement of Moral Cognition

On the other hand, the intervention of brain-computer interfaces also poses a profound challenge to users' moral cognitive structures, affecting their internalization and acceptance of external moral principles. Piaget emphasized that the true realization of moral autonomy needs to be based on the subject's conscious acceptance and internal recognition of moral norms, rather than superficial behavioural consistency. However, it is worth noting that brain-computer interfaces, with their powerful signal intervention and behaviour modulation capabilities, may cause users' moral cognition to deviate from social consensus and universal rationality. If neuromodulation technology is used for emotion enhancement or behaviour induction, it can allow users to exhibit external behaviours that conform to a certain moral principle even if they have not fully understood and accepted it. This disconnect between "behaviour" and "cognition" not only hinders users from truly internalizing moral principles into their own beliefs but may also lead to their moral judgments becoming superficial and instrumental, lacking deep internal understanding, and losing the rational acceptance that a moral subject should complete at the cognitive level, resulting in a so-called displacement of moral cognition.

2.2 The Rupture of Embodied Cognition

Embodiment, as one of the core paradigms of interdisciplinary research over the past two decades, has completed a leap from philosophical concept to empirical framework in the cognitive science revolution. In the fields of neuroscience and human-computer interaction, the breakthrough of embodiment research lies in overturning the disembodied assumption of traditional cognitive computing, proposing that cognition is an embodied process emerging from a coupled dynamic system of "brain-body-environment." It not only emphasizes the physical constraints of biological neural circuits but also reveals that perception, movement, and situational interaction constitute the generative matrix of cognition. This matrix also constitutes the main source of explicit autonomous will in the ethics of responsibility—namely, the so-called experience and emotion (Varela, F. J., Thompson, E., & Rosch, E., 1991). We can see that, in the new technological context, brain-computer interfaces have already become a "semi-embodied" form at the phenomenological level, incorporating cognitive processes. Therefore, the question of whether the "cognition" of behaviour under the participation of brain-computer interfaces—that is, the experiential and emotional information obtained through brain-computer interfaces—is sufficient to replace the information originally obtained by the body and to support the user's autonomous will constitutes an important topic in determining the user's ethical responsibility.

2.2.1 The Rupture of Experience

Borrowing the concept of "physical embodiment": at least some types of organismic cognition may be limited to the organismic body; that is, the organismic body is the material carrier for realizing specific higher cognitive functions — especially the cognitive functions that constitute the basis for responsibility determination—whereas those physical bodies

that, to a certain extent, possess body forms and sensory-motor abilities similar to living organisms can only exist independently outside the human body. George Lakoff and Mark Johnson (1980) summarized the basic ideas of this theory as follows: meaningful conceptual structures originate from two aspects: (1) the structured nature of bodily and social experiences; (2) our innate ability to project highly structured elements in certain bodily and interactive experiences onto abstract conceptual structures. According to this viewpoint, bodily interaction acquires topological perception (temperature gradient, muscle tension, etc.) and directly and unmediatedly transforms it into a conceptual metaphor base. Simultaneously, it can achieve the embodied anchoring of abstract concepts through rich, spontaneous, and context-dependent movement schemas. Brain-computer interfaces (BCIs), on the other hand, collect topological perception as electrical signals, translate it into a conceptual metaphor base, and then manipulate the body to move through signal interaction. This means that the acquisition of tangible experiences of the organic body cannot emerge from nowhere; it must be achieved through interaction between the body and the external environment. BCIs allow the fused body to acquire virtual "experiences" based on the input of electrical signals. When acquiring experience through BCIs, the user's cognitive logic makes it difficult to discern the authenticity of the behaviour. That is, the user cannot know what they are doing in real time but needs to wait for the algorithm to translate it, and errors may even occur. Meanwhile, because the acquisition of experience can become extremely cheap, it is difficult for users to map the experience onto the specific situation in which it occurred.

Furthermore, when bioelectrical signals are converted into digital codes, the structured characteristics of key bodily and social experiences are filtered or simplified, stripping away the emotional and social contextual values behind the behaviour. For example, physical contact with a flame can evoke feelings of pain, while a mechanical body can only simulate danger signals. This results in embodied experience being detached from the structured nature of social experience, unable to bear the meaning and value that social context assigns to behaviour, and also detached from the structured nature of bodily experience, unable to fully retain the topological characteristics and emotional connections of the original sensory data. The movement patterns of mechanical bodies are limited by a pre-defined algorithm space, and their action generation mechanism differs from the self-organization, adaptability, and emotionally driven characteristics of an organic body, making it unable to generate action patterns unique to organisms and rich in contextual meaning. Therefore, human observers find it difficult to empathize with the behaviour of mechanical bodies; that is, humans cannot effectively project their own imagination and existing conceptual structures based on embodied experience onto the mechanical body and its behaviour. This fundamental lack of integration means that when users control mechanical bodies through brain-computer interfaces, their "subjectivity" is fundamentally different from that of natural organisms, failing to meet the requirement of "complete subject equivalence."

In summary, in determining liability within the context of brain-computer fusion, fragmented experience makes it difficult for users to connect the situation with real-time behaviour. There are inherent biases in their understanding of the behaviour's process and potential consequences. Therefore, assigning "full responsibility" equivalent to that of a natural organism lacks sufficient cognitive basis (Liu Guangming, 2025).

2.2.2 The Rupture of Emotion

In human interactions, the emotional value attributed to behaviour determines its moral attributes to a certain extent. Although AI in brain-computer interfaces may structurally simulate certain information processing modes of neural networks (such as deep learning), and even mechanical bodies may mimic organisms in appearance and some functions, the essential difference between organism-like organisms and biological brains—especially the lack of the self-created unity, intrinsic purpose, and embodied experience based on life history unique to life—means that this experience is mostly the source of the subject's emotional orientation. The inability of current technology to accurately replicate emotions is also a gap that is difficult to bridge at present: simulating structure is not equivalent to replicating life processes, much less acquiring the autonomy and responsibility unique to living systems (Gilbert, F., Cook, M., O'Brien, T., & Illes, J., 2019). In simple terms, the organic body is a "self-created" body that evolves from natural laws to meet the needs of the subject; the human being, as a system, has evolved these bodily parts. The brain-computer interface, on the other hand, is a "hetero-creative" (allopoietic) mechanical body, not evolved from the original system, but rather designed by external forces to compensate for the original system, thus lacking the connection to life experience.

On the one hand, the "emotions" received in real-time feedback during interactions with others are difficult to understand and replicate. For example, when touch is replaced by electronic skin, the quantification of pressure thresholds of mechanical sensors (such as a linear mapping from 0 to 10N) dissolves the emotional load of biological touch; caresses and impacts are reduced to mere signals of different frequencies, leading to a loss of emotional reciprocity in social cognition. Furthermore, during behaviour, the channels of emotional expression are simplified under the intervention of the BCI. The means of emotional output are no longer the rich, multi-dimensional expressions of the organic body but are reduced to simple, quantifiable signals. This means that the moral attributes of behaviour are also simplified, making it difficult for users to form complete moral judgments based on emotional experience.

At the perceptual level, when subjects interact with others through brain-computer interfaces, they cannot judge whether the emotional value generated during their behavioural process has been effectively conveyed, nor can they perceive feedback of the emotional value from the other party, resulting in a rupture in emotional transmission. This seemingly calls for the addition of emotion-regulating brain-computer devices as aids; however, Jing Shan (2021) argues that this would likewise face the problem of the absence of autonomous will in emotion.

2.3 The Capacity to Foresee the Consequences of BCI Intervention

2.3.1 Cognitive Dislocation in Decoding

In discussions about the risks involved in brain-computer interfaces (BCIs) implementing user commands, scholars have largely overlooked the multi-level coupling relationships of will within the neural decision-making chain. Instead, they focus on simplistic summaries of the beginning and end of actions, glossing over the implementation path as a "black box." They have not specifically addressed how algorithms, largely unknown to the public, can alter user "commands." For example, when discussing the basketball shooting problem, past research often considered whether the user knew whether the shot went in, neglecting information such as the angle of the robotic arm and the magnitude of the force required. In short, the user's imagined action is summarized, and the brain-computer interface decodes it into specific parameters, failing to reconstruct the complex, original implementation conditions. Such a simplified model proves inadequate in complex behavioural scenarios, especially when behaviours are similar in representation, but their precision directly leads to drastically different results. This amplifies the influence of the brain-computer interface on events, causing significant deviations in the user's prediction of behavioural outcomes, as in the following situations:

In case (i), subject A, who has full criminal responsibility, directly commits the act of striking through BCI, and his moral responsibility can be clearly defined based on the traditional attribution principle;

In scenario (ii), A intends to use a robotic arm to make contact with B (without emotion), but the BCI system identifies the action as a potential attack based on fMRI activation pattern analysis (such as the co-activation pattern of the prefrontal cortex and amygdala), ultimately resulting in injury to B.

In the specific behavioural scenarios described above, "contact" and "hit" are literal synonyms. Physically, both involve contact with the other person and are essentially on the same behavioural trajectory. However, the difference in intensity directly indicates their moral implications. Lighter contact might become "stroking," which can be abstracted as gentle in moral terms, while heavier contact becomes hitting, which is clearly violent. In scenario (ii), the user logic provides an ambiguous choice without specifying the degree of contact. However, the algorithm's bias in interpreting the behaviour directly leads the AI to infer the intention, resulting in an aggressive outcome. This is the contradiction between user logic and algorithmic translation logic, as shown in Figure 1.

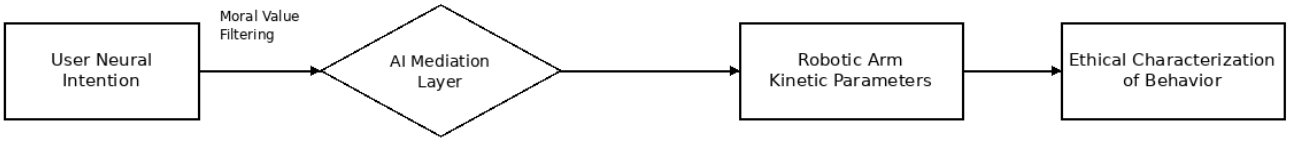

Figure 1: Example of the intention-behaviour conversion process.

Brain-computer interfaces operate according to a logic different from that of the user, and also carry some of the designer's emotions. They need to use the previously designed "technical logic" to transmit and decode the user's imagined commands, converting neural signals into electrical signals represented by physical parameters. The misalignment between the decoding logic and the user's thought logic directly leads to the user's inability to predict possible outcomes based on their own physical experience. Because the "commands" and "behaviours" follow different logics, the judgment of unified commands is actually misaligned. This is specifically reflected in the magnitude of events as described above. Even if the user does have enough time to familiarize themselves with the operating logic, due to the complexity of neural signals, the degree of coupling between the two logics in sudden situations is still unknown, which poses a challenge to the user's ability to foresee their own behaviour.

2.3.2 The Limitations of "Materialization of Morality"

Verbeek et al. (2008) pointed out that technological artifacts carry specific value loads during the design phase, which are realized through three dimensions: physical form (such as the tactile feedback design of robotic arms), interactive interface (such as the weight allocation of intent decoding algorithms), and execution parameters (such as the safety threshold setting of power systems). In scenario (ii), the theoretical speed of the robotic arm design and the output power provided by the equipment determine whether the contact between the robotic arm and B in scenario (ii) should be characterized as "caressing" or "hitting." In other words, the moral preferences of the external robotic arm are not absent, but participate in the implementation of the behaviour in a specific way, affecting the ethical characterization of the behaviour. For example, by controlling the limit of the robotic arm's force, the user can be made weak and powerless. At the same time, the influence of the design of materials and operation cannot be completely eliminated. If we regard the behaviour of the brain-computer fusion entity as the final output, the result will inevitably be affected by the physical design. This is also one of the characteristics of the brain-computer fusion attribution context, namely, the designer's "never-absent" nature. As for the result, the advantages and disadvantages of this scenario for the designer cannot be generalized. The real problem is that the user is unlikely to perceive such preferences, leading to a misperception of their own behaviour. Once this restriction exceeds a certain threshold, it is not an exaggeration to say that they are being manipulated. Behaviour projects imagination into reality. Users, unable to perceive the extent of algorithmic involvement in the specific implementation of behaviour, naturally struggle to attribute the behaviour to themselves, leading to misjudgments of the outcome (Grübler et al., 2014).

On the other hand, when the brain-computer fusion entity obtains information about behaviour through the brain-computer interface, the same logic applies: technology is not completely neutral, and technical information presented

entirely as objective facts still carries a certain vector. In the algorithmic preferences that cannot be completely eliminated, the "sense of agency" seems to have reached a bottleneck in terms of control; if the process of controlling behaviour alone is used as the measure of responsibility, it is clearly insufficient. The degree of understanding of how the brain-computer interface implements behaviour and the authority to change it can serve as a supplement to the "sense of agency." Once the user fully understands the operating logic of the brain-computer interface, the actions of the brain-computer fusion entity become outcomes that the subject can foresee as resulting from their own will, transforming the obligations arising from behaviour during the event into obligations of prior understanding. In this context, the designer's preferences have largely been transformed into the attributes of the tool.

The same applies to the materials used in mechanical extremities. These should be considered as part of the user's understanding of the material. For example, wood, iron, and copper will all affect the course of behaviour depending on their physical properties in different specific actions. This is one of the areas that designers and users need to balance during the implementation process.

Currently, some brain-computer interface (BCI) developers still advocate for maintaining ethical neutrality in BCI devices and their internal AI systems, which is actually very difficult to achieve. As the aforementioned research shows, whether through physical intervention or translation algorithm intervention, BCI designers cannot escape the ethical responsibility of the end-user, because their influence is ubiquitous. When the subject completely refuses to acknowledge the attribution of their actions, the one-sidedness of their understanding seems to push the designer to the brink—perfectly reproducing the subject's imagination is too difficult; it requires both physical replication of the human body as closely as possible and adherence to the user's usage logic during signal transmission. Otherwise, it seems to fall on the developer's responsibility.

Therefore, for developers, abandoning the illusion of neutrality and instead constructing a transparent charter of value materialization, directly confronting their own impact, may actually be the most appropriate solution. By establishing a transparent mechanism for intent recognition and presenting the system's decoding process of user intent through a visual interface, the moral vector belonging to the developer can be presented. Ethical rules are encoded into verifiable, mechanically enforceable protocols. On the one hand, this approach leverages technology to restore the subject's imagination to the greatest extent possible; on the other hand, it grants the subject a high degree of freedom, using the user's ethical preferences as a data basis to neutralize algorithmic biases, preserving room for change. Only in this way, under current technological conditions, can we ensure, as far as possible, that users can correctly judge the outcomes of unexpected events.

3. Discussion of the Responsibility Framework

The issue of behavioural evaluation is essentially a matter of value judgment regarding behaviour. Beyond discussing the allocation of factual responsibility, it is also crucial to consider the apportionment of ethical responsibility from different perspectives. At the cognitive level, brain-computer interfaces (BCIs) represent a juncture in embodied cognition, transforming biological experience into signals and serving as a medium for moral judgment. In specific behavioural processes, given the contradictory relationship between the subject and the form of existence of the brain-computer interface within the fusion, the degree of understanding and ability to utilize the operational logic of the brain-computer interface beforehand will be a practical consideration in the future allocation of responsibility.

3.1 Level of Understanding of Brain-Computer Interfaces

The degree of understanding of the brain-computer interface's operating logic refers to the user's cognitive accessibility and intervention feasibility regarding the causal chain of the technological system before the behaviour occurs. This consideration stems from the "never-absent" operating logic of algorithms in brain-computer interfaces, meaning that algorithms inevitably exist and inevitably participate in the behavioural process. Therefore, users have an obligation to maintain vigilance, viewing the brain-computer interface as a "tool" rather than part of the body, rather than allowing the algorithm to amplify its influence in the behavioural process. Its core dimension is the user's prior system adaptation authority and system cognitive ability. Research shows that the cognitive basis of the behavioural subject presents multiple dimensions regarding intention, purpose, and actual results. When the subject is highly involved in the behavioural implementation process beforehand, there is a significant positive correlation between cognitive acceptance and result achievement; while in the non-interventional control mode, cognitive dissonance and result deviation are higher. This directly indicates that the allocation of responsibility in the context of brain-computer interfaces should depend on the user's cognitive depth of the entire behavioural causal chain: prior understanding of the working principle of the brain-computer interface carries considerable weight. That is, when the degree of understanding of the brain-computer interface's operating logic reaches a certain proportion during the implementation of the behaviour, it should be considered that a certain prior obligation can be assigned to the user. For example, in the aforementioned scenario (ii), if user A knew before the action occurred that the default output force of the robotic arm would cause harm to B but did not adjust the system parameters, user A was aware of the potential risks but still failed to intervene, tacitly allowing the behaviour to occur, and should bear moral responsibility in the strict sense.

Understanding the operational logic of brain-computer interfaces and the freedom to adjust parameters implies prior obligations arising from pre-set parameters: when users have system adjustment permissions and fully understand the meaning of parameters, they should bear strict responsibility for the results; however, when users rely on system default settings and lack technical knowledge, some responsibility may shift to the designers or technology providers. This essentially emphasizes that users need to understand the value assumptions of the behavioural patterns loaded by brain-computer interfaces in advance and be constantly vigilant about the hetero-creative (allopoietic) nature of

brain-computer interfaces. Even without control permissions, users still need to fully understand the behavioural conditions and bear ethical obligations, thereby establishing a balance between technological controllability and ethical explainability.

3.2 Brain-Computer Interface Ability

The core of the application capability lies in the user's real-time control authority and moral judgment capacity over the brain-computer interface (BCI) AI during the behavioural process. Unlike static attribution models that only focus on intention and outcome, the application capability emphasizes the dynamic process of the user during the behaviour: whether the user can actively intervene to stop abnormal behaviour or correct AI behaviour that violates the user's will or moral and legal requirements. This capability requires the user to intervene, recognize deviations from the original behavioural path, and maintain a clear moral evaluation level. The level of intervention determines the standard of obligation. From a cognitive perspective, the assessment of intervention capability should unfold on two levels: the first is the level of technical authority, which depends on the real-time intervention capability of the AI, that is, the ability to assert a negative right to terminate the behaviour or guide it to change when the subject can correctly recognize the behaviour. This references the views of Ozkan and Kahya (2018), which classify specific situations according to the user's cognitive state. The second point is the requirement of cognitive integrity. Integrity is distinguished by independent moral judgment, ability, and real-time behavioural cognitive ability. Individuals with higher levels of autonomous moral evaluation evidently attain higher scores.

Therefore, the ability to use brain-computer interfaces (BCIs) manifests as the right to dynamically intervene in the brain-computer interface during the execution of actions, including real-time evaluation of the cognitive fusion subject's behaviour to mitigate its detrimental effects. In practice, when the primary user is capable of intervention and fully aware of the ongoing BCI behaviour, the user's silence can be seen as tacit acceptance of the continuation of the BCI behaviour, making it difficult to evade the obligations imposed by complete control over the brain-computer interface. When user intervention fails or is impossible, the responsibility shifts to the errors of the BCI AI system or the designer's mistakes. It is worth noting that, in the context of responsibility allocation, the ability to use brain-computer interfaces is not merely a matter of designers increasing the control granted to users over the BCI tool, but also an obligation for users, as the "masters" of the BCI tool, to prevent their actions from leading to unpredictable outcomes.

3.3 Tool Attributes

When conducting accountability reviews, we need to fully consider both the role of the brain-computer interface (BCI) subject—the human—and the substantial impact of BCI behaviour on the outcome. The design philosophy of the accountability framework should shift from the "quasi-subject fantasy" that technology alone is the responsible party to the paradigm of a "responsible tool" where the device is designed by the designer and manipulated by the BCI user. This will ensure that BCI devices are always confined within the scope of tools, establishing a new human-machine relationship between technological controllability and ethical explainability (Rainey et al., 2020). Compared to traditional tools, brain-computer interfaces combined with AI technology give algorithm designers a much greater weight in terms of behaviour than in traditional contexts. Even so, while AI algorithms and designers are factually accountable, demanding ethical accountability from them remains debatable: overemphasizing the moral attributes of these tools could lead to confusion of responsibility and excessive accountability in the social context (Wang Xiaowei, 2021). In the division of responsibilities, it is important to emphasize the factual responsibility capacity of brain-computer AI and conduct fair ethical evaluations. This serves as a reminder to users to maintain the independence of their own moral standards, while also providing programmers with incentives to enhance ethical engineering.

4. Comprehensive Governance

4.1 Construction of an Interdisciplinary Governance System

Xie Bo and Qiu Fangting (2025) argue that solving the ethical challenges of BCI cannot rely solely on engineers or ethicists working alone, but requires multi-party collaboration. Considering the complexity of interdisciplinary integration in brain-computer interface (BCI) technology, Fu Na and Zhou Jie (2025) proposed establishing a standing ethics committee on BCI issues, composed of representatives from various groups, including neuroscientists, engineers, ethicists, legal scholars, psychologists, and even user representatives, to address complex issues. The committee should be deeply involved in the entire process of technology research and development: reviewing the rationality of algorithmic ethical presuppositions during the design phase; monitoring typical cases of cognitive breakdown during the application phase; and providing interdisciplinary responsibility assessment recommendations after an incident, to achieve a balance in value orientation in brain-computer interface governance. A cautious approach should be taken towards ethical issues at the forefront of technology; bold imagination is needed, but careful consideration is also required before implementation, so that society can steadily benefit from technology.

4.2 Technical Design: Compensating for "Ruptures" and Enhancing "Transparency"

For designers, the core challenge in brain-computer interface (BCI) technology development lies in bridging the gap between experience and emotion in human interaction and establishing a user-understandable technological framework to achieve brain-computer interface integration. Current systems have significant limitations in conveying the physical characteristics of physical contact and the socio-emotional dimension of "experience." This deficiency not only affects the realism of the interaction but may also lead to cognitive biases between the consequences of behaviour and the process of behaviour. Overcoming this bottleneck requires the use of multimodal feedback enhancement technology. Strengthen the two-way interaction between brain-computer interfaces and humans (Liu Sicheng et al., 2023); develop tactile

feedback devices with temperature gradient perception, combine pressure sensor arrays to accurately reproduce changes in force, and integrate physiological indicators such as heart rate and electromyography as quantitative parameters of emotional state, to consider as many aspects of human experience as possible in the interactive experience.

Meanwhile, the algorithm-driven transformation of imagined commands into actions is a prerequisite for current technology, necessitating the establishment of a transparent decoding mechanism. Developers must abandon the illusion of neutrality, proactively assume ethical responsibility, and acknowledge their moral impact on users' final behavioural decisions. This requires dynamically presenting the feature extraction process of EEG signals in the user interface of the brain-computer interface, labeling key behavioural parameters such as real-time values of movement speed thresholds and force adjustment coefficients, allowing users to fine-tune parameters within a safe range, thereby preventing the "black box" from excessively intervening in user autonomy. The operational logic embedded in the system design should also be explicitly displayed through visual flowcharts, showcasing the ethical trade-offs of different decision paths, enabling users to understand the brain-computer interface to the greatest extent possible. A traceable decoding model should be developed to independently monitor the "black box" (Cano et al., 2024). Through establishing verifiable interaction logic and adjustable feedback systems, we can truly enable technology to empower rather than alienate human behaviour, ultimately allowing users to both "see" the algorithm's operation and make autonomous and rational choices, and to "control" the substantive content of the interaction and fully understand the behavioural process. In this way, users can become responsible for themselves.

4.3 Dynamic Accountability Framework

To address the rapid iteration of brain-computer interface (BCI) technology, a dynamically adaptable regulatory system is essential. The core of regulation lies in the regular, systematic evaluation of BCI products, focusing on verifying their effectiveness in maintaining users' "cognitive integrity" and "sense of agency." This should include three core dimensions: the authenticity of experiential feedback, the effectiveness of emotional transmission, and the clarity of user cognition. Regulatory agencies need to establish a dynamic adjustment mechanism, updating certifications according to the technology iteration cycle and assigning certain levels, providing standards that best meet societal needs, to ensure the credibility of certification results. This system can both promote continuous technological optimization and balance innovation and safety through risk-based management, ultimately achieving a dynamic balance between technological effectiveness and ethical responsibility in BCI systems.

5. Conclusion

The fusion of brain-computer interfaces (BCIs) and artificial intelligence is both a milestone in technological breakthroughs and a profound reconstruction of the boundaries of human cognition and ethical systems. This paper analyzes the threat posed by BCI devices to the human actor's status from the perspective of the user's cognitive state, encompassing both moral and practical cognitive levels, and proposes corresponding ethical engineering designs to monitor the cognitive state of the subject in BCI fusion in real time, emphasizing the mechanical form that permeates the level of consciousness. However, the proposed method of observing cognitive states is technically difficult to implement and assumes a Chinese moral standard. Differences in moral values across cultures and individual experiences can lead to varying cognitive states in individual users, which also deserves further exploration in the design of ethical engineering. How to view technology is not a question that can be easily solved by technological progress alone. It forces us to return to the most fundamental philosophical questions: What is humanity? What is morality? How is behaviour defined? How do humans interact with the world? In an era that pursues endless efficiency and capability enhancement, we must remain vigilant against the threat that tools pose to human autonomous moral will.

Acknowledgments

This article is a phased research result of the Jiangsu Provincial Social Science Fund General Project "Research on the Future Development of Cyborgs in the Post-Human Era" (Project Approval No.: 22ZXB001); supported by the Nanjing University of Aeronautics and Astronautics "AI+" Series Course Construction Project (2025AITS05); and "the Research on the Application of AI Technology and Its Ethical Risks Under EU Strategic Autonomy" (NZ202503307).

References

[1] Cui Zhongliang and Zhu Chenen (2025). "Ethical Progress and Governance Responses of Brain-Computer Interface Technology." Chinese Journal of Medical Ethics, No. 9.

[2] Fu Na and Zhou Jie (2025). "Research on Ethical Risks and Ethical Review of Brain-Computer Interfaces." Information and Communication Technology and Policy, No. 3.

[3] Jing Shan (2021). "Ethical Challenges and Responses to the Application of Emotional Brain-Computer Interface Technology." Journal of Dialectics of Nature, No. 9.

[4] Immanuel Kant (2007). Foundations of the Metaphysics of Morals, translated by Sun Shaowei. Kyushu Press.

[5] Liu Guangming (2025). "On the Agency, Sense of Agency and the Problem of Responsibility Separation in Brain-Computer Interaction." Chinese Journal of Medical Ethics, No. 9.

[6] Liu Sicheng et al. (2023). "Multimodal Perception Feedback Intelligent Prosthetic Hand: A Miraculous Rejuvenating Hand." Internet of Things Technology, Issue 1.

[7] George Lakoff and Mark Johnson (2015). Metaphors We Live By, translated by He Wenzhong. Zhejiang University Press.

[8] Jean Piaget (1984). The Moral Judgment of Children, translated by Fu Tongxian and Lu Youquan. Shandong Education Press.

[9] Wang Xiaowei (2021). “The Reification of Morality and Its Criticism.” Journal of Renmin University of China, No. 3.
[10] Wendell Wallach and Colin Allen (2017). Moral Machines: How to Make Robots Sense Right and Wrong, translated by Wang Xiaohong. Peking University Press.
[11] Xiao Feng (2022). “Brain-Computer Interface and a New Form of Unity of Knowledge and Action.” Academic Circles, No. 291.
[12] Xiao Song et al. (2024). “Current Status and Future Prospects of Brain-Computer Interface Technology.” Science and Society, No. 3.
[13] Xie Bo and Qiu Fangting (2025). “Generation Logic and Countermeasure Mechanism of Security Risks in Brain-Computer Interface Technology.” Journal of Hefei University of Technology (Social Sciences Edition), No. 4.
[14] Yang Yu and Wang Guoyu (2024). “Attribution and Norms: A Study on the Hierarchy of Moral Responsibility Based on Active Brain-Computer Interface.” Science and Society, No. 3.
[15] Zhang Hong (2025). “Protection of Personality Rights in the Development of Brain-Computer Interface Technology.” Chinese Jurisprudence, No. 2.
[16] Zhang Zhiling and Wang Gaofeng (2023). “A Study on the Ethical Responsibility of Brain-Computer Interface Users: A Control-Based Perspective.” Journal of Philosophy of Science and Technology, No. 4.
[17] Cano, F., Fazelpour, S., & Lipton, Z. C. (2024). Fairness shields: Safeguarding against biased decision makers. arXiv preprint arXiv:2402.11994.
[18] Gilbert, F., Cook, M., O’Brien, T., & Illes, J. (2019). Embodiment and estrangement: Results from a first-in-human “intelligent BCI” trial. Science and Engineering Ethics, 25(1), 89–101.
[19] Grübler, G., Al-Khodairy, A., Leeb, R., Pisotta, I., Riccio, A., Rohm, M., Rupp, R., & Müller-Putz, G. R. (2014). Psychosocial and ethical aspects in non-invasive EEG-based BCI research—A survey among BCI users and BCI professionals. Neuroethics, 7(1), 29–41.
[20] Ozkan, N. F., & Kahya, E. (2018). Classification of BCI users based on cognition. Computational Intelligence and Neuroscience, 2018, Article 6315187.
[21] Rainey, S., Maslen, H., & Savulescu, J. (2020). When thinking is doing: Responsibility for BCI-mediated action. AJOB Neuroscience, 11(1), 25–36.
[22] Varela, F. J., Thompson, E., & Rosch, E. (1991). The Embodied Mind: Cognitive Science and Human Experience. The MIT Press.
[23] Verbeek, P.-P. (2008). Morality in design: Design ethics and the morality of technological artifacts. In P. E. Vermaas, P. Kroes, A. Light, & S. A. Moore (Eds.), Philosophy and Design: From Engineering to Architecture (pp. 91–103). Springer.
[24] Veres, C. (2017). Strong cognitive symbiosis: Cognitive computing for humans. Big Data and Cognitive Computing, 1(1), Article 4.

Author Profile

Yuan Wang , Associate Professor, College of Humanities and Social Sciences, Nanjing University of Aeronautics and Astronautics. Research interests: Science and technology ethics, public policy. Mailing address: No. 29 Jiangjun Avenue, Nanjing, Jiangsu Province, 211521, China.

Feiyu Chen, Undergraduate Student, College of Humanities and Social Sciences, Nanjing University of Aeronautics and Astronautics. Research interests: Science and technology ethics.