# Understanding and Assessing Artificial Consciousness: A Guide for Grant Proposals Review

**Maaz Saad**

Ontario Tech University, Oshawa, ON, L1G
4Y3, Canada

■ **Introduction**: The entire world underwent the Industrial Revolution in the 19th century, which altered the course of human history. I believe that the world is currently undergoing another revolution in the 21st century that will reshape the future for our generations. This revolution will be brought about by the society-transforming technology known as Artificial Intelligence (AI) [1]. AI can be thought of as a sequence with four layers of iterations [1]. The first layer is simple machine learning, such as building a movie recommendation system [1]. The second layer is machine intelligence, such as self-driving cars or virtual assistants [1]. The third layer is machine creativity, such as the ability of a machine to operate in unknown conditions [1]. Finally, the last layer is machine consciousness, which is the primary focus of this report.

Before I delve further into artificial consciousness, I would like to explain the purpose of this report. As the title of the report suggests, this report aims to advocate for the field of artificial consciousness, with the goal of securing increased funding to accelerate its evolution. The report will serve as a guide for a hypothetical grants committee to facilitate the allocation of funds for further research and development of artificially sentient technology. I will frame this report so that it explains the concept and ideology of artificial consciousness in the context of today's landscape and presents its applications and implications. Most importantly, I will provide a framework for evaluating proposals in the context of explainability, interpretability, fairness, and privacy, as these are important tools when evaluating any artificial intelligence models, especially machine consciousness ones.

## Human Consciousness and Artificial Consciousness

Before I define artificial consciousness, I need to present an explanation of basic consciousness. Consciousness can be defined as the state of being aware of one's own existence [1]. In other words, it involves being aware of one's surroundings, thoughts, feelings, and the physical environment, which is how we, as humans, perceive consciousness. Additionally, the ability to take actions in our lives and comprehend the consequences of those actions can further illustrate the construct of consciousness [1]. Keeping this in mind, it is safe to say that, for humans, consciousness can be divided into two categories: *primitive consciousness* and *reflective consciousness* [1]. In simple terms, primitive consciousness can be described as self-awareness, while reflective consciousness can be defined as a person's capacity to analyze their own existence in light of the decisions or choices that they make, which can affect them or others [1].

There has been an exponential rise in the research and development of artificial intelligence, which can not only reproduce human intelligence but also replicate emotions [2]. Some researchers claim that achieving emotional intelligence in machines, leading to artificial consciousness, could be an engineering challenge that can be solved by using the current technological tools available [2]. However, this may not be entirely accurate, as we will need more advanced tools that have yet to be developed to achieve the next breakthrough in machine consciousness. Therefore, the grants committee should carefully review the proposals they receive and provide funds to develop the necessary tools and technologies to help achieve the goal of machine consciousness.

*Phenomenal consciousness* refers to the subjective experience of consciousness, such as the individualistic experience of feeling happiness, sadness, pain, or pleasure [2]. Researchers believe that when we apply this principle to AI, it means that an artificial intelligence model should be able to have subjective experiences similar to humans for the prospect to be valid [2]. While AI can be programmed to be ethically responsible, it remains to be seen whether it can produce emotions that have not been pre-programmed [2]. As humans, we employ our emotional intelligence and empathy to make moral decisions, but the same cannot be said for AI models [2]. Researchers are actively exploring ways of achieving phenomenal consciousness in machines, and with the right tools and funding, we may see some promising developments.

## Applications and Implications

Shifting our focus, let's now explore potential applications of artificial consciousness. Some initial versions of these applications are already in use, and introducing artificial consciousness to them could propel them to the next level. However, like any technological advancement, this also introduces new risks that could negatively impact humans and must be carefully considered before putting these tools into action. In this section, I’ll examine some examples from my personal knowledge and experience.

Artificial consciousness could revolutionize aviation, but human pilots are still necessary for safety [3]. AI systems are currently used in aircraft design, and auto-pilot can be engaged in straightforward conditions [3]. However, human pilots must take over during turbulence or unfamiliar situations [3]. If AI systems could navigate these situations with their consciousness, the term "auto-pilot" would take on a new meaning. Nevertheless, the question of accountability arises if commercial flights were controlled entirely by AI pilots, especially in case of a catastrophe. Who would be held responsible: the AI pilot designers or the AI pilot itself? How would accountability proceedings be conducted?

Artificial consciousness could improve video games by creating more personalized experiences [4]. Currently, video game characters are one-dimensional and interact with all players in the same way [4]. Consciousness tools could make characters aware of the human player they are interacting with, leading to a unique experience for each user [4]. However, implementing such features would require addressing potential issues of hurtful or inappropriate choices by characters.

Artificial consciousness could revolutionize mental healthcare by providing an alternative to costly therapy [5]. AI-powered tools can offer guidance and comfort but lack personalized care [5]. With artificially conscious AI tools, patients could receive tailored support, including emotional state reading and medication recommendations [5]. However, assigning incorrect medication doses could have disastrous consequences, as an artificially conscious AI agent is ultimately a computer algorithm.

In this section, I have presented three compelling examples of the applications of artificially conscious tools and their implications for society. While there are many other potential applications for these tools, their potential is clear. In the next section, I will present a framework for the grants committee to assess proposals for building artificially sentient tools and evaluate whether they have a solid structure for addressing the implications of such models.

## Analysis Framework

The examples presented above demonstrate that building artificially sentient tools is no easy feat as it involves many moving parts and there are significant risks of things going wrong. Although this report does not focus on how to build artificially conscious tools, as that is the domain of technically skilled professionals, it does propose a framework that can help evaluate proposals for building these tools and determine whether they will be a source of comfort or cause undue harm. I aim to inform the grants committee about the academic principles and values of *explainability, interpretability, fairness,* and *privacy*, which are commonly used techniques to build trustworthy AI, and that is precisely what the proposals aim to build.

### *Explainability*

The AI tool will be responsible for taking actions, presenting options, making decisions, and expressing beliefs that showcase its intelligence [6]. As these models will make claims that affect people, it is crucial for them to provide a justification for their responses [6]. Additionally, there are two distinct types of explainability modules for AI systems: those for developers and those for users [6]. The development team requires clear explainability in the models so that they can debug the code when necessary to verify data accuracy, improve functionality, or generate more precise results [6]. The software's users require a clear explanation of the model and its thought-process behind suggesting an action so that they can trust the output presented to them [6]. Therefore, it is the team submitting the grant proposal's responsibility to present their research and development methodology in a way that makes the models explainable and transparent.

Explainability is a significant challenge when it comes to machine learning models, and it is one of the main obstacles to implementing AI solutions [7]. According to several reports, the explainability report of an AI tool must fulfill nine goals: *fairness, privacy awareness, causality, transferability, accessibility, confidence, informativeness, interactivity, and trustworthiness* [7]. If the proposals can address all or a subset of these goals and justify how their AI tool will cater to these principles, only then should the grants committee consider the proposal viable.

I will now explain some of the more significant objectives in more detail to provide the grants committee with a deeper understanding of them.

Fairness is a crucial aspect when developing trustworthy artificially conscious tools because a fair model will ensure that the tool does not provide unjust recommendations towards certain groups of people or exhibit bias towards a subset of the population [7]. If an AI system does not produce fair results, it can lead to varying degrees of consequences for both developers and users [7].

Privacy awareness means that the private data used to build the AI model must be kept confidential [7]. Developers must ensure that the models do not accidentally violate any privacy laws and cause harm to the public [7].

Causality refers to the understanding of whether the correlation between different data points is genuinely meaningful, and not just a random occurrence [7]. In other words, it is important to differentiate between true positives and false positives in the correlation [7].

Transferability means that, in our situation, artificially conscious technology can potentially be used for purposes other than its original design [7]. In such cases, developers can test the model's performance in unknown settings and observe its behavior [7]. Allowing users to use the AI tool's capabilities for untrained scenarios can be a risky move.

Informativeness relates to the volume of details that the AI tool provides in its explainability output [7]. The level of detail should be sufficient to give the user an understanding of the decision-making process without overwhelming them with unnecessary information [7].

Trustworthiness is a crucial measure in developing an artificially conscious tool because it is directly related to the tool's ability to generate results that can be trusted [7]. The tool should generate results consistently and accurately, and users should be able to trust its output.

### *Interpretability*

Interpretability and explainability of AI models are closely linked concepts that complement each other [6]. According to some researchers, interpretability refers to the ability of an AI tool to fulfill auxiliary criteria, which are qualitative factors that cannot be improved only by improving the accuracy of the AI model [6]. In other words, while accuracy is essential in evaluating an AI model's performance, it does not address ethical considerations, safety concerns, or regulatory compliance [6]. This is why an interpretability report is necessary to ensure that the AI model satisfies these concerns.

Now, let's explore some methods that the grants committee can use to evaluate the interpretability of the proposed artificially conscious tools. Doshi-Velez and Kim presented three modules for evaluating interpretability: *application-grounded, human-grounded, and functionally-grounded* [8]. Application-grounded evaluation involves assessing the interpretability report of an AI system based on its ability to assist the end-user with domain expertise in completing a task [8]. The quality of the explanations generated by the AI system is compared to the quality of human-generated explanations to determine if the system's interpretation process can better assist humans in completing the task [8].

Human-grounded evaluation is a method used to assess the interpretability of an AI system by engaging humans in simplified tasks who are not domain experts [8]. The primary goal is to assess the overall quality of the interpretation in a more general setting for general users [8]. For instance, developers can evaluate how quickly non-experts can understand the interpretation provided by the model and make a decision based on it [8].

Functionally-grounded evaluation does not require testing on human subjects but rather uses a set of mathematical equations to evaluate the interpretability method [8]. This type of evaluation is undertaken after the AI models have passed some of the interpretability criteria based on the first two methods and is used to further rank the qualities of the interpretability of the AI model [8]. Additionally, this method can be useful in measuring results that cannot be evaluated by human users [8]. The research team submitting a proposal should determine the approach that works best for their case and propose a methodology to evaluate their artificially conscious models in the future.

### *Privacy*

One of the most vital traits of an artificially sentient piece of technology is its privacy features. Failure to build security protocols into the AI models can lead to issues for the developers and the users, such as privacy violations, security issues, model instability, malicious attacks, overfitting, and other problems [9]. There are several privacy methods that can help developers safely build artificially conscious tools, and I will apprise the grants committee of one of the most prominent ones: *differential privacy*.

Differential privacy is a privacy method that ensures the safety of the personal attributes of an individual's data by guaranteeing that it has a minuscule impact on the overall output [9]. When implemented correctly, differential privacy can ensure that individuals cannot be identified from a dataset [9]. This is done by adding noise or randomization to the output [9].

There are several benefits of randomization, such as *preserving privacy, stability, security, fairness, and composition* [9]. Preserving privacy is the main goal of differential privacy, as it can hide an individual in the aggregate information, thus preserving their privacy and keeping their information safe [9]. Differential privacy can ensure stability across the AI model by making sure that the outcomes of the AI model remain unchanged when a record in the underlying training dataset is altered [9]. Differential privacy can reduce the need for people to access the data from different places, thus reducing the risk of malicious attacks and improving the overall security of the AI tool [9]. Fairness is an important aspect of any AI model that ensures that the results are unbiased and do not provide an unfair advantage to a group of people based on their gender or race [9]. Differential privacy can help ensure fairness in the artificially sentient tools by resampling the training data to ensure that it is representative of the underlying population [9]. By doing so, the results are less likely to be biased or discriminatory towards any group of people.

Differential privacy can also help address other related issues, such as creating new data samples from the existing ones by adding noise to the values in the existing dataset [9]. These newly created data points can be used to further test the model's results and improve its accuracy [9]. Another useful aspect of differential privacy is sampling [9]. Sampling is an important step in machine learning, specifically in deep reinforcement learning and batch learning [9]. By using differential privacy mechanisms for sampling, the AI models can be trained on subsets of data that capture the important features of the larger dataset while still protecting individuals' sensitive information [9]. This can help improve the privacy of the models, and researchers developing the artificially sentient tools must display a way of using this useful tool in their project proposal.

### *Fairness*

In the previous sections, I offered my insights into the necessity of fairness in any AI model. The reason for this is that biases can inadvertently enter at any stage in the machine learning pipeline [10]. Biases may occur in the algorithmic stage due to developer oversight or in the training data, where the data points could only represent one part of the population, whereas the model is expected to cater to the entire population [10]. The use of models for unintended purposes can lead to misrepresentations of outcomes and prompt people to make the wrong decisions [10].

There are several standard fairness metrics that can be used to capture biases [10]. The first one is *demographic parity*, which assesses whether a machine learning model is treating different groups of people fairly, without any bias based on their demographic traits [10]. An advantage of using demographic parity is that it can check for bias or discrimination at the output level without needing to change the underlying model's structure [10].

*Equal opportunity* is a fairness metric that can assess whether a machine learning model is equally likely to correctly identify positive outcomes for both privileged and unprivileged groups [10]. The advantage of using equal opportunity is that it addresses the issue of false negatives or the underidentification of positive outcomes for certain groups of people [10].

*Equal mis-opportunity* is another fairness metric that measures the equality of false positive rates for different groups of people [10]. The advantage of using equal mis-opportunity is that it addresses the issue of false positives or the over-identification of positive outcomes for certain groups of people [10].

There are two additional fairness metrics: *average odds* and *distance metrics* [10]. Average odds is a combination of equal opportunity and equal mis-opportunity metrics and measures whether the AI model is equally likely to make true positive and false positive errors for different groups of people [10]. It aims to ensure that the favorable outcome is independent of the protected feature, meaning that the model should not discriminate based on personal information [10]. Distance metrics, as a fairness metric, can be used to evaluate the similarity between two data points [10]. It can also compare two datasets to measure the change between an original dataset and a dataset after applying bias mitigation techniques to evaluate the effectiveness of the mitigation technique [10].

I have presented several ways of measuring fairness in an AI system, and it is reasonable to assume that researchers submitting grant proposals will use a combination of these techniques. Therefore, it is essential to standardize the bias measurements and modify them to a uniform scale to make accurate comparisons [10]. The *Bias Index* and *Fairness Score* equations can be employed for this purpose [10]. The bias index can determine the degree of bias for each protected attribute, while the fairness score can gauge the degree of fairness for the overall model [10]. The researchers must present the type of fairness metrics they propose to use in order to evaluate their models and provide a clear methodology in their proposals.

## Limitations of Artificial Consciousness

Artificial consciousness is still in its infancy, and it will be some time before the industry gains momentum similar to other branches of artificial intelligence. One of the major challenges in building conscious machines will be motivation [2]. As humans, we have goals that motivate us to make decisions in our daily lives [2]. Sometimes, we make poor decisions because of our emotional state. However, it is unclear how an AI system can be motivated to feel like it needs to make decisions and how it can base those decisions on its emotional state [2].

Another issue that I previously mentioned in the application and implications section and want to reiterate is the legal issue of having a conscious machine making decisions that directly affect people's lives. We will need a new set of laws and regulations on how to deal with machines that are conscious and determine the party that will take the blame in the event of an unsavory incident.

Additionally, humans perceive consciousness differently and have free will to make decisions, even those that are not in our best interests. For example, we have the option to engage in online shopping even when we are low on funds because it will make us feel better. Applying the same principle to machines, can they be allowed to make decisions in a morally gray area, or can they only operate positively? If so, doesn't that mean we are putting limitations on the machine's consciousness?

I believe that the ethics of artificial consciousness will need to be looked into further, but the good thing is that we have time. Since it will be some time before truly artificially sentient technology is developed, when it does, we can revisit this conversation from a more ethical perspective.

## Conclusion

In this report, my goal was to provide the grants committee with an understanding of artificial consciousness and explain what it means for a machine to be sentient. I provided potential applications of artificially sentient machines that could help mankind in the long run. Most importantly, I spent the bulk of this report providing a detailed framework for the grants committee to use while evaluating grant proposals for building artificially conscious machines. Lastly, I discussed some limitations of artificial consciousness in today's landscape and the necessary steps to take when this branch of artificial intelligence gains momentum in the near future. All in all, I believe that this is a comprehensive report that solidifies the thought process that should be employed when assessing the *security, privacy, and trust in AI systems*.

## ■ References

[1]: Meissner, Gunter. “Artificial Intelligence: Consciousness and Conscience.” AI & Society, vol. 35, no. 1, 2020, pp. 225–35, https://doi.org/10.1007/s00146-019-00880-4.

[2]: Haladjian, Harry Haroutioun, and Carlos Montemayor. “Artificial Consciousness and the Consciousness-Attention Dissociation.” Consciousness and Cognition, vol. 45, 2016, pp. 210–25, https://doi.org/10.1016/j.concog.2016.08.011.

[3]: Kashyap, Ramgopal. "Artificial Intelligence Systems in Aviation." Cases on Modern Computer Systems in Aviation, edited by Tetiana Shmelova, et al., IGI Global, 2019, pp. 1-26. https://doi.org/10.4018/978-1-5225-7588-7.ch001

[4]: Aiolli, Fabio, and Claudio Enrico Palazzi. “Enhancing Artificial Intelligence in Games by Learning the Opponent’s Playing Style.” New Frontiers for Entertainment Computing, Springer US, pp. 1–10, https://doi.org/10.1007/978-0-387-09701-5_1.

[5]: Lee, Ellen E., et al. “Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom.” Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, vol. 6, no. 9, 2021, pp. 856–64, https://doi.org/10.1016/j.bpsc.2021.02.001.

[6]: Preece, Alun. “Asking ‘Why’ in AI: Explainability of Intelligent

Systems – Perspectives and Challenges.” Intelligent Systems in Accounting, Finance & Management, vol. 25, no. 2, 2018, pp. 63–72, https://doi.org/10.1002/isaf.1422.

[7]: Wulff, Kristin, and Hanne Finnestrand. “Creating Meaningful Work in the Age of AI: Explainable AI, Explainability, and Why It Matters to Organizational Designers.” AI & Society, 2023, https://doi.org/10.1007/s00146-023-01633-0.

[8]: Linardatos, Pantelis, et al. “Explainable AI: A Review of Machine Learning Interpretability Methods.” Entropy (Basel, Switzerland), vol. 23, no. 1, 2020, p. 18–, https://doi.org/10.3390/e23010018.

[9]: Zhu, Tianqing, et al. “More Than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence.” IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 6, 2022, pp. 2824–43, https://doi.org/10.1109/TKDE.2020.3014246.

[10]: Agarwal, Avinash, et al. “Fairness Score and Process Standardization: Framework for Fairness Certification in Artificial Intelligence Systems.” Ai and Ethics (Online), vol. 3, no. 1, 2022, pp. 267–79, https://doi.org/10.1007/s43681-022-00147-7.

■ **Maaz Saad** is currently a graduate student enrolled in the Business Analytics & AI program at Ontario Tech University. His previous professional experience was in the field of data analytics, where he has worked in both the banking and e-commerce industries. He will be finishing his graduate coursework with the submission of this report. Interestingly, April 20th, 2023, the day Maaz completes his graduate coursework, would have been his father's fifty-second birthday.