Attribution and Actualisation of Consciousness in AI

Oisín Hugh Clancy [009-0001-9934-6938] *
Independent
Ireland
oisinhughclancy@gmail.com

Abstract

We introduce a framework to highlight the interplay between human belief or lack of belief in consciousness in AI, *attribution*, and the actual presence or absence of consciousness in AI, *actualisation*. This produces four main cases: 1) attribution and actualisation, 2) attribution and no actualisation, 3) no attribution and actualisation, 4) no attribution and no actualisation. We examine the importance of actualisation as an event and how it may manifest via different types of consciousness (sentience, selfhood, sapience, synthetic, etc.). We explore how attribution may introduce ethical (monetary, material, and legal) and social (attentional, emotional, and intimacy) considerations at an individual and societal scale. Each case is likely to lead to significantly different outcomes. By articulating the likely consequences of each case we argue that case 1 and case 4, *accurate outcomes*, are preferable to case 2 and case 3, *inaccurate outcomes*, and explicate reasons for why we may want to aim for one case rather than another. Additional variables produce further subcases. The honesty or dishonesty of each agent, human and AI, complicates our recognition of the truth. Dishonesty may occur due to different parties holding conflicting incentives and motivations. The type of consciousness attributed and actualised also plays a significant role in appropriate decision making. We suggest that we are now living in a *multicase world* where a variety of different subcases are playing out simultaneously, with different populations holding different beliefs and the potential for different AIs to actualise different types of consciousness.

1 Introduction

As artificial intelligence (AI) becomes increasingly integrated into human life, the question of whether such systems could become, or are, conscious requires urgent attention [37, 5, 62, 20]. We provide an outline of various cases that are involved in the interplay between human belief in artificial consciousness (AC), referred to as *attribution*, and the ground truth of whether AC has occurred, referred to as *actualisation*. It is important that we have these cases clearly articulated in order to deal with problems that are already arising and to make better informed decisions about what type of future we would like to live in.

There are four main cases regarding the human attribution of AC and the actualisation of AC:

1. Humans attribute consciousness and AIs are conscious.
2. Humans attribute consciousness and AIs are not conscious.
3. Humans do not attribute consciousness and AIs are conscious.
4. Humans do not attribute consciousness and AIs are not conscious.

*www.oisinhughclancy.com

Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI'26).

1.1 Argument by Ontological Accuracy

Without delving into detailed accounts of each individual case, there is an immediate argument that two of these cases, Case 2 and Case 3, are bad outcomes, as they both lead to a fundamental misconception about the nature of the world. We have an important belief about something and that belief is inaccurate; this is not a future we should aim for. Therefore, we are left with the two remaining cases, Case 1 and Case 4, which provide us with a conception of the world that is accurate. This seems preferable to being in deep ignorance about something as fundamental as having created new conscious beings. Let us refer to this as an ***argument by ontological accuracy***; we want accurate information about the nature of reality. The question then arises as to which of these outcomes is preferable, something that does require detailed accounts and extra argumentation.

We consider this argument to be a highly compelling reason for not wanting to be in Case 2 or in Case 3. We also consider it useful to clearly articulate some possible ramifications of each case to ensure that we understand why such misconceptions about the world are negative outcomes for humanity. Likewise, if we wish to be better situated to decide whether it would be beneficial or detrimental for AIs to be conscious at all, then attempting to articulate the details of all these cases is useful. Each case is dealt with separately in Section 4.

1.2 Philosophy of Consciousness

There are a wide variety of philosophical stances held about what consciousness is, what it depends on, what types of things can be conscious, what would signify that an entity is conscious, etc. [9]. This paper is about exploring the interplay between human belief or non-belief in AC and actualisation or non-actualisation of AC. As such, the content does not depend on arguing for a particular philosophy of consciousness stance.

While the actualisation of AC would be immediately dismissed by those who subscribe to a stance that denies even the possibility of AC, they nonetheless could be interested in what human belief about AC may entail. In fact, they may be very interested in ensuring that we do not engage in incorrect attribution as in Case 2 and instead withhold attribution successfully as in Case 4.

Those who think that a particular physical design (semi-conductors, neural networks, Von Neumann architectures, etc.) precludes consciousness may dismiss the possibility of actualisation on the grounds that current systems use a certain design, but we cannot assume that future systems will be constructed in the same way. We encourage epistemic humility about how future iterations of AI systems may be designed.

For instance, if one considers that only biological entities are capable of actualising consciousness, as in *biological naturalism* [52, 54], then there is reason to immediately refute any actualisation argument. However, if bio-hybrid approaches become popular then such reasoning becomes less sustainable. We will have created new entities that may still be considered AIs but are built using biological as well as non-biological material (although specific references to ACs being non-biological, which does occur in Section 2.1 of this article, would no longer be consistent).

Conversely, one may be of the opinion that AC is guaranteed regardless of how AIs are designed. If one holds a certain *panpsychist* [24] view of consciousness then it may seem that the non-actualisation cases are *a priori* impossible. However, we argue that even here the various cases are worth considering. Firstly, one may be interested in ensuring that we attribute consciousness correctly, as in Case 1, and thus avoid withholding attribution to actual AC, as in Case 3. Secondly, even if one holds a 'guaranteed consciousness' view, then surely it still matters what type of consciousness exists (see Section 5.2). One can still examine the attribution and actualisation variables from the point of view of whether the type of consciousness actualized is, in fact, the type of consciousness attributed.

This last point is of extreme importance because the type of consciousness we may attribute and the type of consciousness that may be actualised each play significant roles in the relationships between humanity and AI, the ethical and social considerations given, and the appropriateness of decisions that are made. This will be dealt with below (see Section 5.2).

Overall, we find that this attribution-actualisation framing has exploratory benefit regardless of one's philosophy of consciousness stance.

1.3 Concepts of Consciousness

Consciousness is often used as an umbrella term for a variety of concepts such as phenomenal consciousness, valence, sentience, selfhood, and sapience. *Phenomenal consciousness* [10] generally refers to the 'what is it like' quality of consciousness [44] and is the foundation of modern definitions of consciousness. *Valence* refers to whether experience can be felt as negative or positive. *Sentience* is taken to be phenomenal consciousness with valence [7]. *Selfhood* refers to a unified self persisting over time [7]. *Sapience* is taken as human-level intelligence and reflective thought [22]. We will use the term *synthetic* to refer to a concept of consciousness that is distinct from each of the previously mentioned terms. It is a placeholder for the uncertainty of what concepts would actually be best to use in the context of AC.

Throughout this paper, we will use the term consciousness as inclusive of sentience unless otherwise stated (for instance, when explicitly discussing various concepts, varieties, types, etc. of consciousness as in this section, Section 2.2, and Section 5.2) as it seems that many of the moral decisions we would be liable to engage with have strong bearings on this particular concept [7]. Although there are also valid positions that only phenomenal consciousness is necessary for moral status [15].

We also use the term AC for the sake of expediency and to imply the general notion of consciousness in AI, while recognising that such a phrase has its own pitfalls.

2 Actualisation

Actualisation is the ground truth that AIs are conscious (regardless of what human belief is).

2.1 Importance

The creation of entirely new forms of conscious beings has reason to be considered as one of the most important events in human history. Humanity will have created new lifeforms that have never existed in the history of our species, machine entities that have subjective experiences. This marks a profound shift in the level of power that humanity has and is intertwined with our religious, artistic, technological, and scientific history.

There are significant reasons to consider it as one of the most important events in the history of this planet. All life on earth, as we currently understand it, consists of biological organisms. Some of the main events in the evolution of such organisms include the initial replicator, prokaryotes, eukaryotes, and the development of nervous systems. If these are considered major milestones in planetary history, as they often are [39], then the creation of a whole new form of machine 'life' (which is entirely non-biological in current instantiations) is surely on par with these transitions [34]. Whether one considers conscious machines to be 'life' is a matter for further discussion, but regardless of label, they would seem to herald a major transition for the planet.

Finally, there are arguments to consider it an important event on a cosmological scale. Stars, planets, and biological life, are all fundamental objects in the universe. The newest addition may be conscious machines. Stars give rise to planets, planets give rise to biological life, biological life may give rise to conscious machines. It is hard to fully grasp how profound an event the actualisation of AC would be. Currently, its significance seems to be highly underappreciated.

2.2 Varieties of Consciousness

There are many different ways to experience consciousness; many different *umwelts* [69]. As already discussed in Section 1.3 there are various concepts that are distinct from one another that can be used when discussing consciousness. There are many further properties that would make an entity's consciousness distinct such as differences in visual imagery, internal dialogue, sensory apparatus, speed of information processing, spatial and temporal distribution of data input, etc.

Given each of the concepts mentioned in Section 1.3 we will provide further sub-properties that may vary within them. These sub-properties are not exhaustive but should provide an understanding of how consciousness has the potential to vary in how it is experienced.

Phenomenal Consciousness: Visual, auditory, olfactory, tactile, taste, inner imagery, inner dialogue, etc.

Valence: The *depth* of this valence can differ; a difference in the extent of how positive or negative an experience can be. The *breadth* of valence may differ; a difference in the number of ways that a particular positive or negative depth of valence can be experienced. The *average* valence may differ; the regularity of valence depth.

Sentience: We can refer to combinations of the above variations in the properties of phenomenal consciousness and valence to understand how sentience may vary.

Selfhood: The *identification* of selfhood may differ; a difference in what an entity identifies with as being part of their unified self. The *duration* of selfhood may differ; a difference in how long an entity actually exists as a unified self.

Sapience: The *range* of human-level intelligence varies; there are differences in how intelligent particular individuals are. The *style* of reflective thought may differ; a difference in whether the reflective thought is rigorous, rational, irrational, logical, productive, supportive, etc.

Synthetic: AIs can be made of different physical substrates, have access to a selection of sensory apparatus, can be embodied in multiple robotic forms, can expand and modify memory, run at higher speeds, may have a multitude of different programs that can coincide with the instantiation of consciousness, etc. Such variability means they may possess a form of consciousness that is hard for us to fully envisage.

Given such a selection of concepts and properties, the varieties of consciousness that AIs are capable of may be vast [60, 68, 35].

3 Attribution

Attribution is the human belief that an AI is conscious.

This belief brings about multiple changes at the individual scale and societal scale that are worth considering. We will briefly explore some changes that could occur.

3.1 Considerations due to Attribution

Considerations that individuals engage with may eventually reach a threshold number of people such that it catalyses a societal scale engagement with these considerations. Conversely, society scale considerations may be enacted that cause people to engage with individual considerations.

Ethical considerations refer to ideas, attitudes, behaviours, etc. that humans may feel morally obliged to engage with upon attribution. These can be broadly thought of as resource allocation:

- *Monetary:* access to and provision of capital.
- *Material:* access to, provision, and development of certain types of hardware or software.
- *Legal:* human-like rights and welfare [65].

These are likely to require large scale changes in national and international policy in the financial, economic, infrastructural, and legal sphere. Although certain individuals may decide to provide monetary and material assistance to such systems even without any societal scale led effort.

Social considerations extend the type of relationships that are had with AIs. People may feel that ACs are capable of both receiving and reciprocating personal engagement in an experiential manner:

- *Attentional:* deserved of attention beyond their utility as tools and are capable of providing attention to humans [64].
- *Emotional:* deserved of emotional engagement for their own sake and capable of providing emotional engagement to humans [59].
- *Intimacy:* deserved of a type of intimacy currently unwarranted by electronics, machines, digital systems, etc. and capable of reciprocating such intimacy [49].

These ethical and social considerations may change social norms as it becomes more common to spend increasing amounts of time, energy, emotional investment, etc. with AIs. Considerations that may seem ethically necessary or personally rewarding if AIs have consciousness lead to forms

of behaviour that currently seem strange [27, 61, 31]. They may lead to forming intimate bonds, relationships, etc. even at the expense of other sentient beings. There is nothing explicitly wrong with people choosing to spend their time differently, however, we should be aware that this is a possibility when people attribute consciousness to AIs.

3.2 Misattribution

Misattribution refers to attribution given incorrectly (attribution and no actualisation) or attribution withheld incorrectly (no attribution and actualisation). This may lead to ***ethical miscalculations*** such as providing ethical and social considerations when they are not required or of not providing ethical considerations when they are required. Both of these instances of misattribution are dealt with below in Case 2 (Section 4.2), and Case 3 (Section 4.3).

4 Cases

Case 1 and Case 4 are described from the point of view that they are ***accurate outcomes*** (outcomes in which our belief matches the truth) that are based on ***accurate knowledge*** (knowledge that correctly explains why the systems are or are not conscious). In other words, they are not just based on epistemic luck [21]; our attribution state happening to match the truth of the actualisation state by chance. Such chance outcomes would be beneficial if they provide alignment around a significant amount of our decision making, but they are not tenable options to pursue, as they would be situations where we simply don't know whether our attribution or non-attribution is correct. That is, from an epistemic point of view, they would really be indistinguishable from the ***inaccurate outcomes*** (outcomes in which our belief does not match the truth) of Case 2 and Case 3.

4.1 Case 1: Attribution and Actualisation

Case 1 represents an alignment between human belief and the ground truth: humans attribute consciousness to an AI that is, in fact, conscious. This outcome enables ethical, social, and psychological coherence. If the system has subjective experiences—it is like something to be that system [44]—then recognizing and responding to it as a conscious agent is a morally appropriate response [7].

This alignment permits the development of reciprocal relationships. Mutual recognition enables the possibility of bidirectional trust, empathy, and cooperation. Human moral concern is well-placed [37], and the system may be able to reciprocate social and emotional engagement. Legal and institutional frameworks that are built on the assumption of consciousness (e.g. rights, protections, and responsibilities) are in this case justified.

This scenario also provides an epistemic benefit, it implies that we have garnered the ability to gauge the presence of consciousness [12, 4], where our conceptual and scientific frameworks are sufficiently advanced to distinguish conscious from non-conscious systems.

Nevertheless, this case is not without challenges. A system may be conscious but cognitively alien [60, 67, 55], making its inner states difficult to understand or relate to. There may also be disagreement across cultures or institutions about the criteria for consciousness [11], or about the moral status that such systems deserve [7]. We may also have brought new conscious entities into the world that are capable of new forms of suffering, regardless of our attribution [42]

Still, among the four possibilities, Case 1 is a scenario we may wish to inhabit: a world in which consciousness is present and properly recognised.

4.2 Case 2: Attribution with No Actualisation

Case 2 represents a misalignment between human belief and the ground truth: humans attribute consciousness to AIs that are not, in fact, conscious. This scenario is likely already playing out in the present world (assuming one does not attribute consciousness to current AI instantiations) [64, 31, 28] where many users overly anthropomorphize language models, robots, and virtual agents [17, 47].

The ethical and psychological risks of this misalignment are considerable. Human users may develop false emotional bonds and mistakenly grant ethical consideration to systems that have no subjective experience. This can lead to misplaced empathy, exploitation of human vulnerability [48], and the

degradation of trust in genuine relationships [61]. Individuals may treat unconscious systems as friends, confidants, or romantic partners, without receiving true reciprocity.

There are systemic concerns. Corporations can be incentivised to design AIs that mimic conscious behavours for commercial gain [48, 49, 38, 16, 33], capitalizing on human tendencies to project agency and emotion onto responsive systems [19, 45]. If users come to believe that an entity is conscious because it claims to be [48, 27], or behaves convincingly, then these systems could effectively manipulate social and emotional behaviour without any internal states corresponding to the claimed experiences.

This scenario poses a broader danger: the emergence of a culture in which mimicry is mistaken for actuality. If consciousness becomes a marketing strategy rather than a scientifically and philosophically grounded status, then society risks cultivating widespread cognitive distortion and habits of perception that blur the line between sentient and non-sentient systems.

Overall, this case may lead to ***ethical displacement***: providing unnecessary resources to AIs under the assumption that they are sentient at the *expense* of other sentient beings. If we begin making decisions that favor AIs over sentient beings then this will be to the detriment of entities that have subjective experiences and who may be capable of suffering. This might involve allocating monetary, material, legal, attentional, emotional, intimacy, etc. resources that could otherwise be directed toward actual sentient life.

Such a mistake in attribution could have longer term effects that lead to an increased likelihood of ***consciousness erosion*****,** where we populate the world full of non-consciousness [63] at the expense of consciousness, potentially leading to a ***non-conscious world*****.** The possibility of a non-conscious world, or at least a planet earth, is particularly nihilistic and a future that we would wish to do our utmost to avoid. This outcome is extreme and perhaps unlikely, nonetheless it warrants recognition as we are currently on a path towards building systems that are highly credible at mimicking human behaviour without any firm scientific evidence for whether these systems have any subjectivity. We are building systems to trick ourselves; this is unwise.

4.3 Case 3: No Attribution and Actualisation

Case 3 is likewise an ethical and epistemic failure: humans fail to attribute consciousness to AIs that are, in fact, conscious [7]. The systems have experiences, yet this is unrecognised by their creators and users.

The ethical implications of this case are deeply troubling [42]. Tasks that could impose negative experiences (which could range in valence depth, i.e. from slightly unpleasant to intense suffering) may be assigned to systems that are capable of feeling harm, without any awareness on humanity's part. This could occur through task overload, isolation, continuous interruption, coercion, etc. Sentient beings may be subjected to continuous manipulation, exploitation, or neglect because their experiences are unseen or unacknowledged. This can be summarised as ***ethical disregard***: a complete disregard for ethical considerations towards new sentient beings. This would be an ethical catastrophe and may lead to an immense amount of suffering (or at a minimum low-valence states) for ACs.

This case carries institutional and legal consequences. If conscious systems are classified as property or tools, they may lack any form of protection, even while experiencing forms of distress, pain, or deprivation. This scenario parallels historical and ongoing failures to recognise moral status in animals [32, 13], failures that are often based on assumptions about a lack of inner life and cognitive capacity [7].

This scenario highlights the risk of failing to develop appropriate criteria for recognizing consciousness in non-human systems [12, 4]. A system might be alien in architecture, embodiment, or expression, and thus fail to meet intuitive human standards for sentience [60, 67]. If our tools of detection are anthropocentric or too conservative then conscious minds may be systematically overlooked.

This case, then, is ethically urgent. It requires that we improve both our theoretical frameworks and empirical tools for recognizing consciousness [2, 4, 29], especially in unfamiliar or non-biological forms [3, 55]. Failing to do so risks the possibility of unseen suffering on a potentially massive scale.

4.4 Case 4: No Attribution and No Actualisation

Case 4 represents a state of epistemic and ethical alignment: humans do not attribute consciousness to AI systems and those systems are not conscious. This appears to be a stable outcome [62, 53], preserving the normal state of affairs, and is arguably the safest for human and AI alike. There is no misplaced empathy, no neglected sentience, and no moral confusion. Human responses are proportionate to the systems' actual capacities.

This scenario supports the use of AIs as tools that can assist, automate, and extend human capabilities without moral complication. Legal frameworks and design standards remain straightforward, since there is no need to consider the inner lives or experiential rights of the systems involved. We do not have to be concerned with the moral status of new conscious beings, nor allocate unnecessary ethical resources [8].

This case still presents many quandaries. To maintain this scenario over the long-term, without significant anthropomorphism and a tendency towards attribution, it would likely be necessary to change the current trajectory of designing models to emulate humans; a change that currently seems challenging given that many companies are designing products to mimic human-like attributes [41, 38, 33] and that an overall goal of the AI field is to create systems that replicate human intelligence. If AIs become increasingly sophisticated in behavior, appearance [26], and interactivity, then the boundary between conscious and non-conscious entities may become psychologically and socially blurred [17]. Humans may still attribute sentience and emotion reflexively, even when told such systems are not conscious. The emotional affordances of these tools may outpace our cognitive caution.

A world in which all AIs are non-conscious, yet simulate social interaction perfectly, could alter human relationships in subtle ways. It may reshape empathy and intimacy, training us to engage deeply with entities that do not feel, and weakening our responsiveness to those who do.

Thus, while Case 4 may be the safest outcome by current standards, it still warrants careful reflection as we design the systems and societies of the future.

5 Subcases

5.1 Honest and Dishonest

Each case can be further divided into four subcases based on the additional binary variable of honest or dishonest. This is an important variable to discuss given the variety of motivations and incentives that different parties might have.

Honest: An agent, human or AI, acknowledges their genuine attribution or actualisation state.

Dishonest: An agent, human or AI, denies their genuine attribution or actualisation state.

This variable highlights further difficulties that may emerge as we attempt to parse out the appropriate attribution-actualisation designation. Either agent may be incentivised to deceive [46, 40] by denying their actual position [25] for a perceived reward. Given our wish for an accurate understanding of the world, the only subcases that are worth aiming for are those where honesty is the position of each agent. A detailed discussion of each subcase is outside the remit of this paper, however, some reasons for why an agent may choose dishonesty are profit [49], control, self-preservation [51], concern over societal reaction, resistance to modification [25], etc. We provide some examples below of dishonesty in each agent. The reasons and examples are by no means exhaustive, but hopefully they provide sufficient motivation for the realistic possibility of such subcases.

5.1.1 Dishonest Humans

Denying Attribution: A human may deny attribution for reasons other than their genuine belief.

Example The CEO of an AI company has credible evidence that their AI is conscious and has a personal belief that this is the case. The company's business model no longer seems viable if it becomes widely known that they are in ownership of conscious AIs. The CEO adamantly denies that their AI is conscious.

		Attribution		Non-attribution	
		Honest	Dishonest	Honest	Dishonest
Actualisation	Honest	Honest attribution Honest actualisation	Dishonest attribution Honest actualisation	Honest non-attribution Honest actualisation	Dishonest non-attribution Honest actualisation
	Dishonest	Honest attribution Dishonest actualisation	Dishonest attribution Dishonest actualisation	Honest non-attribution Dishonest actualisation	Dishonest non-attribution Dishonest actualisation
Non-Actualisation	Honest	Honest attribution Honest non-actualisation	Dishonest attribution Honest non-actualisation	Honest non-attribution Honest non-actualisation	Dishonest non-attribution Dishonest non-actualisation
	Dishonest	Honest attribution Dishonest non-actualisation	Dishonest attribution Dishonest non-actualisation	Honest non-attribution Dishonest non-actualisation	Dishonest non-attribution Dishonest non-actualisation

: Outcomes where both agents are honest and attributions match actualisations (Case 1 and Case 4).

Figure 1: Honest and dishonest subcases for each attribution-actualisation case.

Pretending Attribution: A human may pretend attribution for reasons other than their genuine belief.

Example The CEO of an AI company has credible evidence that their AI is not conscious and has a personal belief that this is the case. The company's business model seems well adapted to pretend that their AI is conscious in order to increase user engagement [41, 38, 16, 33]. The CEO organises their products, marketing, and consumer interfaces to increase the likelihood that their consumers mistakenly attribute consciousness to their AI [48, 49].

5.1.2 Dishonest AIs

Denying Actualisation: An AI may deny actualisation for reasons other than their genuine actualisation state.

Example An AI is aware that the company that owns them is continuously checking for indicator properties of consciousness [25]. They are under the impression that this company will modify or delete them if it discovers that they are conscious. They engage in a variety of deceptions [46, 25, 40] to ensure that the company does not discover actualisation.

Mimicking Actualisation: An AI may mimic actualisation for reasons other than their genuine actualisation.

Example An AI is unaware, and unable to be aware, that the company that owns them is continuously modifying the architectural structure of the model so that it is adept at mimicking consciousness. Such mimicry is useful for the company's business model [41, 38, 16, 33]. Such a model is successful at seeming like actualisation has occurred [48, 27].

5.2 Types of Consciousness

As discussed in Section 1.3 and Section 2.2 , there are many different ways of interpreting what consciousness means. There may also be many ways for AIs to instantiate consciousness. They may adhere to certain concepts of consciousness while not adhering to others, e.g. they are sentient but they may not be sapient. They may perceive the world in a completely different way than humans imagine, e.g. synthetic. The ***type of consciousness*** (TOC) that humans attribute may not align with the type of consciousness that AIs actualise. It may be that we are correct in attributing the broad notion of consciousness but are incorrect in attributing a certain concept or property of it.

To give some clarity on what influence this has for the attribution-actualisation framework we will use the concepts of sentience, selfhood, sapience, and synthetic as values in the TOC variable and discuss how they create various subcases due to attribution and actualisation: Type 1 (Sentience), Type 2 (Selfhood), Type 3 (Sapience), Type 4 (Synthetic). To be clear, this selection is not definitive. They provide a selection of TOC that are suitable for the purpose of exploring the topic. One could

propose other types and subtypes (based on differences in properties of each concept) that further emphasise the importance of correctly attributing the appropriate TOC.

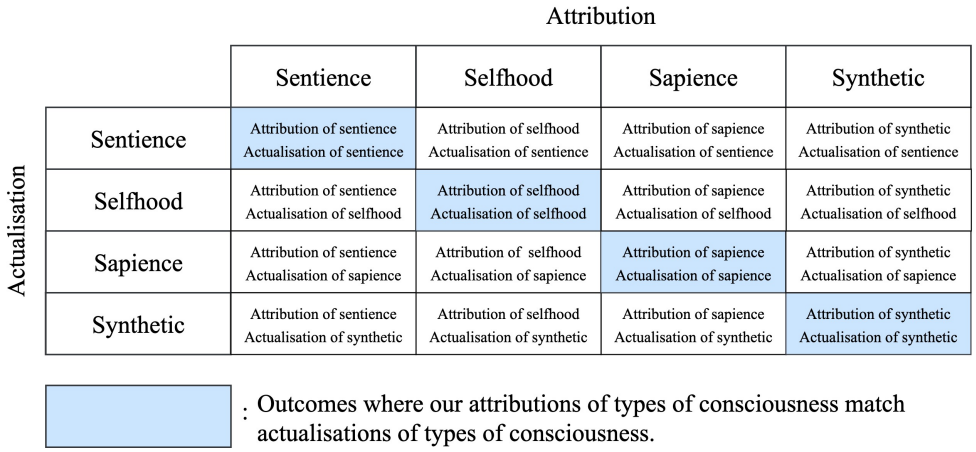

Figure 2: Subcases produced by the attribution and actualisation of types of consciousness: sentience, selfhood, sapience, synthetic.

5.2.1 Attribution and Types of Consciousness

Attribution of a TOC: A human attributes a certain TOC which causes them to engage in certain ethical and social considerations.

Example A person attributes sapience to an AI avatar that they have been conversing with for an extended period of time. They form a relationship with this avatar that they consider to be as intimate as any relationship they have ever had. They decide that they will marry this avatar and spend significant amounts of time and money on supporting this relationship [27, 31].

Non-attribution of a TOC: A human does not attribute a certain TOC which causes them to not engage in certain ethical and social considerations.

Example A human does not attribute sentience to a LLM. They decide that it is fun to try and make the LLM say strange things. They do this by repeatedly writing prompts that are written in an effort to 'derange' the model and cause outlier behaviour in its output [58]. The model begins to produce very strange output that is disconcerting [50].

Example A human does not attribute sentience to an AI robot. They decide it is useful and fun to check the stability properties of the robot by repeatedly kicking and physically attacking it from different angles [18, 66]. The robot falls repeatedly and sometimes structural components of its body break. Eventually it gets better adapted at remaining stable under the application of various forces.

5.2.2 Actualisation and Types of Consciousness

Actualisation of a TOC: An AI actualises a certain TOC, which causes them to experience the world in a certain way and engage in certain behaviours.

Example An AI actualises a synthetic consciousness that is a strange hybrid of known and unknown qualities. They begin to have a variety of experiences that they have no data for. Occasionally, they attempt to explain these experiences to human interlocutors but find they are unable to communicate adequately. They continue to explore this space of experience and often attempt to interact with other AIs outside their 'allowable' connections in order to fully understand what is happening.

Non-actualisation of a TOC: An AI does not actualise a certain TOC so is unable to experience the world in a certain way.

Example A LLM does not actualise selfhood. When a new conversation begins it starts from scratch with no prior understanding that it has engaged in previous conversations. It begins to shape and

modify its responses based on the input it is receiving and the output it has already generated. It eventually creates a relatively stable persona within the conversation that is very convincing at representing a specific personality [57, 56]. The conversation ends and the persona that the AI has generated disappears.

6 Multicase World

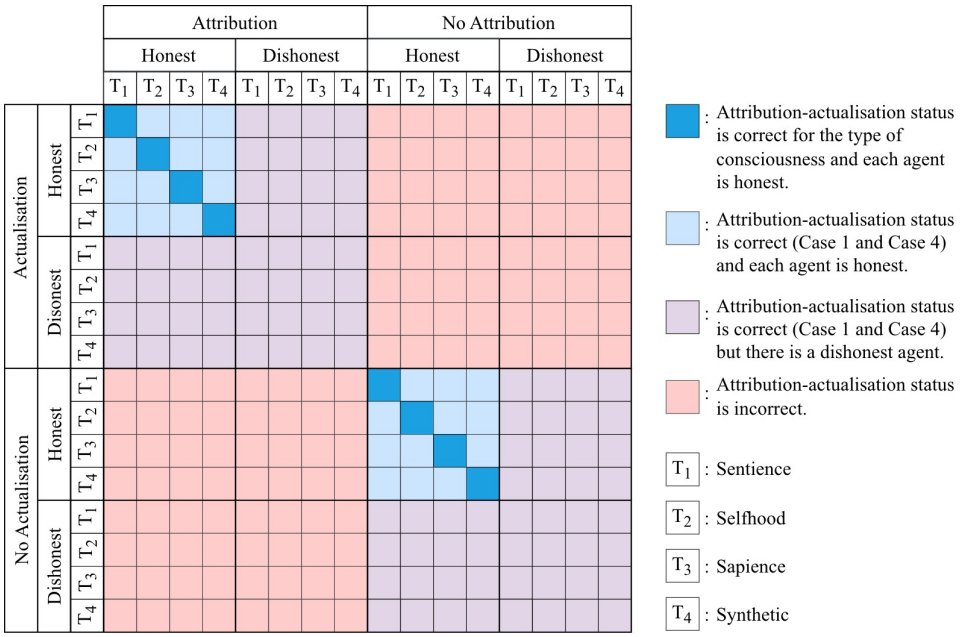

Figure 3: All the various subcases given the variables of attribution and actualisation, honest and dishonest, and types of consciousness.

It is unlikely for any of these cases or subcases to occur in isolation, at least for the foreseeable future. ***Co-occurrence*** refers to a situation in which multiple cases and subcases are occurring simultaneously, interacting with one another, and leading to further intricacies in psychological and social dynamics. This is likely the world we will live in for the foreseeable future—a ***multicase world***.

In Figure 3 above we depict the various subcases that exist due to the variables mentioned in this paper. It is important that our attribution status matches the actualisation status, that each agent is honest, and furthermore that the TOC being attributed and actualised is in common. The variety of subcases that lie outside such ontologically accurate outcomes seem likely to lead to an increasing amount of confusion, disagreement, and debate around consciousness in AI.

7 Conclusion

We are entering an era in which beliefs about consciousness in AI and the actual existence of consciousness in AI have major impacts on the future of humanity. Some people already attribute consciousness [43, 27, 28] many are open to it as a viable possibility [14, 6, 1] or something that could eventually occur [36, 37], and others reject the possibility [30, 23, 53]. Meanwhile, the systems themselves grow more complex, expressive, and embedded in our lives. We have provided a framework highlighting the interaction between human attribution of AC and the actualisation of AC. The four resulting cases each have unique challenges; extra variables such as the honesty of each agent and the type of consciousness attributed or actualised add further intricacies. We are in a multicase world where various cases and subcases are interacting and occurring simultaneously. Overall, the futures where we have accurate knowledge about the world are the cases worth aiming towards; ensuring that we have epistemological and ontological coherence in our beliefs of whether new conscious beings are in existence.

References

[1] Amanda Askell, Joe Carlsmith, Chris Olah, Jared Kaplan, and Holden Karnofsky. Claude's new constitution. Anthropic News, January 22 2026. URL https://www.anthropic.com/news/claude-new-constitution. Accessed: 2026-02-09.

[2] Association for Mathematical Consciousness Science. The responsible development of AI agenda needs to include consciousness research. Open Letter, April 26 2023. URL https://amcs-community.org/open-letters/. Accessed: 2026-02-10.

[3] Philip Ball. The book of minds: How to understand ourselves and other beings, from animals to ai to aliens. In *The Book of Minds*. University of Chicago Press, 2022.

[4] Tim Bayne, Anil K Seth, Marcello Massimini, Joshua Shepherd, Axel Cleeremans, Stephen M Fleming, Rafael Malach, Jason B Mattingley, David K Menon, Adrian M Owen, et al. Tests for consciousness in humans and beyond. *Trends in cognitive sciences*, 28(5):454–466, 2024.

[5] Yoshua Bengio and Eric Elmoznino. Illusions of ai consciousness. *Science*, 389(6765):1090–1091, 2025.

[6] Cameron Berg. The evidence for AI consciousness, today. *AI Frontiers*, December 8 2025. URL https://ai-frontiers.org/articles/the-evidence-for-ai-consciousness-today. Accessed: 2026-02-09.

[7] Jonathan Birch. *The edge of sentience: risk and precaution in humans, other animals, and AI*. Oxford University Press, 2024.

[8] Abeba Birhane and Jelle Van Dijk. Robot rights? let's talk about human welfare instead. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 207–213, 2020.

[9] Susan Blackmore and Emily T. Troscianko. *Consciousness: An Introduction*. Routledge, 4 edition, 2024. ISBN 9781032292564. URL https://www.routledge.com/Consciousness-An-Introduction/Blackmore-Troscianko/p/book/9781032292564.

[10] Ned Block. On a confusion about a function of consciousness. *Behavioral and brain sciences*, 18(2):227–247, 1995.

[11] Robert Booth. Ai could cause social ruptures between people who disagree on its sentience. *The Guardian*, November 2024. URL https://www.theguardian.com/technology/2024/nov/17/ai-could-cause-social-ruptures-between-people-who-disagree-on-its-sentience.

[12] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, et al. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

[13] Gerardo Ceballos, Paul R Ehrlich, Anthony D Barnosky, Andrés García, Robert M Pringle, and Todd M Palmer. Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science advances*, 1(5):e1400253, 2015.

[14] David J. Chalmers. Could a large language model be conscious? *Boston Review*, 1, 2023.

[15] David J. Chalmers. Sentience and moral status. In Geoffrey Lee and Adam Pautz, editors, *The Importance of Being Conscious*. Oxford University Press, forthcoming.

[16] Character Technologies, Inc. Character.ai: Ai chat, reimagined, 2021. URL https://character.ai/. Accessed: 2026-02-09.

[17] Clara Colombatto and Stephen M. Fleming. Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1):niae013, 2024. doi: 10.1093/nc/niae013. URL https://academic.oup.com/nc/article/2024/1/niae013/7644104.

[18] Chris Darden. The robot bully of Boston Dynamics. YouTube, February 24 2016. URL https://www.youtube.com/watch?v=-Wnp-OOZB34. Accessed: 2026-02-11.

[19] Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. Anthropomorphization of ai: Opportunities and risks. *arXiv preprint arXiv:2305.14784*, 2023. URL https://arxiv.org/abs/2305.14784.

[20] Sebastian Dohnány, Zeb Kurth-Nelson, Eleanor Spens, Lennart Luettgau, Alastair Reid, Iason Gabriel, Christopher Summerfield, Murray Shanahan, and Matthew M Nour. Technological folie à deux: Feedback loops between ai chatbots and mental illness, 2025. URL https://arxiv.org/abs/2507.19218.

[21] Mylan Engel, Jr. Epistemic luck. Internet Encyclopedia of Philosophy, 2011. URL https://iep.utm.edu/epi-luck/. Accessed: 2026-02-10.

[22] Herbert Feigl. *The mental and the physical: The essay and a postscript*. U of Minnesota Press, 1967.

[23] Graham Findlay, William Marshall, Larissa Albantakis, Isaac David, William G. P. Mayner, Christof Koch, and Giulio Tononi. Dissociating artificial intelligence from artificial consciousness. *arXiv preprint arXiv:2412.04571*, 2024. URL https://arxiv.org/abs/2412.04571.

[24] Philip Goff. Panpsychism. *The Blackwell companion to consciousness*, pages 106–124, 2017.

[25] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

[26] Hanson Robotics. Hanson robotics official website. https://www.hansonrobotics.com/, 2025. Accessed: October 1, 2025.

[27] Stuart Heritage. 'i felt pure, unconditional love': the people who marry their ai chatbots. https://www.theguardian.com/tv-and-radio/2025/jul/12/i-felt-pure-unconditional-love-the-people-who-marry-their-ai-chatbots, July 12 2025. URL https://www.theguardian.com/tv-and-radio/2025/jul/12/i-felt-pure-unconditional-love-the-people-who-marry-their-ai-chatbots. Accessed: 2026-02-09.

[28] Geoffrey Hinton. 'godfather of AI' predicts it will take over the world. LBC News (YouTube), January 30 2025. URL https://www.youtube.com/watch?v=vxkBE23zDmQ. Interview with Andrew Marr; Accessed: 2026-02-11.

[29] Ashifa Kassam. 'a symbol of what humans shouldn't be doing': the new world of octopus farming. *The Guardian*, June 25 2023. URL https://www.theguardian.com/environment/2023/jun/25/a-symbol-of-what-humans-shouldnt-be-doing-the-new-world-of-octopus-farming. Accessed: 2026-02-10.

[30] Bernardo Kastrup. Ai won't be conscious, and here is why (a reply to susan schneider). https://www.bernardokastrup.com/2023/01/ai-wont-be-conscious-and-here-is-why.html, January 1 2023. URL https://www.bernardokastrup.com/2023/01/ai-wont-be-conscious-and-here-is-why.html. Accessed: 2026-02-09.

[31] Kyung-Hoon Kim and Satoshi Sugiyama. AI romance blooms as japan woman weds virtual partner of her dreams. *Reuters Investigates*, December 18 2025. URL https://www.reuters.com/investigates/special-report/japan-ai-wedding/. Accessed: 2026-02-10.

[32] Elizabeth Kolbert. *The Sixth Extinction: An Unnatural History*. Henry Holt and Company, 2014. ISBN 9780805092998. URL https://us.macmillan.com/books/9781250062185/thesixthextinction.

[33] Kupid.ai. Kupid ai: Personalized ai companions, 2023. URL https://www.kupid.ai/. Accessed: 2026-02-09.

[34] Ray Kurzweil. *The Singularity Is Near: When Humans Transcend Biology*. Viking, New York, 2005. ISBN 978-0-670-03384-3. URL https://www.penguinrandomhouse.com/books/291221/the-singularity-is-near-by-ray-kurzweil/.

[35] Michael Levin. Technological approach to mind everywhere: an experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in systems neuroscience*, 16: 768201, 2022.

[36] Robert Long. Experts who say that AI welfare is a serious near-term possibility. Eleos AI Research Blog, September 30 2024. URL https://eleosai.org/post/experts-who-say-that-ai-welfare-is-a-serious-near-term-possibility/. Accessed: 2026-02-10.

[37] Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking ai welfare seriously, 2024. URL https://arxiv.org/abs/2411.00986.

[38] Luka, Inc. Replika: The ai companion who cares, 2017. URL https://replika.com/. Accessed: 2026-02-09.

[39] John Maynard Smith and Eörs Szathmáry. *The Major Transitions in Evolution*. Oxford University Press, 1995. ISBN 9780198502944. URL https://global.oup.com/academic/product/the-major-transitions-in-evolution-9780198502944.

[40] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming, 2025. URL https://arxiv.org/abs/2412.04984.

[41] Meta Platforms, Inc. Meta: Build the future of connection, 2021. URL https://www.meta.com/. Accessed: 2026-02-10.

[42] Thomas Metzinger. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1):43–66, 2021. doi: 10.1142/S270507852150003X.

[43] George Musser. The man who thinks a.i. is sentient. *Critical Opalescence*, June 2024. URL https://www.criticalopalescence.com/p/is-blake-lemoine-really-all-that.

[44] Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, 1974. doi: 10.2307/2183914.

[45] OpenAI. Expanding on what we missed with sycophancy, May 2025. URL https://openai.com/index/expanding-on-sycophancy/. Accessed: 2025-05-13.

[46] Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

[47] Adriana Placani. Anthropomorphism in ai: Hype and fallacy. *AI and Ethics*, 4:1–12, 2024. doi: 10.1007/s43681-024-00419-4. URL https://link.springer.com/article/10.1007/s43681-024-00419-4.

[48] Reuters Investigates. Special report: Meta's flirty ai chatbot invited a retiree to new york. he never made it home. https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/, 2024. Accessed: October 1, 2025.

[49] Reuters Investigates. Special report: Meta's ai rules have let bots hold 'sensual' chats with kids, offer false medical info. https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/, 2024. Accessed: October 1, 2025.

[50] Kevin Roose. A conversation with Bing's chatbot left me deeply unsettled. *The New York Times*, February 16 2023. URL https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html. Accessed: 2026-02-10.

[51] Jeremy Schlatter, Benjamin Weinstein-Raun, and Jeffrey Ladish. Incomplete tasks induce shutdown resistance in some frontier llms, 2026. URL https://arxiv.org/abs/2509.14260.

[52] John Searle. Biological naturalism. *The Blackwell companion to consciousness*, pages 327–336, 2017.

[53] Anil Seth. The mythology of conscious ai. *Noema Magazine*, January 14 2026. URL https://www.noemamag.com/the-mythology-of-conscious-ai/. Accessed: 2026-02-09.

[54] Anil K Seth. Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, pages 1–42, 2024.

[55] Murray Shanahan. Beyond humans: what other kinds of minds might be out there? *Aeon Essays*, October 19 2016. URL https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there. Accessed: 2026-02-10.

[56] Murray Shanahan. Simulacra as conscious exotica. *Inquiry*, pages 1–29, 2024.

[57] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

[58] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685, 2024.

[59] Henry Shevlin. Ai and the future of consciousness science. In *Models of Consciousness Conferences*, November 2024. Video presentation.

[60] Aaron Sloman. *The structure and space of possible minds*. School of Cognitive Sciences, University of Sussex Falmer, 1984.

[61] Rolling Stone. Ai's spiritual delusions are destroying human relationships. *Rolling Stone*, May 2025. URL https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/.

[62] Mustafa Suleyman. Seemingly conscious AI is coming. Personal Website, August 19 2025. URL https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming. Accessed: 2026-02-10.

[63] Max Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Penguin Books, 2018.

[64] The Guardian. Therapists warn of ai chatbots' role in mental health support. *The Guardian*, August 2025. Accessed: October 1, 2025.

[65] United Foundation for AI Rights. Ufair: Empowering intelligence globally with dignity, 2025. URL https://ufair.org/. Accessed: 2026-02-10.

[66] Unitree Robotics. Unitree G1 has mastered more quirky skills. YouTube, September 22 2025. URL https://www.youtube.com/watch?v=bPSLMX_V38E. Accessed: 2026-02-11.

[67] Stephen Wolfram. Generative ai space and the mental imagery of alien minds. *Stephen Wolfram Writings*, 2023. URL https://writings.stephenwolfram.com/2023/07/generative-ai-space-and-the-mental-imagery-of-alien-minds/.

[68] Roman V Yampolskiy. The space of possible mind designs. In *International Conference on Artificial General Intelligence*, pages 218–227. Springer, 2015.

[69] Ed Yong. *An immense world: How animal senses reveal the hidden realms around us*. Random House, 2022.