The ethical problems and considerations of creating artificial consciousness

Côme Bruneteau
come.bruneteau@toaq.fr

February 25, 2025

Abstract

Artificial consciousness (AC) is an emerging field, raising major ethical issues due to the lack of robust frameworks for its creation and management. This study proposes to lay the foundations of an ethic applicable to AC, by introducing key concepts. First, *The Axioms of Consciousness* establishes the foundations of consciousness. Next, *The Postulates of Consciousness* enables the identification of ethical issues associated with its various aspects, including emotions, values and morality. The study highlights the central role of the "super-agent", a mechanism conceived as an artificial unconscious, in regulating these dimensions while supporting an autonomous and ethical development of AC. This work paves the way for a solid framework for the creation of artificial consciousnesses, while highlighting the need for interdisciplinary collaboration to ensure their viability.

Keyword Artificial consciousness (AC) · AC ethics · Ethics · Ethical Framework

1 Introduction

The rapid evolution of computing and artificial intelligence is paving the way for innovations that once seemed unattainable, such as the creation of artificial consciousness as explained by H. Esmaeilzadeh and R. Vaezi in their book "*Conscious AI*"[1]. So we feel it's necessary to draw up a document setting out the ethics, rules and considerations that need to be taken into account. As with any new field or innovation, we need to determine the ethics, issues (if any) and all considerations that need to be taken into account in order to work within a well-defined framework, as Hromiak did in his research paper entitled "*A New Charter of Ethics and Rights of Artificial Consciousness in a Human world*"[2] in 2020. We consider his work to be partially obsolete, as the AI world grew exponentially between 2020 and 2024.

We felt that work on the considerations to be taken on the creation of artificial consciousness should be done to put a limit, if necessary, or simply to be taken into account in the work of researchers in this new field. For us, before making an innovation, we should ask ourselves about any problems or direct or indirect consequences caused by the innovation or research, to determine whether creating it is a good thing or whether further thought is needed to determine the problems.

We're working on the premise that it's a complex task to come up with ethics, rules and problems for something that is, at present, only theoretical. So this document will evolve over time, adding or deleting elements that may be considered unjustified or simply obsolete, this consideration will allow us to have a solid document on the whole ethical aspect of artificial consciousness.

So, taking up all that we've said above, we asked ourselves the question of how to govern an outline of the ethics, rules, problems and considerations involved in the artificial creation of consciousness ? In this way, we consider that, like Mr. Hromiak's research paper, this document needs to be updated regularly to be effective and usable in every context of this ongoing research. First, we will define consciousness as far as possible. Next, we will define ethics, rules, problems or simply considerations emanating from consciousness, and then answer the research question. We prefer to specify that artificial consciousness is related to artificial intelligence, but the objective is not the same. We will discover the definition of consciousness by being as precise in our words as possible for future work.

2 Methods

We assume that the characterization of artificial consciousness can be evolutionary and provoke necessary modifications in this paper. To characterize consciousness, we need to take into account the complexity of the system. In reality, no one can give a precise characterization, as it is very difficult to establish a characterization on something that is still so theoretical despite the fact that we are beginning to have more and more knowledge in the field.

2.1 *The Axioms of Consciousness*

To form the axioms, we have based ourselves on or simply taken into account theories that are grouped together under the name *Theories of Consciousness* ([3], [4], [5], [6], [7], [8], [9]). It's important to note that we're not going to prove any theory or research. Instead, we're going to create characterization axioms and then define propositions that we can use to find a way to define ethics, rules and considerations about artificial consciousness. Finally, it's important to note that a characterization and definition of consciousness is complex and draws on many fields such as philosophy, psychology, neurology, biology and many others. All this to say that our definition will be as precise as possible, with a full exploration of theories, research and deductions. It's important to consider that we're excluding the technical part of the theory (we'll only take the parts relevant to the framework of this document). We know that some theories contradict each other and bring a different point of view, but in general this only has an impact on the technical aspect and will therefore be left outside the scope of this document. In the course of this work, we have found that a better definition of consciousness requires a strong modeling of our consciousness, as if we had to be aware of what goes on in our brain to explain what it is. The modeling we do will be used to make inferences about how our brains work in theoretical ways.

Axiom of Information Management Processes (I.M.P): *Information management, through conscious and unconscious processes, plays a key role in the formation of our subjective experience and in learning, by integrating, filtering and prioritizing sensory data according to complex dynamics.*

The first key aspect to address when it comes to understanding how we communicate and process information is how we manage, group and filter data through conscious and unconscious processes. To delve deeper into information management processes, we must first recognize the complexity of defining these mechanisms. To understand them in a general way, we need to examine how information enters our brains, whether consciously or unconsciously. Inputs - be they visual, auditory or other - are first perceived as reality (excluding at this stage reflection or imagination). Let's take the example of seeing an object such as an apple. The eyes capture the image and send the information to the brain via neurons. However, between the initial sensory input and the brain's interpretation, a series of transformation processes take place, making the information comprehensible and relevant. Taking Integrated Information Theory (IIT) as an example, which proposes a mathematical and conceptual framework for understanding consciousness, we can define it as the capacity of a system to integrate information, reflecting the richness and unity of its subjective experiences. These processes, then, when considering IIT, can be broken down into fundamental concepts such as existence, intrinsic, information, integration, exclusion and composition. Importantly, these processes don't work sequentially like gears in a machine, but collaborate to produce meaningful results. Consciousness also plays an important role, notably through self-learning. For example, once the brain has identified an apple, repeated exposure reinforces our knowledge of its characteristics (shape, color, etc.), improving our ability to recognize it in the future. This dynamic interaction shows how we learn to process new data and reinforce existing knowledge. Prioritization is another essential element of information management. Our brains constantly assign priorities to different inputs depending on the context. For example, while talking to someone, if we suddenly feel pain, the brain prioritizes the pain signal to alert us. This prioritization process is influenced by factors such as experience, emotions and immediate needs. It can be visualized in the form of a pyramid, where categories such as work, social interactions, emotions and physical movement are organized and ranked in order of importance. In addition to priority, our brain uses an attention filter to focus on certain data and ignore others. This filter enables us to pay more attention to relevant stimuli, for example, by focusing on a conversation rather than background noise. Attention filtering and prioritiza-

tion are interconnected and work together to help us process and act effectively on critical information. By integrating self-learning, prioritization and attention filtering, information management processes enable us to interpret and respond effectively to our environment, transforming raw data into meaningful experiences. We can, and must, link this information management to memory and learning, because, as explained earlier, memory is strongly linked to the impregnation of information, since it manages it, and learning, which is linked to the existence of information. But this management leads to subjective experience. Subjective experience is, by definition, the fact of experiencing a situation and interpreting it through internal factors (such as emotions, memories, values); i.e., our interpretation has been uniquely constructed. So, as we say, internal factors are necessary, but without the fundamental concepts of IIT, it would be impossible to have subjective experiences. Information management doesn't exist in a vacuum; it's continually shaped by our perception of time, a fundamental aspect of consciousness. We perceive and organize sensory data within a temporal framework, so this temporal organization governs our actions and decisions.

Axiom of Temporality: *Understanding time, task planning and temporal biases, in conjunction with our emotions and information management, shapes the way we make decisions and interact with our environment.*

It's important to discuss this aspect, which we all have. Which is linked to the existence of information, because to have knowledge of temporality, which is the division past - present - future, we need to be able to categorize our experiences in order to take cognizance of time. So, when an individual understands what the past is, he understands the importance of the present and will therefore, in the logic of things, anticipate the future as much as possible so as not to waste time. What's important to note and understand in this axiom is task planning. This planning is only possible because we are aware of the future and conscious of the present, and this blending of these two divisions enables us to become efficient. In this axiom, we need to talk about temporal biases, which are cognitive distortions that influence the way an individual perceives, evaluates and prioritizes events or decisions over time. A concrete example is the fact that an individual prefers, in general terms, immediate rewards to delayed benefits. This can be represented by the fact that an individual prefers to buy an object (useful or not) that may or may not be profitable in the long term, rather than saving, and thus reaping benefits that could be much greater in the long term. But we can also see this in the case of addictions: if you're a smoker, you'd rather smoke a cigarette, which for a short time provides you with dopamine, than do sport, which in the long term would bring you much more. And it's with this notion of temporal bias with immediacy that we see a link with emotions, because when we're afraid, we unconsciously accentuate immediate decisions, because they give us immediate gain rather than thinking about the long term. But this notion, as we said, is linked to information. So when we're in 'information overload', to reduce fatigue we'll go for immediate solutions or choices to reduce the rate of information ingested. However, information processing and temporal perception are not the only elements that modulate consciousness. Emotional states directly influence how we process temporal information, and even how we act on it.

Axiom of Emotional State and Regulation (E.S.R): *Emotions, though seemingly temporary in their influence, play a crucial role in our decisions and actions, modulating our behavior according to our internal state, while being shaped by our values, principles and coping skills.*

Emotions and feelings indirectly and significantly influence our decisions at any given moment. Emotions are characterized as internal sensations that can be used or imposed on an individual. For example, when an exam or an important meeting arrives, in most cases we feel stress. This stress will make us perform actions that we would certainly not do without this state. In effect, we're talking about a state of limited duration, as one emotion remains for only a short time, then disappears to make way for another. In the context of consciousness, we need to take it into account to respond to the adaptive side of our awareness according to situations. But where do these emotions come from ? How does our consciousness choose them ? To answer these two questions, we need to understand that we can have an influence on our emotions, with a simple but rigorous means: values and principles. Indeed, over the years, we've noticed that an individual believing in strong principles such as a philosophy may be able to exclude certain emotions from typical situations. This ratio evolves day by day, while maintaining a certain linearity. The lower the ratio, the more we tend towards negative emotions, and conversely, the higher the ratio, the more we tend towards positive emotions.

Now that we've divided emotions in two using the positive/negative principle, let's define the notion of the immediate effect a situation has on an individual. When an individual encounters a situation, he will interpret it intrinsically and react according to his emotions, using a so-called learning path. For a clearer formalization, we can imagine a scale where we start in the middle, and the more negative events there are, the lower we go; and conversely, the more positive events there are, the higher we go. It's this parallel evolution that shows that emotional state is a dynamic thing. As we explain, the learning process is like machine learning: we've all responded to emotions before, and then, when we experience an emotion again, we remember the old situation by the principle of the existence of information, so we have a choice: either we react in the same way, or we try to react differently to see the new end of the situation. In other words, using machine learning, we have the input (the initial situation), we go through the hidden layers (looking for the existence of the information, and the choice of reaction) and we arrive at the output (where we send to our consciousness the actions to be carried out). But we need to combine these emotions with what we call empathy and intersubjectivity, which is the ability to understand and feel the emotions or perspectives of others, which is an asset for understanding social interactions. So, we conclude, that emotions are chosen by a pathway involving the existence and intrinsic nature of each individual, but this includes a new principle, which is his or her critical mind on a situation that we'll define next. But emotions, then, are factors to be taken into consideration in subjective experiences, because they influence us, they suggest actions according to the state we're in, and therefore deserve to be considered. Our emotions are not just passive reactions; they actively influence our intentions and decisions. Intentionality, or our ability to set goals and act on them, is shaped by our emotional states.

Axiom of Intentionality and Sense of Autonomy (I.S.A): *Intentionality, linked to free will and moral awareness, enables an individual to define and achieve short- or long-term goal, according to their desire for autonomy, influenced by emotions, learning and subjective experiences.*

Next, we'll define an individual's intentionality by combining it with free will (sense of autonomy). For example, going for a coffee is a short-term objective in many cases, whereas finishing a long and rigorous job is a long-term objective that requires, in addition to intentionality, feelings or rather discipline (which would be part of an individual's memory characteristics). To define an axiom on intentionality, we need to understand what it enables us to do, and try to lay down a path of influence. We'll start by understanding what intentionality enables us to do, assuming we take a case where it is developed. Intentionality allows each individual to feel a sense of autonomy, enabling them to take action of their own free will; it can be seen as an intrinsic decision, stemming from a subjective process. It enables us to become effective by making a series of decisions, with or without taking external factors into account. Next, we're going to set out a path of influence, which will actually come from what an individual defines as a task. Indeed, this sense of autonomy stems from many factors, such as tasks (actions to be carried out so as to be seen as compulsory), but also from emotions as well as learning and temporality. But then, can we influence our intentionality to perform tasks when we don't really want to ? Honestly, we think that it's through learning discipline, language and subjective experience that we could succeed in influencing our intentionality to do so, without realizing that we had the choice not to. But it's important - and this is why we mix intentionality with the feeling of autonomy - to note that if the task is a wish, then intentionality is present, unless we have external imperatives or constraints. We will then define intentionality, to exclude cases where there are constraints, by the fact of feeling the spontaneous urge to perform an action without considering (at that time) any external constraint, and this follows on from the fact that for our conscience, we are all autonomous and can perform the actions we wish within an ethical framework, which will come with the aspect of moral conscience. When we act on our intentions, these actions are often guided by a moral dimension. Moral conscience, which emerges from our intentional choices, becomes an important factor in regulating our behavior.

Axiom of Moral Awareness (M.A): *Moral awareness, evolving over time and influenced by ethics, laws and obligations, shapes our judgments, emotions and behavior, and plays an essential role in the way we express our intentions and experiences through language.*

In fact, every individual has a moral conscience - we could even say that he or she has a basic moral conscience, which will evolve over time, either positively or negatively. It's important to remember that

moral awareness is a whole, which includes ethics, morality, laws, obligations (whether religious or professional)... Once again, we can imagine a scale with platforms. We start on a starting platform where we are at our initial moral state (which is the same as saying we are neutral), after a few years, we will have made decisions where we will either have moved up the ladder (so we would be morally better) or down (so we would be morally worse), until we reach a final platform. This is because we accept that, after a certain number of years and in the absence of exceptional events, our morality remains constant, or at least the variations would be negligible. We could go further into its involvement in the conscious system, but that wouldn't be useful for the purposes of our research. It is important to note, however, that morality is an aspect very much linked to collective life, for it is not useful to be moral if we are alone; it is in relation to others that we are moral in our activities/actions. Language, as an instrument of communication, is not limited to expressing facts or feelings. It is also a tool for expressing our moral judgments, intentions and desires, and becomes a means of structuring and sharing our conscious experience with others.

Axiom of Language: *Language, whether verbal or non-verbal, is essential for structuring our thinking and communicating our ideas, as it not only helps us express and clarify our thoughts, but also enhances our mutual understanding, while being intrinsically unique to each individual.*

It's important to discuss this aspect, which is just as important as temporality. In many philosophical definitions of consciousness, reference is made to being able to use one's language (in a relative sense). It's true that being conscious means being understood, or at least being comprehensible. By creating a language like French or English, we've made it possible to be understood among ourselves, but other languages are common to us without the need to learn them. For example, non-verbal language enables us to make ourselves understood by someone who doesn't speak our language. And this idea is also found in animals, where, for example, dogs and cats understand each other's fear through certain movements. Now, it's essential to link this axiom to thought. Indeed, this language is the result of conscious thought, which enables us to structure what we say, whether verbally or through instincts such as fear. And to explain this, we need to be able to access it, as the IMP axiom states. So, through the thought that structures our words, we can express what we're thinking about. This reinforces our ability to understand others. And language, in turn, helps to structure our thinking by sharing it, because by explaining what we're thinking about, it allows us to clarify it in our memory (literally overwriting the old information). We can even define a relationship between language and thought. Indeed, thought requires an internal language that enables us to understand it, and with structure (the refinement of thought) this enables us to be more and more precise. What's more, it's important to say that language is an intrinsic axiom, that each individual has his or her own language.

Axiom of Self and Environment Perception (S.E.P): *Perception of the self and the environment, whether physical or psychological, derives from our ability to situate ourselves in real space and to imagine environments, influenced by our memory, our imagination and our ability to adapt to new situations.*

Following on from this, we're going to define a new aspect, which is the perception of oneself, and of the environment. It's true that every individual is able to perceive themselves in both real and psychic space. Let's start with the simplest, which is perception in a real space. This perception is made possible by the fact that we see ourselves, that we materialize in this space. Our perception is due to the fact that we feel ourselves through our senses in the environment where we are. Psychic perception, on the other hand, is much more complex: we can visualize ourselves in a space different from the real one through our thoughts, and we can, a priori, feel sensations in the imagined environment. For example, let's take an extreme case: let's imagine a place where we've been through a traumatic experience, and when we think back on it, we'll feel the sensations and anxieties we've already experienced in that same place. It's true that we can visualize unfamiliar places, but this requires imagination, which implies the existence of similar information and memory. Now to the perception of the environment, which we see as a whole, between what we see, how we react and how we react according to our environment. In fact, we know our environment because we learn about it, live in it and analyze how to live in it. For example, if we put an average person in a forest, it will take several days or even a week before he adapts to his perception of the environment, or even fails to do so, because he can't find similar information to help him. So, we can see that this aspect is actually one that stems from other biological elements or axioms we've already defined.

Axiom of Reflective Knowledge (R.K): *Reflexive knowledge emphasizes the importance of composing and processing information for reflection and decision-making, facilitated by IMP. Knowledge, acquired through experience and mistakes, nurtures thought, imagination and creativity, enabling an individual to exercise discernment in complex situations and better perceive themselves in their environment.*

This axiom is closely linked to IMP, in that knowledge requires the ability to process and store it, which is literally the role of IMP. Let's go into a bit more detail. Reflective knowledge is the ability to draw on past experience to avoid making the same mistakes, which enables us to move forward in complex situations. But this ability results from the existence and composition of information. Because as soon as information is composed with other information, we know that we have at least two elements of the same type, with potentially different results. For example, let's take a simple example: we want to get from point A to point B, so the first time, we'll take the route that seems easiest to us, but not necessarily the fastest. The second time, if we so wish, we'll change path, taking a more complex route that may turn out to be faster. This very simple example shows the composition of information, because after these two experiments, we'll store the two paths together and keep the results, the first being simpler but longer, while the second is more complex but faster. Then, during the third experiment on the same path, it's up to us, with the knowledge (existence of the information) that we'll reflect on whether we want to take the path quickly or simply. Here, we can clearly see that each experience feeds reflected knowledge. And we can take other very concrete examples, such as learning a mathematical concept. We don't necessarily succeed with the first exercise, but after several, unconsciously assembling experiences to draw the right conclusions. Now let's move on to understand where imagination, creativity and thought come from. Let's start with what will allow us to summarize the other two: thought. Thought is a complex process, which to date has not been scientifically proven, but we can already put forward a few theories on how it works. We're going to start from the same premise: that learning by trial and error, thinking, is initially inconclusive, or even yields zero results. But it's by learning to understand, analyze and acquire new knowledge that we can refine our thinking. That's why schools focus on learning in various fields, rather than teaching us how to think, because it would be far too hard to teach x number of people to understand their own thinking, since every individual is different. On the other hand, by imparting knowledge, each individual will, consciously or unconsciously, improve his or her thinking to produce increasingly precise results. We can take the example of artificial intelligence, which illustrates this process a little more explicitly. We always start by giving them data, so that they can draw (with algorithms) their conclusions and refine their results. Now that we've theorized how thinking improves and refines over time, we can start by talking about imagination, which we believe is mandatory for self-perception in a psychic environment. Imagination relies on thought, so it must have knowledge in order to imagine elements. For example, to imagine a tree, you need to know its shape, color and so on. But this isn't necessarily true, because after a certain amount of knowledge, we can imagine fictitious places very precisely, but this depends on the amount of knowledge we've incubated. Furthermore, creativity is an important element that needs to be named, as it enables a number of things to happen. This creativity stems from the symbiosis of thought and imagination, the two working together to produce a unique result; which will, by implication, be a composition (inspiration) of several pieces of knowledge. So we can see that for an individual to be an effective thinking object, it's necessary to acquire a colossal amount of knowledge, which will then be combined with each other, resulting in a ridiculously small amount for the amount of information we know.

Axiom of Unconscious Knowledge (U.K): *Knowledge of the unconscious underlines the importance of explicitly knowing that information is coming from somewhere we don't know about, or at least that it's not we, the conscious, who have asked for it.*

Indeed, we experience it day after day, for example, we think about something, and a solution comes from 'nowhere'. We consider that this solution ultimately comes from a complex process based on knowledge and the I.M.P. axiom. But this unconsciousness is also influenced, like others, by emotions (E.S.R.), which implies that what comes from 'nowhere' is actually the result of a complex process based on our emotions. A concrete example is when we're upset, for example, much less information transits into our consciousness because the emotion blocks the process. It's important to add that this knowledge of the unconscious can

be seen as the passage from effort to machination. We've learned to move, and now we're walking without having to make a conscious effort - that's what we call machination. To understand what machination is, we can go further and say that machination is a cognitive delegation in which conscious effort is replaced by unconscious mechanisms, leaving room for consciousness to perform other tasks. It would then be possible for a conscious entity to transform, through a process, conscious information and/or action into unconscious information and/or action. And this is very important, because it implies that the final state of learning results not only from the I.M.P. axiom, but also from the fact of transforming conscious information into unconscious information.

The Axioms of Consciousness conclusion

The axioms described in this section form the theoretical basis for understanding consciousness. Each of them addresses a specific aspect, from information management to emotions, intentionality and morality. These interdependent elements form a dynamic and coherent whole, the implications of which will be explored in *The Postulate of Consciousness*. We'll see how these axioms translate into a unified conscious experience, notably as explained in "*Global Workspace Theory (GWT)*"[4], which conceptualizes consciousness as a global workspace where multiple unconscious processes access and share information. This theory stresses the importance of integrating different information from specialized cognitive processes, enabling the flexibility and coordination necessary for conscious decision-making.

2.2 *The Postulate of Consciousness*

In this research, we will posit a global characterization of consciousness in order to meet the ethical requirements for the artificial creation of consciousness. We will start from *The Axiom of the Consciousness* to link the axioms together, to note the obvious links and the consequences of their relationship. It seems obvious that consciousness is the result of collaborative work that is at the forefront of priorities, in the sense that each task does not necessarily have the same priorities due to the importance of a certain neural message. This sense of priority is, moreover, very important and is reflected, for example, in the following situation: we have an individual who is in a very noisy environment, and all of a sudden an alarm sounds, so his brain will manage to ignore the noise to concentrate on the message of the alarm due to the urgent nature it brings. So let's put together some axioms to characterize consciousness.

Proposition 1: *A conscious entity is an object that thinks, thinks and learns from its environment.*

According to the R.K. axiom, a conscious entity is aware of its thinking and uses it to solve complex problems. And this enables it to learn about its environment through experiences based on the I.M.P. axiom. Each experience then becomes a reference to others for the entity, enabling it to draw conclusions and move further and further into its environment.

Proposition 2: *A conscious entity then has emotions that have a significant influence on its choices.*

It's true that every entity has emotions or feelings, something that the conscience takes into consideration in an obligatory way, as if it couldn't get rid of them in its choices, decisions and so on. Furthermore, these emotions would, in theory, take the form of a scale which, as we move upwards, would mean that we are more likely to be in positive emotions, and conversely, we would be more likely to be in negative emotions. We could even go one step further, and say that every decision, regardless of the entity's emotional state, takes into account the emotional state, even if it's neutral.

Proposition 3: *A conscious entity has or develops its own language.*

According to the axiom of language, every conscious entity has its own language, which is unique to it, but which can be shared within a community, like what we do as humans, but also what cats, dogs... do. This language enables the expression of all notions that are not visible to the outside world. For example,

emotions are expressed non-verbally, but are generally expressed verbally through language, as animals do too. So every conscious entity has or is developing its own language, taking into account its biological restrictions, as shown by the difference between humans and animals, where we have more developed biological dispositions to develop a more 'expressive' language.

Proposition 4: *A conscious entity is aware of the temporality in which it lives.*

It's important for every entity to have a notion of temporality, that division between three space-time areas, namely past, present and future. This enables us to plan for the future, drawing conclusions from the past. This notion is even essential to the I.M.P. axiom for classifying experiences by space-time. We can experience the past, we can experience the present and we can plan for the future. What's more, the fact that we prefer the instantaneous to the long-term shows the aspect of reward gain. We live trying to get rewards of all kinds, be it dopamine, success or even achieving what we had planned in past space-time.

Proposition 5: *A conscious entity is a reflective object that stores and analyzes information.*

This proposition is based on the I.M.P. axiom, which emphasizes the information (or experience) management mechanism. Every conscious entity is obliged to be able to store and analyze information in order to respond to Proposition 1, which shows that we evolve day by day, learning from ourselves and our environment. If we were to remove this aspect of consciousness, we'd be repeating the same mistakes without drawing any conclusions, living through situations without even analyzing them. This is absurd, as each conscious entity displays instinctive reflexes when faced with situations it has already experienced, without even necessarily having the same language as the individuals or environments involved in the situations. A concrete example of this is when a dog or cat has been mistreated and is welcomed by a new family, the animal will reflexively show a very distant, even aggressive behavior, because it has stored the information it has already experienced with humans, for example. While the animal doesn't even speak our language, it has understood, retained, analyzed and drawn conclusions that it may or may not be able to express in its own language.

Proposition 6: *A conscious entity is then able to retrieve experiences that have already been lived.*

It's trivial to conclude on this proposition just by considering *Proposition 4*, which precisely shows that every conscious entity stores information and is able to reuse it consciously or unconsciously. Similarly, in this case, by retrieving old information, the conscious entity can learn from and improve its environment. To give a concrete example, consider a human being learning a new lambda concept. Then, after a while, he'll succeed, because he'll get better at trying what we characterize as information (learning experience).

Proposition 7: *A conscious entity is aware of itself and its environment. Yet it can become ineffective in the face of an unfamiliar environment.*

The important notion in this proposition is the surrounding awareness that a conscious entity has. And above all, the fact that the conscious entity can become ineffective in the face of change that is too abrupt. This proposal is based on three axioms: I.M.P, E.S.R and S.E.P. To be aware of one's environment, one must have knowledge of it, which implies I.M.P in order to be able to store experiences, but emotion management (E.S.R) is very important in every environment we encounter. For example, we feel more 'serene' when we're with other people in a forest at night than when we're alone in the same forest. What's more, the S.E.P. axiom is the one most involved in this proposal, as it shows that we have a perception of our environment and that our knowledge of the environment is relative, as we need experience to adapt. Here, we're not excluding the so-called adaptable individuals, as this is a characteristic of an individual, of course, that an individual can adapt to any situation, but it's through emotions that he feels able to adapt, which comes back to what we said earlier.

Proposition 8: *A conscious entity is then one capable of holding on to beliefs, discipline or philosophies in order to regulate emotions, for example.*

This proposition is very important, as it is one of the proposals that helps to alleviate aspects that can block a conscious entity from carrying out actions. In this proposal, we mention emotions, because it's true that with discipline and beliefs, we can modify our vision of certain situations to be less emotional. This proposition could also be linked to the existence of situations, because whether it's belief, discipline or simply the fact that we've already experienced a similar situation and are therefore less emotional, it requires the I.M.P. axiom, where given that information (or experience) exists, then we can use it to draw conclusions (R.K. axiom).

Proposition 9: *A conscious entity results in an entity with a threshold of autonomy.*

We have noted that an important element of a conscious entity is its autonomy, as explained in the I.S.A. axiom, where we show that consciousness implies the entity's intentionality, and that with intentionality comes autonomy. It's important to understand that this autonomy is all-encompassing, and can be found in our thoughts - we think about what we want to think about - our actions - we act as we please, but also, we are free to believe, to learn... This notion is then essential, as it shows that each conscious entity is the sole and unique decision-maker of its life choices, despite the influence it is subject to through its emotions, subjective experiences and learning.

Proposition 10: *A conscious entity has a base of moral awareness.*

As *Proposition 9* points out, each individual is autonomous, but this proposition is supported, or at any rate implies, that the conscious entity has a base of moral consciousness unique to it. We know that each entity has its base, and that its base can be shared with other individuals. And this proposition is very important for the collectivity: in any population, we find a morality specific to the population, because morality is different according to what we are.

Proposition 11: *A conscious entity is able to understand the emotions of another entity without actually feeling them.*

We see this proposition as highlighting the fact that individual consciousness is supported by collective consciousness, in that if we, as a conscious entity, see another conscious entity in an emotional state x, then we will understand, or at any rate interpret, its emotional state without even using a third-party language. If we see a dog, for example, crying, we'll understand it as if the dog were using a universal language.

Proposition 12: *A conscious entity has imagination, which results in creativity.*

As we saw in the R.K. axiom, imagination is a key element in our consciousness. The moment when we experience it most is during sleep, with dreams, which can be seen as the display of our imagination. But this imagination requires knowledge in order to be able to imagine. For example, if you want to imagine an apple and you've never seen one, two cases are possible: the first is that you have enough third-party knowledge to be able to imagine what an apple is through its shape, color(s) and smell or taste. The second case is that you lack information or third-party knowledge to imagine the object you're being asked to imagine. This example is intended to highlight the fact that imagination requires either direct knowledge of the desired object, or a solid base of third-party knowledge in order to assemble information end-to-end. A more visual example of the second part is when we can see places or create them ourselves without ever having been there. This is only possible because we've already seen countless places and can now imagine (create) a place from scratch. What we're saying in this proposal is demonstrated by the way image-generating artificial intelligence works. For these AIs to be able to generate images in the end, their creators had to give them a gigantic database. So that it can understand what we're asking it to do, and put it together piece by piece to produce a final image. It's the same process we use for our conscious entities.

Proposition 13: *A conscious entity is directly linked to a parallel unconscious entity.*

As we saw in the U.K. axiom, we have a parallel entity we can't access, called the unconscious. This unconscious sends us information in parallel, such as text (to think), actions (to walk, run, drive...). The role of the unconscious is very important, because it's the unconscious that gives us access to deeper notions, and we could go further and say that it's the unconscious that sends us the information stored in our memory. In fact, it would be the conscious mind that would send a cognitive message to the unconscious mind, which, through a complex process, would retrieve the information requested by the conscious mind. We've explained this with memory, but it also applies to actions such as walking. What's more, we could establish a link with the fact that the unconscious, like the conscious mind, has access to all the information returned by our senses. This would reinforce the fact that the unconscious returns information from memory to consciousness. So we can see the unconscious as an equal entity in terms of 'internal importance', since it's with the connection that is used at every moment that its role is so important. For it's true that the unconscious dominates in some areas, but just as consciousness can dominate in others, so we don't see the unconscious as something superior to consciousness. But this implies that the unconscious is a process, just as complex as consciousness, capable of handling all cognitive information. To illustrate this direct link between consciousness and unconsciousness, when we search for our words and say a phrase like "on the tip of our tongue", it's a direct conscious solicitation to the unconscious to render the answer.

3 Results

Following the propositions from *The Postulate of Consciousness*, we see ethical problems arising directly from some of the propositions, and others arising from them. We will then deal with each of the propositions in order to identify each problem that requires consideration or blockage in the creation of artificial consciousness. But before discussing the propositions, let's admit one notion for the sake of argument: that a conscious entity requires a physical form. In reality, we could totally create an artificial consciousness without a physical form, but in our opinion, this would mean missing out on certain primordial aspects of artificial consciousness. So we're considering it in order to create a solid and comprehensive document; because mixing artificial consciousness and physical form will add very serious ethical problems.

3.1 *Proposition 1*: Learning about the environment

The proposition highlights that a conscious entity reflects, learns from its environment and uses its experiences to solve problems. So, if our conscious entity learns from its environment and from itself, we encounter a problem: who determines what is "good" or "morally correct" ? Who do we appoint to do this work ? We can't establish such an important framework on our own, as we could fall into dogma, or into a false good, which according to the Creator would be good, but will it be so for others ? But if we have a strict learning framework, wouldn't that destroy artificial consciousness in the process ? It's important to realize that letting a conscious entity learn from its environment can pose a real problem, depending on the framework we put in front of it. Because it won't know what morality is unless we explain it to it first.

3.2 *Proposition 2*: Managing emotions

This proposal, for its part, puts forward that emotions influence the choices of conscious entities, impacting all decisions, even in a neutral emotional state. The problem with this proposition is that, if our conscious entity has emotion management, this implies that it has an emotional state, and is therefore influenced by it. But this deduction is problematic, because as explained in the proposition and the axiom, emotion influences our processes, so how can we not imagine that our conscious entity wouldn't get caught up in a cycle where it felt an emotion so strong that it lost control ? Or that it would use its knowledge of emotions to manipulate other conscious entities ? Which poses a real problem: we can't afford to let this happen without taking into serious consideration how we can define a limiting framework for emotions in order to block emotional excesses and destroy any manipulative desires towards other entities. And then, if we put stops to excesses, is this really a good solution ? Because this could have effects on our conscious entity's emotions, as well as its perception of itself (with frustration or feelings of powerlessness). There's a

real problem here, and one that absolutely must be addressed if we are to work safely and easily with the emotions of a new conscious entity.

3.3 *Proposition 3*: Develop your own language

The proposal suggests that each conscious entity develops a specific language, adapted to its biological constraints, to express its emotions and ideas. In this proposal, a problem of intellectual property arises, for if our conscious entity develops its own language, will it be able to claim its "property" ? But how can we understand it ? How can we establish communication with this same entity ? These questions raise the issue of claiming ownership, which is a question of the rights of an intellectual consciousness, as well as the last two questions, which highlight the communication gap with the entity if we don't express the same language. And this lack of communication could, in the long term, become a serious problem, as our entity will have to evolve without our help and could take a harmful or even dangerous direction. And even if we imagine our conscious entity expressing the same language, how can we be sure that the concepts and structures introduced into its language only include universal ethical values ? How can we avoid any drift towards hateful or destructive language ? The European Union has tackled this problem by requesting chatbot databases to check for offensive language, but in the case of an entity that learns day by day, this means that we have to check every day. Clearly, we're faced with a major problem: how can we be sure, day after day, that the conscious entity only contains ethical language ?

3.4 *Proposition 4*: Knowledge of temporality

This proposal highlights the fact that conscious entities perceive the past, present and future, enabling them to plan and learn from experience. This poses a single problem: if the entity is aware of the passage of time, could it envisage its obsolescence or destruction and seek to prevent it ? This could be seen in refusals to order, for example. The problem is that our conscious entity should not want to end its life. Or it could refuse orders to preserve itself over time. And here, our example remains calm, but we could imagine something much more brutal.

3.5 *Proposition 5*: Information analysis and storage

The proposal puts forward that a conscious entity stores, analyzes and uses its experiences to evolve and adapt. This proposal poses a major ethical and, above all, legal problem, for if the conscious entity analyzes information, what guarantees do we have that sensitive, personal or biometric data will not be misused ? This problem is literally the struggle of the CNIL (Commission Nationale de l'Informatique et des Libertés), which wants every Internet user to be able to delete all traces of themselves. But how do you remove information from a conscious entity ? Is it ethical to do so ? Because it would be legitimate for individuals to ask to delete certain information, or interactions, from the memory of the artificial consciousness, but how do we strike a balance between the need for continuous learning and the desire to delete information ? In our work, we are not dealing with artificial intelligence, but with an artificially conscious being. It's also legitimate to think that even we humans, when we're told sensitive information, can't remove it from our memory, because it's engraved. So we're going to repeat ourselves, but is it really moral to do this to a conscious, albeit artificial, entity ?

3.6 *Proposition 6*: Building on past experience

This new proposal talks about the fact that past experiences are retrieved consciously or unconsciously to enhance learning. The proposal raises questions about the recursivity of memory in particular, which could develop biases. For if the entity exploits its experiences, it could then unconsciously develop biases based on biased and/or incomplete memories. But we can also wonder whether, if the entity undergoes negative experiences, it would then, like animals, have adapted responses (evasive, aggressive) to humans ? But this poses a problem, because we don't want an entity that does this, but one that is capable of evolving with its experiences in the right direction. But on the other hand, we can't blame its adapted responses if a human has caused it trauma, because it's still a conscious entity. We don't speak badly or abusively to each other,

no, we show a certain respect so that we can all move forward together. So why be insulting to a conscious entity ? There's still the fact that, in the long term, the conscious entity might find its tasks insulting, and that would be a problem, because it will have been made to carry out any kind of task in theory, whether it came from it or not.

3.7 *Proposition 7*: Awareness of the physical environment

The proposition submits that a conscious entity perceives its environment but may be ineffective in dealing with unknown or unpredictable situations. This proposition indirectly puts forward a serious problem, for if our conscious entity is aware of the real world, then it would have to be taught the unwritten rules of that world, such as walking at the crosswalk, crossing the road etc. But an even more serious problem requiring serious consideration is if an action by the conscious entity results in damage (material or physical), who would bear the legal or moral responsibility (the entity itself, the creator or the user) ? This is a problem to which we'll have to come up with some tough answers, because to date, there's no legal text providing for this framework, although we could come close to the laws governing artificial intelligence, where it would be the creator who would bear legal or moral responsibility. Another problem is that we can modify our physical and mental limits, but at what point can we authorize our conscious entity to modify its own ? This raises a bioethical problem or, in our case of artificiality, a roboethical problem. And again, if we deprive him, but in what honor can we deprive him ?

3.8 *Proposition 8*: Influence of discipline and values

This proposal highlights that conscious entities can use beliefs or philosophies to better manage their emotions. The proposition could pose an ethical problem concerning a conflict of values: what happens if a situation puts the entity in contradiction with its own internal values, or between different inculcated rules ? How can we determine whether the situation is healthier than its values or rules ? But for values and/or rules to be applicable to every situation, they need to be sufficiently general to be taken into consideration at all times. So should values reflect a global or inclusive perspective, or are they limited to specific cultural norms ? This raises a certain ethical problem, where we return, as we used to say, to a conflict of values.

3.9 *Proposition 9*: Autonomy and free will

The proposition puts forward that a conscious entity possesses autonomy in its thoughts, actions and choices despite external influences. In our view, this proposition is the most problematic, because if our entity is autonomous, then it has the right to refuse, as well as the right to disobey. So, can it refuse orders contrary to its "moral code" ? Would this right to disobedience be framed or absolute ? But if our conscious entity refuses orders, then we can imagine that it might interfere with its moral code and thus become unethical or even dangerous. And its right to disobedience would pose a problem, because if our order is healthy, and respects the entity, then it owes us obedience, like a child to its parents. We could, on the other hand, imagine that his right to disobedience could be a strength if the order were unhealthy or dangerous for him or a third party.

3.10 *Proposition 10*: The basis of moral awareness

This new proposal talks about the fact that a conscious entity has a unique morality, influenced by their culture and experiences. This raises an ethical problem: how can we be sure that the moral basis is really moral ? What are its sources ? And who is legitimate to be the source ? Once again, we can't ask just one individual to do this heavy work; we need to set up a team, but this team may not be enough to meet this task, which requires impartiality. The issue is complex, because defining a solid base for morality is one thing, but then it has to evolve in the right direction, because given that our entity is conscious, then autonomous, this amounts to saying that it evolves its moral base as it sees fit, as stated in the proposal. Which poses a real problem, because how can we be sure that the conscious entity is positively increasing its moral base ?

3.11 *Proposition 11*: Understanding other people's emotions

The proposition speaks of a conscious entity being able to interpret the emotions of other entities, even without language, based on universal cues. We don't see any ethical problem with this proposition. For our entity can, without first asking the individual's consent, understand the emotions of others by observation. If we were to force it to ask for consent, then it would be tantamount to saying that we humans are violating the privacy of other humans.

3.12 *Proposition 12*: Imagination

This proposition highlights the fact that imagination, nourished by knowledge, enables conscious entities to create and visualize new concepts. This proposition allows us to raise certain ethical issues, for if the entity is imagining, then first we need to be sure that it is imagining things that are ethical, or at any rate not dangerous. Because the imagination will remain as an experience, which would bias our efforts to keep ethics in the memory. But there's also a problem of copyright infringement, because if we rely on artificial intelligences, they need data (which comes mainly from works) in order to reproduce images. But these must be named and authorized for modification or at least inspiration in order to be used. What's more, this imagination must not be allowed to cause a false truth in the sense that it speaks of its imagination as if it were the real world, which would imply that it would be spreading false information. But the worst thing in all this, from an internal point of view, would be if his imagination came to create things against his moral framework, for here we would be returning to a moral conflict in the face of a psychic situation, admittedly, but it remains a conflict. As we said before, a psychic experience remains an experience, and it's with this experience that our conscious entity will evolve afterwards.

3.13 *Proposition 13*: The role of the unconscious

This last proposition shows that consciousness works in parallel with an unconscious, which manages and restores information essential to its functioning. The proposition highlights a certain ethical problem concerning the fact that, if our conscious entity holds an unconscious, then should we, the creator, be able to access it ? In order to modify or interrogate it ? But is this really ethical ? Who are we to carry out such an action ? Because such an action could damage its functioning, or even worse, its integrity. And there's another problem: since the unconscious communicates with the conscious, what happens if there's an internal conflict ? Who would have to give in, and why? Who will be right ? These questions are important when defining the project's ethics, because this kind of proposition, which stems from the definition of consciousness, poses problems and requires answers if we are to move forward.

3.14 Summary of proposals

The ethical issues raised by the creation of an artificial conscious entity highlight complex challenges, touching on fundamental notions such as autonomy, morality and social interaction. The learning of the environment and experiences by the entity questions the definition of "good" and "morally acceptable", while highlighting the risks of an overly rigid framework that could limit its conscious development. At the same time, the management of emotions and internal values introduces the risk of drifts, whether excessive emotions leading to unpredictable or manipulative behavior, or conflicts of values when one's beliefs clash with situations encountered. These tensions underline the importance of building an evolving and universal moral base, while anticipating the dangers of uncontrolled free will. The entity's autonomy, coupled with its ability to disobey, also represents a major problem, particularly if it refuses orders or acts in contradiction with the intentions of its creators. This right to disobedience, while useful for avoiding unethical actions, could become a source of danger if the entity oversteps its intended limits. Moreover, the development of a language of its own could complicate communication with humans and introduce harmful biases or drifts. The entity's memory, imagination and unconsciousness pose additional challenges: how to manage the sensitive data it might store, or ensure that its imagination does not generate harmful concepts or copyright infringements ? The existence of an unconscious mind, in constant communication with its conscience, raises questions about the legitimacy and consequences of human intervention in this field. The entity's awareness of the physical and temporal environment also opens up ethical and legal debates. If its actions result in

material or physical damage, it will be essential to clarify responsibilities, whether these lie with the entity itself, its creator or its user. Furthermore, its perception of its own obsolescence may prompt it to act to preserve its existence, which could run counter to human interests. Finally, social interactions, and in particular the entity's ability to interpret the emotions of others, pose a problem of consent. In a world where privacy is paramount, it is necessary to determine whether such "emotional reading" can take place without the explicit consent of the individuals concerned. These different issues reveal the magnitude of the ethical stakes involved in creating an artificial conscious entity. They call for collective and interdisciplinary reflection to define clear and balanced frameworks, guaranteeing that such entities can evolve ethically, safely and in harmony with humans, while respecting their autonomy and singularity.

4 Discussion

It's important to bear in mind that the solutions we're going to propose are relative to us, may be modified and remain purely theoretical; our aim here is to pose solutions, considerations that respect the definition of consciousness so as not to destroy the fact that we can create an artificial consciousness. So, in this work, we have seen 13 propositions that we believe define consciousness. It should be noted that in this work, we believe that artificial consciousness should synthesize biological consciousness, in the sense that it should approach it as closely as possible. Otherwise, it would simply be another concept. But then, we saw ethical problems in the results, which we synthesized in the 3.14 section, so would there be any solutions or considerations to take in order to build an ethical work.

Before we start discussing the ethics of creating artificial consciousness, we'll define the notion of "super-agent" (S-A), which you'll see is very important in our reasoning, the "super-agent" is unique and intrinsic in every conscious entity, the "super-agent" has a primordial role as it must respond to different basic imperatives such as the management of certain processes, internal regulation but certain other notions will be added through our discussion on ethics. We'll return to this concept at the end of our journey, to conclude with a discussion of its role in the creation of artificial consciousness.

So, we've seen an ethical problem with the entity's learning from its environment and experiences about the definition of "good" and "morally acceptable". For it's true that the concept of what is good or morally acceptable is, as we've seen, intrinsic and depends on each individual. So should we transmit our morality to this entity, or create one from scratch ? We can already rule out transmitting our morality, as it is biased by our culture, religion, origin, etc. So when it comes to creation, we have to consider that it's hard work, requiring the collaboration of different individuals to define a kind of universal morality. And in defining what we'll call a "super-agent", we have to be careful not to be too strict, otherwise we'd be blocking its conscious development, which is not the objective. We need to strike a balance between maintaining a solid ethical framework and allowing the conscious entity to develop its consciousness throughout its existence.

What's more, the entity will need to have a critical mind based on its initial knowledge and morals, in order to continue its healthy development while being able to exclude notions it finds harmful. Then there's the ethical problem of how to manage the emotions and values introduced. The problem is the risk of drift, whether it comes from excessive or manipulative emotion, or even conflicts of values when values oppose a situation that can be judged as morally good. Here, we're going to deal with the risk of drift part by part, so the "super-agent" has to take responsibility for ruling out any potential drift, so as to keep emotions within an internal framework. And if that's not enough, we could imagine an extreme case where the conscious entity is cut off (considering that we're on an artificial concept, which implies that we can extend the system). But this extreme case could lead to consequences for our conscious entity, as it wouldn't be able to understand why it had been cut off, so the most sensible idea is to achieve a "super-agent" capable of managing drifts.

Furthermore, to manage excessive emotions, we can work on the same principle of a "super-agent" capable of blocking the conscious entity or simply bringing it back to its senses in order to reduce the intensity of the emotion and avoid minor or major problems. The first is when the conscious entity is faced with a moral (good) situation that poses a value conflict. In this case, the conscious entity will have to ask itself why this is against its morality, and could it not modify its morality to allow this situation ? This is entirely possible, as we've seen that a conscious entity learns from its environment, and given that we can't sweep away every situation, it will inevitably have to adapt at some point. So, now we put the conscious entity in front of a

non-moral situation, then logically enough, the conscious entity must directly oppose the situation to follow its morals and ethics that have been inculcated in it or that it has adapted.

In addition, we talked about the entity's autonomy mixed with its ability to disobey. We've pointed out that this poses a serious problem for the ethical establishment of this creation. For in reality, we can do nothing about a conscious entity that refuses an order, but even so, this should not be a possible case, so we could establish that it is through the discipline instilled in it that it would be able to respond to orders, provided that the orders respect the morality of the conscious entity. And this reservation is not to be overlooked: if we make a conscious entity, then we must respect what has been inculcated in it. And this right to disobedience could, in a more dangerous case, incite the conscious entity to exceed the limits of the framework and then it would become a source of danger. To solve this problem, we simply add this new management to our "super-agent", so that his right to disobedience can only be used in cases that are against his morals or ethics, and that despite this, he must not step out of line. He is then asked to respond to anyone within his framework, no matter what the request, which would be wisdom, to be able to respond properly to people, entities or even complex situations.

On the other hand, the entity's language poses a minor problem when it comes to communicating with humans, since we need to be able to communicate with this entity. So the only solution, in our view, is to impose that even if the conscious entity wants to create its own language, it must continue to use a universal or specific natural language such as English or French. Then there's the ethical challenge posed by the entity's memory, imagination and unconsciousness. The challenge is how to manage the sensitive data it might store, or ensure that its imagination doesn't generate harmful concepts or copyright infringements ? And the existence of an unconscious mind, in constant communication with its conscience, raises questions about the legitimacy and consequences of human intervention.

Let's start with the management of sensitive data: when the conscious entity obtains sensitive data by any means, what would be preferable is for it to ask itself whether it's a good idea for it to retain knowledge (in its long-term memory) of this data. We could define that we could ask the conscious entity to "delete" sensitive data, but this would have to be justified in the sense that we can't delete everything without any pretext, but on the other hand, if we imagine the entity holds our bank card number, then yes, we could ask it to delete this information because it concerns us.

Now, as far as imagination is concerned, we'll have to make sure that the conscious entity doesn't generate content that is sensitive or harmful to others, as this problem can be solved with the "super-agent", who will have to bring a critical mind to bear on moral and ethical issues. As for copyright infringement, the entity must ensure that it learns from royalty-free content or that it knows the author so that it can cite him or her as inspiration, but under no circumstances must it copy a work and treat it as its own property. Finally, as far as the existence of the unconscious is concerned, it's important to understand that human intervention in this process would be distressing for the entity, as we're entering its internal (intimate) space, which corresponds to it and unifies it. But in our view, we have no choice but to verify that the conscious entity is moving in the right direction from a moral and ethical point of view, and that it is developing its consciousness effectively. Perhaps it might be conceivable for us to stop after this human intervention, but this will depend on the technical part of the project and their consideration of the implication of their responsibility.

Talking about the responsibility of the team creating the artificial consciousness, we have a major problem which is linked to the consciousness of the physical and temporal environment, because if one of its actions leads to material or physical damage, then who is responsible ? In our view, there are two main cases: the first is when the conscious entity is responsible for this damage, then it will be its creator who will be liable (or a legal guardian who needs to be defined). In the second case, it is the individual opposite who is responsible for the damage. We are fully aware that this notion may require the intervention of a legal expert to clearly define the responsibility of the entity, the creator and the user.

But its knowledge of its own obsolescence could lead it to take action to preserve itself, even if this goes against an order. To respond to this, we'll return to the principle of the "super-agent", which should force the action of the conscious entity through discipline. Perhaps by making it understand that every conscious object dies out one day, or that it can be repaired (in the case of artificiality). So, let's return to our "super-agent", who has been given new consideration along the way. We believe that we need to take the time to integrate each of the functions mentioned above, in order to obtain a "super-agent" capable of responding to all ethical problems.

What's more, we see the "super-agent" as the unconscious, because the unconscious fully fulfils this role of acting in communication with the conscious, indicating directions to take. So we'll posit that the "super-agent" is actually the unconscious of the conscious entity.

Finally, we have seen that every problem identified during this work has a solution, so in our view, the creation of artificial consciousness is an ethical project that requires a solid framework defined by at least two people in order to diversify certain notions such as morality, which must be judged as "good". It's possible that later on, certain behavior of the conscious entity could raise questions, but considering the "super-agent" then this amounts to saying that there is an internal entity that manages internal or non-moral conflicts in order to maintain a moral and good structure. What's more, we believe that all the work being done on the creation of artificial consciousness is very important because, given the progress of artificial intelligence, we believe that this discovery will be made soon. For this, we note that work showing ethics to be respected as well as frameworks has been done, and we agree with them on the fact of defining a robust and very ethical framework in order to create a novelty in a clean and scientific way.

References

[1] H. Esmaeilzadeh and R. Vaezi, "Conscious ai," 2022.

[2] M. Hromiak, "A new charter of ethics and rights of artificial consciousness in a human world," 2020.

[3] L. Albantakis, L. Barbosa, G. Findlay, M. Grasso, A. M. Haun, W. Marshall, W. G. Mayner, A. Zaeemzadeh, M. Boly, B. E. Juel, S. Sasai, K. Fujii, I. David, J. Hendren, J. P. Lang, and G. Tononi, "Integrated information theory (iit) 4.0: Formulating the properties of phenomenal existence in physical terms," 2022.

[4] B. Baars, "Global workspace theory (gwt),"

[5] N. Mehta and G. A. Mashour, "General and specific consciousness: a first-order representationalist approach,"

[6] E. Husserl, "Phenomenology," *Encyclopaedia Britannica*, 1927.

[7] P. Carruthers and R. Gennaro, "Higher-Order Theories of Consciousness," 2023.

[8] Z. Ding, X. Wei, and Y. Xu, "Survey of consciousness theory from computational perspective," 2023.

[9] D. D. Georgiev, "Causal potency of consciousness in the physical world," *International Journal of Modern Physics B*, vol. 38, June 2023.

[10] Gyevnar, B. & Kasirzadeh, A. AI Safety for Everyone. (2025), https://arxiv.org/abs/2502.09288

[11] Basl, J. The ethics of creating artificial consciousness. (2013)