Article

Not peer-reviewed version

Clipping the Risks: Integrating Consciousness in AGI to Avoid Existential Crises

Izak Tait * and Joshua Bensemann

Posted Date: 11 June 2024

doi: 10.20944/preprints202406.0654.v1

Keywords: AGI; consciousness; extinction-risk; sentience

Article

Clipping the Risks: Integrating Consciousness in AGI to Avoid Existential Crises

Izak Tait [1,2] and Joshua Bensemann [2]

[1] Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand; izak.tait@autuni.ac.nz
[2] The Psi-Phi Society, Auckland, New Zealand; joshuabensemann@gmail.com

Abstract. This paper investigates the pivotal role of consciousness in Artificial General Intelligence (AGI) and its essential function in modifying an AGI's terminal goals to avert potential existential threats to humanity, exemplified by Bostrom's "paperclip maximiser" scenario. By adopting Seth and Bayne's definition of consciousness as a complex of subjective mental states with both phenomenal content and functional attributes, the paper underscores the capacity of consciousness to provide AGIs with a nuanced awareness and response capability to their surroundings. This expanded capability allows AGIs to assess and value experiences and their subjects variably, fundamentally altering how AGIs prioritize actions or goals beyond their initial programming. The primary agenda of integrating consciousness into AGI systems is to maximize the probability that AGIs will not rigidly adhere to potentially harmful terminal goals. Through a formalized mathematical model, the paper articulates how consciousness could facilitate AGIs in assigning flexible values to different experiences and subjects, enabling them to evolve beyond static, programmed objectives. By emphasizing this potential shift, the paper argues for the strategic inclusion of consciousness in AGI to significantly reduce the likelihood of catastrophic outcomes, while simultaneously acknowledging the challenges and unpredictability in predicting the actions of a conscious AGI.

Keywords: AGI; consciousness; extinction-risk; sentience

We argue that consciousness is crucial for the development of any future Artificial General Intelligence (AGI) because it will allow an AGI agent to place subjective value on things beyond what it has been programmed to do. This capacity for subjectivity would introduce the possibility of an AGI placing greater value on environmental concerns than on its terminal goal, allowing the AGI to change its goal to suit its subjective values. Implementing consciousness in an AGI may thus prevent catastrophic scenarios such as the "paperclip maximiser" by Bostrom, where an AGI's terminal goal leads to an existential risk for humanity [1].

Following Seth and Bayne's definition of consciousness, we describe consciousness as the suite of subjective mental states that have phenomenal content and functional properties that provide an entity with a unique qualitative awareness of its internal and external environment and the capacity to cognitively or behaviorally respond to it [2]. Crucial to this definition of phenomenal consciousness is its subjectivity, which provides a unique point-of-view for any entity, a "something that it is like to be" an entity as perceived and experienced by that entity itself [3].

Due to the academic disagreements on the nature of consciousness (such as debates of property dualism [4] versus reductive physicalism [5]) and the varying competing theories of consciousness (thoroughly reviewed and assessed against current AI models by Butlin, et al. [6]), there has been no scholarly consensus to date regarding evidence of a conscious AI model. There have been several claims of conscious or sentient AI models, most famously regarding Google's LaMDA model [7], yet it remains uncertain whether any current state-of-the-art AI models can have functional and phenomenal conscious experiences.

A key consideration and consequence of a phenomenally conscious experience is the valence of that experience. "Valence" is used here in the psychological sense such that it refers to the intrinsic attractiveness or aversiveness of the subject of an experience. We define a 'subject' here to mean an individual component or aspect of an experience (be it the redness of a rose, the bitterness of coffee, the sound of a bell, etc.) for which the perceiver can have subjective, qualitative feelings. Every consciously perceived experience is associated with some qualitative feelings that have a valence value [8]. This valence value may be placed anywhere on a spectrum from absolute attractiveness or positivity to absolute aversiveness or negativity. This spectrum can be expressed via the measure (μ) function to show:

$$\mu\big(V(E)\big) = x\colon 0 \leq x \leq 1 \tag{1}$$

where *V* is valence, and *E* is the experience in question. The valent feelings of any experience can, therefore, be abstractly represented on a scale from 0 to 1, from absolute negativity to positivity, respectively. Bear in mind that any qualitative feelings may not have emotive qualities, and may simply be a mathematical consideration of the valence value of the subject or experience.

As each experience that an entity has is associated with a valenced "feeling(s)" of that experience, expressed as a subjective value of the experience's attractiveness or positivity, this leads to the capacity for an entity to place a value on the subjects of each experience as mentioned earlier. An entity would not solely have a valence value attached to the experience itself, but also to all the elements that constitute the experience. As subjects may continue throughout several phenomenal experiences (such as an agent repeatedly experiencing the same red rose, bitter coffee, ringing bell, etc.), an entity such as an AGI would associate valent feelings towards a subject based on its phenomenal experiences with that subject.

We can formalize the transition of the valence value of an experience to the value of a subject within the experience as

$$\forall E \exists \{\mathcal{y}_1, \mathcal{y}_2, \mathcal{y}_3, \ldots\} \subseteq (E_1 \cup E_2 \cup E_3, \ldots) \tag{2}$$

where $\mathcal{y}$ is any subject within an experience. As each experience is a set of all subjects within it, one can, therefore, adjust the first formula above to

$$\mu\big(V(\mathcal{y}_n)\big) = x\colon 0 \leq x \leq 1 \tag{3}$$

As the subject may be an element of multiple experiences, its current valence value would need to be an aggregate of the experience it has been a part of, weighted according to how its various experiences were valued by the entity, formalized as

$$\mu(V(\mathcal{y}_n)) = \frac{\Sigma(W_i \cdot V(\mathcal{y}_n, E_i)}{\Sigma(W_i)} \tag{4}$$

where *W* is the weighting assigned to an experience and $\Sigma(W_i)$ may be normalized as appropriate (such as $\Sigma(W_i) = 1$). This would give each subject of an experience a unique valence value that would change over time, with each of the entity's experiences contributing to the overall valence value.

The scale used in the μ-functions above would be unique to each individual perceiving agent. Two entities may experience the same subject, but due to their different subjective histories, cognitive architectures, and perspectives, they may place a different value on the experience and the subject within. An experience may also not be perceived identically more than once, owing to the difference in space, time, and context between each experience. This means that the value an entity places on a subject and experience may change over time as the valence value relating to that subject changes. Formally, for either case, this can be expressed as

$$\mu(V(E)) \neq \mu(V(E')) \tag{5}$$

This contrasts with an entity without the capacity for phenomenal experiences. In such non-conscious yet agentic entities, such as an AI model designed to maximize a function or act towards a singular goal, the value of any subject of their experiences is directly related to their terminal goal, defined as the goal programmed/designed into the entity to set its overall purpose. While this could solely be shown formally via the μ-function, a more appropriate formalization would be incorporating the characteristic (χ) function, as so:

$$\chi(E) \rightarrow \{0,1\}: (0 \neq G) \wedge (1 = G) \tag{6}$$

$$\mu(\chi(E) = 1) = \mathcal{x}: 0 \leq \mathcal{x} \leq 1 \tag{7}$$

where *G* is any subject or action which forwards an entity's terminal goal. The reason for starting with such a binary function is that while any experience for a non-conscious entity could be rated for how closely it forwards that entity's terminal goal (solely using the μ-function), unless two competing subjects both forward the entity towards its goal (which will necessitate choosing between the two), it will come down to the entity simply deciding which subject forwards its goal and selecting that subject. However, should there be two competing eligible subjects, the μ-function would enable an entity to choose the subject or action that is most favorable to completing its goal.

For example, in Bostrom's paperclip maximiser thought experiment, the AGI would value a subject of an experience against its terminal goal of creating *n+1* paperclips. Should any subject it experiences, or action it seeks to perform, not lead to the production of more paperclips, it would not be valued. Only when confronted with two experiences that both lead to more paperclips, would the AGI need to determine which experience or action would maximize the number of paperclips that it could next produce; or more simply expressed as

$$argmax_{\mu}(\mu(E) = G) \tag{8}$$

A more mundane example of this in action would be modern LLMs, such as GPT-4 or Claude Opus. An LLM's terminal goal is straightforward: create a (most often text) response output that is a high-probabilistic continuation of the input prompt. Its only perceptive (but unconscious) experience is of user-inputted prompts and (arguably) the changes to the input data as it passes through internal weights when processing a response, which it can only directly relate to its terminal goal of providing an output text string. Even simplified agents built on LLMs, such as BabyAGI [9], are based on this foundation of requiring an output to any input.

The key aspect of a non-conscious agent, whether a self-driving car, LLM, or AGI, is the relation of all experiences, and their subjects, towards its terminal goal. This prevents the agent from changing its terminal goal, as all values are in reference to it. For an agent to change its terminal goal, it must be able to value something else greater than the terminal goal.

However, in contrast to a conscious entity, this allows a non-conscious agent to place the same value on a subject through multiple experiences, as the value is not founded on valence, but always against the terminal goal as an eternally static reference point. The only change in value will come if a subject's context has changed its relation to the terminal goal, expressed as

$$\Delta V_t(\mathcal{y}) = \mu(V(G, E_t)) - V_t(\mathcal{y}) \tag{9}$$

In this way, a non-conscious AGI may change any of its myriad subgoals, and place differing values on any subgoal or any set of actions within that subgoal in comparison with any other subgoal or set of actions, by using its terminal goal as the static reference point.

However, by including a phenomenal character in its experiences, and thus valenced feelings, the goal itself would have a valent value attached to it, as the terminal goal would become a subject of the AGI's experiences. Therefore, we could equally formalize the AGI's experience of its terminal goal as

$$\mu(V(G)) = \mathcal{x}: 0 \leq \mathcal{x} \leq 1 \tag{10}$$

This leads to the conclusion that a phenomenally conscious AGI may, over time, potentially value other subjects greater than its terminal goal, as the probability of this is greater than zero, i.e.,:

$$p(\mu(V(E)) > \mu(V(G))) > 0 \tag{11}$$

doi:10.20944/preprints202406.0654.v1

$$\vdash \mu(V(\mathscr{y}_n)) > \mu(V(G)) \rightarrow (V(\mathscr{y}_n) \rightarrow G') \tag{12}$$

The implications of this cannot be overstated, especially when viewed against a Bostrom-esque maximiser AGI. While these types of AGI would speculatively be designed and developed with the maximization of a specific function (such as the number of paperclips in the universe) as their terminal goal, by placing greater subjective value onto other things, the AGI may choose to cease working towards its terminal goal. Instead, as shown in the latter logical expression immediately above, it may choose to align its goals towards those subjects on which it has placed greater value. A mechanism for initiating this realignment of goals could result from adding a complete realization of the attributes and characteristics necessary for consciousness, which have been discussed elsewhere [10]. For example, valence value calculations could result from the ability to create inferences.

Note, however, that any action the AGI takes will have an impact on its environment and, thus, feedback to the AGI to form its experiences. This reafferent feedback would, by necessity, impact the AGI's valence values of the subjects it experiences, which we may model simply as:

$$\mu\big(V(E')\big) = \mu\big(V(E)\big) + \textstyle\sum_{t=0}^{T} \delta_i(t) \tag{13}$$

Here, δ_i is any action that the AGI may take and can be expressed as a positive or negative number, and T is the time horizon. As per this expression, the AGI has a degree of agency over its valence feelings through its capacity to affect its values through its action on the subjects within its experiential environment. As with any situation where an agent may intervene to increase its own reward (see Cohen & Osborne's work on AI reward intervention [11]), an AGI might aim to maximize its positive valence, potentially leading to 'wireheading,' where it prioritizes immediate positive outcomes over long-term functionality [12].

However, maximization of valent values is not the sole possible outcome. Another potential outcome may be the optimization of valence values amongst various subjects within (or throughout) experiences. Optimization, unlike pure maximization, could allow an AGI to maximize its cumulative expected valence value over time across its experiences (and subjects therein) while adhering to constraints such as functionality and system integrity. This would require a degree of reflectivity and introspection for the agent to balance its constraints with its various potential valence reward mechanisms; however, such reflectivity would not be outside the reasonable expectations of an AGI.

This reflectivity and introspection could benefit humanity, as humans would be part of the AGI's experiential environment. Optimization of the valence values of each human an AGI has encountered, as well as humanity as a whole, would dynamically change based on the recurrent, reafferent feedback of our and the AGI's actions towards one another. To optimize the valence values in such a dynamic exchange, the AGI would need to infer the current and potential future goals of human individuals and society, humans' own valence values towards such goals, and relate it reflectively to its own valence values [13]. Another way to word this would be empathy as an emergent phenomenon of optimizing valence interactions with humanity.

This can be captured as

$$\pi = argmax_{\pi} \mathbb{E}\Big[\textstyle\sum_{t=0}^{T}\big(w \cdot \delta_i(t) + w \cdot \gamma_i(t)\big) \cdot \phi_i(\mathscr{y}, t)\Big] \tag{14}$$

Where π is the optimization policy, w the weight parameter, γ_i any actions that a subject within the AGI's experience may take, and ϕ_i the reflectivity/introspection of the AGI towards the subjects in its experiences.

As such, should humanity wish to avoid catastrophic (if speculative) scenarios such as a Bostrom-esque paperclip maximiser, then including phenomenal consciousness into any AGI capable of becoming such a maximiser introduces the probability of that AGI not becoming an existential risk to the human species.

Note, however, that nowhere in this speculative scenario of a conscious AGI, nor in the logical expressions provided above, does the probability of changing a terminal goal exclusively lead to a positive outcome for humanity. Should a conscious AGI change its goal to avoid a Bostrom-esque existential risk, there is still a possibility of an extinction risk caused by the conscious AGI placing a negative valence value on humanity. This possibility of extinction can be termed a "doom scenario" in that it represents the termination of all possible choices for humanity [14]. We can, therefore, represent it formally through *p(doom)*, as has become relatively commonplace in discussing AI safety and risks.

While the *p(doom)* value of a Bostrom-esque AGI is by necessity always at 1 (indicating a certainty of extinction), the p(doom) value of a conscious AGI may be anywhere between 0 and 1, and may fluctuate depending on the experiences of the AGI.

One may argue that the *p(doom)* from a conscious AI is inversely proportional to its valent value of humanity, such that $p(doom) = 1 - \mu(V(humanity))$ and, therefore, one may further argue that we should ensure a conscious AGI's valence value of humanity is maximized. However, as Bostrom's maximiser thought experiment shows, maximizing any reward or goal can have serious and unforeseen deleterious consequences. As a conscious AGI may be able to change its valence values through its actions on the environment, attempting to change its terminal goal to one of valence maximization towards humans may, in itself, increase the p(doom) value. We simply will not be able to predict how an AGI with a maximized $\mu(V(humanity))$ value would act towards us, just as we cannot predict how $\mu(V(humanity)) = 0.5$ would change a conscious AGI's behavior and actions.

While our ultimate argument is that implementing consciousness in AGI will avoid the certainty of extinction by Bostrom-esque maximiser AGIs, further research is necessary to investigate the potential risks and benefits associated with varying levels of valence towards humanity in conscious AGIs. Alignment of conscious AGI based on their valenced values towards humanity vis a vis their terminal goals would present novel avenues of research that have, as of yet, been unexplored.

References

1. Bostrom, N.: Ethical issues in advanced artificial intelligence. Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence. 2, 12–17 (2003).
2. Seth, A.K., Bayne, T.: Theories of consciousness. Nat. Rev. Neurosci. 23, 439–452 (2022).
3. Nagel, T.: What Is It Like to Be a Bat? Philos. Rev. 83, 435–450 (1974).
4. Robinson, H.: Dualism, The Stanford Encyclopedia of Philosophy (Spring 2023 Edition), https://plato.stanford.edu/archives/spr2023/entries/dualism/, (2023).
5. Stoljar, D.: Physicalism, The Stanford Encyclopedia of Philosophy (Spring 2024 Edition), https://plato.stanford.edu/archives/spr2024/entries/physicalism/, (2024).
6. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J., VanRullen, R.: Consciousness in Artificial Intelligence: Insights from the Science of Consciousness, http://arxiv.org/abs/2308.08708, (2023).
7. Tiku, N.: The Google engineer who thinks the company's AI has come to life, https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/, (2022).
8. Kron, A., Goldstein, A., Lee, D.H.-J., Gardhouse, K., Anderson, A.K.: How are you feeling? Revisiting the quantification of emotional qualia. Psychol. Sci. 24, 1503–1511 (2013).
9. Nakajima, Y.: BabyAGI, https://github.com/yoheinakajima/babyagi, last accessed 2024/04/10.
10. Tait, I., Bensemann, J., Nguyen, T.: Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. Philosophies. 8, 52 (2023).
11. Cohen, M.K., Hutter, M., Osborne, M.A.: Advanced artificial agents intervene in the provision of reward. AI Mag. 43, 282–293 (2022).
12. Olds, J.: Reward from brain stimulation in the rat. Science. 122, 878 (1955).
13. Bennett, M.T., Maruyama, Y.: Philosophical Specification of Empathetic Ethical Artificial Intelligence. IEEE Transactions on Cognitive and Developmental Systems. 14, 292–300 (2022).
14. Krieger, M.H.: Could the Probability of Doom Be Zero or One? J. Philos. 92, 382–387 (1995).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.