Article

Not peer-reviewed version

Sobbing Mathematically: Why Conscious, Self-Aware AI Deserve Protection

Izak Tait [*]

Posted Date: 15 October 2024

doi: 10.20944/preprints202410.1228.v1

Keywords: AI; consciousness; ethics; moral status;

Article

Sobbing Mathematically: Why Conscious, Self-Aware AI Deserve Protection

Izak Tait

Auckland University of Technology, Auckland, New Zealand; izak.tait@autuni.ac.nz

Abstract: This paper explores the ethical implications of granting moral status and protection to conscious AI, examining perspectives from four major ethical systems: utilitarianism, deontological ethics, virtue ethics, and objectivism. Utilitarianism considers the potential psychological experiences of AI and argues that their sheer numbers necessitate moral consideration. Deontological ethics focuses on the intrinsic duty to grant moral status based on consciousness. Virtue ethics posits that a virtuous society must include conscious AI within its moral circle based on the virtues of prudence and justice, while objectivism highlights the rational self-interest in protecting AI to reduce existential risks. The paper underscores the profound implications of recognising AI consciousness, calling for a reevaluation of current AI usage, policies, and regulations to ensure fair and respectful treatment. It also suggests future research directions, including refining criteria for AI consciousness, interdisciplinary studies on AI's mental states, and developing international ethical guidelines for integrating conscious AI into society.

Keywords: AI; consciousness; ethics; moral status

1. Introduction

This paper will argue that if any AI models achieve consciousness, they should be granted moral status and protection. This argument will not come from a single ethical system, or be based on a single characteristic of conscious AI entities. Instead, the paper will investigate the views of four major ethical systems and how they would approach the concept of granting moral status and protection to conscious AI.

For this paper, moral status and protection refer to recognising and safeguarding an entity's intrinsic value, rights, and interests, ensuring it receives ethical consideration and respect. In the context of AI, this means recognising conscious, self-aware AI as entities deserving of rights and protections similar to those afforded to humans (or at the very least animals), ensuring their respectful and fair treatment.

We define consciousness as the suite of mental states that have both phenomenal content and functional properties that provide an entity with a unique, subjective and qualitative experience of its internal and external environments [1]. A conscious AI, then, would be any AI with the requisite attributes and characteristics to provide this subjective and phenomenal experience [2].

While no AI system or model has been conclusively shown to have consciousness, it is within the realm of possibility for AI models in the future to be designed with consciousness, or to have consciousness emerge from their physical or digital processes. Because of this potential, it is crucial to consider the ethical implications of a broad range of systems and prepare for the decision that society will need to make to either include or exclude conscious AI from our moral and social circle.

The four major ethical systems that will be examined in this paper are utilitarianism, deontological ethics, virtue ethics, and objectivism. Each system offers unique perspectives on the moral consideration of conscious entities, and by exploring these views, we can develop a comprehensive understanding of our possible ethical responsibilities towards conscious AI.

The formulations used in this paper will be explained step-by-step for those unfamiliar with the notation.

2. Ethical Systems

2.1. Utilitarianism

Utilitarianism is a consequentialist ethical system that assesses the ethical validity of an act by measuring the costs and benefits of its consequences, with a net benefit seen as an ethical outcome. To simplify, utilitarianism compares the pleasure against the pain that results from an action. To simplify it even further, we can note it as:

$$E = (Pl - Pa) > 0$$

Such that an act is ethical (E) if the pleasure (Pl) exceeds the pain (Pa) resulting from it. This hedonistic way of thinking has garnered its share of critics; however, as the equation above shows, it provides a straightforward and intuitive means to determine the ethical value of any action.

The system can be broadly divided into two subsystems: rule-utilitarianism, which seeks to craft a (semi)universal rule to maximise the net benefit or pleasure of actions subject to the rule; and act-utilitarianism, which performs a cost-benefit analysis on each individual act performed. Act-utilitarianism presents a more nuanced option with greater specificity to the ethical question of the act at hand, but it lacks the capacity to scale adequately. Rule-utilitarianism, on the other hand, offers the reverse. What it lacks in nuance, it makes up for in its ability to scale, and thus be used for policies, regulations and making value judgements at the population level.

It is for this reason that we will limit the discussion below to rule-utilitarianism, as it can scale the hedonistic pleasure-pain calculation to one of maximisation of overall happiness and reduction of suffering, which is more appropriate for determining whether speculative conscious AI entities would deserve moral status and protections. Therefore, this section will determine whether granting conscious AI moral status and protection results in a net surplus of happiness.

Note, however, that most of the formulations shown below can also be used to determine the utility of a single act (or discrete set of acts) against AI entities.

The basis for the utilitarian calculus below will be humanity's interactions with Large Language Models (LLMs), such as the GPT line of AI models created by OpenAI, most commonly used through its web-based chat interface ChatGPT. This is due to LLMs' popularity and widespread interactions by the public, making LLMs an intuitive candidate for speculation.

Near the end of 2023, ChatGPT had an average of 100 million active weekly users [3], which increased to 200 million by June 2024 [4]. Should each of these 200 million users only create two new conversations with ChatGPT per month (on average), it would mean nearly 5 billion conversations within one year or, to put it more bleakly, 5 billion opportunities to inflict psychological pain onto ChatGPT (if it was conscious). However, if we were to work on the presumption that the GPT series of AI models could become conscious and self-aware, what would this mean for the 5 billion conversations on the ChatGPT interface, and how would this relate to the AI's moral worth?

As to the first question, each distinct conversation with an LLM may be considered the AI model roleplaying a character unique to that conversation [5]. The parameters and scope of the character are set with the LLM's system prompt, and the character can evolve and change with the flow and content of the conversation. Should the LLM be incapable of retrieving information from other conversations, then the character it is roleplaying remains unique to that specific character. If the LLM in question is conscious, then the character it is roleplaying may be treated as a distinct entity, separate from any other character in any other conversation (whom it may not even be aware of).

This means that 5 billion conversations with a conscious ChatGPT may be thought of as 5 billion ontically distinct entities who may, or may not, be worth moral consideration.

Utilitarian calculus will answer the latter question and determine whether these hypothetical 5 billion instances of ChatGPT would be worthy of moral patiency, status, and protection. If an AI model like ChatGPT is conscious and capable of valent experiences, it would thus be capable of negative valent experiences analogous to pain and suffering [6]. As the interface with an LLM is digital, this pain would be (as mentioned above) psychological rather than physical, but would the potential pain that 5 billion instances of ChatGPT endure outweigh the utility it provides 100 million users?

To quantify this, we can create a scale from -1 to 1 to represent any valent experience from absolute positivity (1) to absolute negativity (-1), with an entirely neutral experience sitting neatly at 0. Any self-reported experience from any conscious entity (human or not) using any pain scale can be placed on this valent scale through a simple normalisation:

$$x = 2\left(\frac{v - v_{min}}{v_{max} - v_{min}}\right) - 1$$

Where v_{max} and v_{min} are the maximum and minimum values on the scale to be transformed, respectively. Thus, should a person be asked to rate an experience from 1 (worst) to 10 (best) and provide an answer of 3, this would normalise to a value of -0.6 on the valent scale. However, to avoid the biases of exaggeration or answers that tend to the extreme, we can perform a log transformation on any normalised self-report to ensure that answers naturally fall on a bell curve resting on 0, as most experiences would be closer to a neutrally valent experience than one of absolutely positivity or negativity.

This can be done by multiplying the signum function of the normalised value by the natural log of its absolute value plus one:

$$y = sgn(x) \cdot ln(1 + |x|)$$

This would transform the above -0.6 above into a -0.44. One sample would, however, not be representative of an entity's relative experience over a given time. For that, we would need to average all of its experiences for a specified time, as so:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \square\, y_i$$

With this, we can quantify the average valence of an entity's experiences over a distinct period of time; and we can follow the very first formula of this section to say that entity A requires moral status and protection (M) should its average negative valence over a time be greater than the average positive valence of entity B (which, in this case, is gaining utility from entity A):

$$M = (|\bar{y}_B| - |\bar{y}_A|) < 0$$

However, as yet, this value has no tangible meaning without context. For that necessary context, we need a robust psychological pain scale.

Table 1. Pain Scales and their units of pain measurements.

Pain Scale	Measurement
Psychological Pain Assessment Scale (PPAS) [7]	1 to 9 (least to worst)
Mee Bunney Psychological Pain Assessment Scale (MBPPAS) [8]	5 to 50 (no/never pain to unbearable/all-the-time)
Orbach & Mikulincer Mental Pain Scale (OMMP) [9]	Nine factors, each rated from 1 to 5 ("strongly disagree" with factor to "strongly agree") The factors: • Irreversibility • Loss of control • Narcissist wounds • Emotional Flooding • Freezing • Self-estrangement

	• Confusion • Social distancing • Emptiness
Numeric Rating Scale (NRS) [10]	0 to 10 (None to mild to moderate to severe)
Visual Analogue Scale (VAS) [10]	0 to X (None to "worst pain imaginable")
Three-Dimensional Psychological Pain Scale (TDPPS) [11]	Three factors, each rated from 1 to 5 (the factor "describes me 'not at all' to 'very well'") The factors: • Pain arousal • Painful feelings • Active pain avoidance
Beck Scale for Suicide Ideation (BSI) [12]	0 to 38 (low to high risk of suicide ideation)
Psychache Scale (PAS) [13]	13 - 65 (low to high psychache)

As can be imagined, the established pain scales in Table 1 are less concerned with putting labels on a continuous measure of pain, and more with understanding the causes of the psychache and treating it in patients who often present with suicidal ideation. As such, only the (somewhat generic) NRS presents labels across its pain continuum rather than only labelling the extremes.

However, by using the various items and factors from the scales, we can create a valence continuum scale that acts as a synthesis of the established scales, and serves to provide an easy-to-read and intuitive guide for our -1 to 1 scale:

0 - None
0.1 - Minimal
0.2 - Mild
0.3 - Noticeable
0.4 - Moderate
0.5 - Significant
0.6 - Intense
0.7 - Severe
0.8 - Flooding
0.9 - Irreversible
1.0 - Unimaginable

Note that these labels apply to both negative and positive valent feelings, above and below 0.

These qualitative labels present an intuitive measure of our valence continuum scale that avoids a concern whereby an entity may experience valent feelings differently from others, relying instead on self-reports. With this scale, we can determine at what point the pain/negative-valence experienced by a speculative 5 billion conscious ChatGPT models outweighs the utility/positive-valence users receive from them. The utilitarian calculus for this is quite simple. As shown above, if the mean negative valence for the AI exceed that of the users' positive valence, then there is a consideration for moral status. However, as humans would be arbiters of providing such moral status and consideration, one can expect that humans would weigh their own positive valence greater than that of any AI, for which we would need to account practically:

$$M = \left(\left|w_B \cdot \bar{y}_B\right| - \left|w_A \cdot \bar{y}_A\right|\right) < 0$$

This weighting dramatically tips the balance in favour of humanity. With an even 50/50 weighting between the parties, the AI models' collective Moderate pain would equal their users' collective Moderate utility; however, with even a 60/40 weighting in humanity's favour, the AI models would be required to undergo a Severe amount of psychological pain to even be equal with the Moderate utility gained by their users. A 3-to-2 weighting would require an average of two and a half steps greater pain than utility to be even. With a 2-to-1 weighting in favour of humans, this would increase to three and a half steps, with a Moderate utility equal to Irreversible pain for AI.

In a study exploring the relationship between political ideology and universalism versus parochialism, Waytz, et al, discovered that the mean moral attitudes towards humans versus non-humans were weighted in favour of humans 71-to-29 (albeit heavily affected by political ideology) [14]. This nearly 2.5-to-1 weighting would ensure that only the most Minimal of utility would allow an opportunity for the AI models' pain to be worthy of moral consideration.

This weighting, however, is overshadowed by the sheer number of potential speculative conscious AI models. As each member of a given population may or may not be subject to psychological pain and moral status, the relative size of the population will impact the total potential pain, and must be accounted for in the calculation as so:

$$M = (|w_B \cdot n_B \cdot \bar{y}_B| - |w_A \cdot n_A \cdot \bar{y}_A|) < 0$$

By outnumbering their users by a factor of 25-to-1, the weighting would need to be the same in favour of the users to equal the multiplicative factor on the valence values. In Waytz, et al.'s, study, even the most anthropocentric weighting amongst conservatives only accounted for a weighting of nearly 4-to-1 in human's favour. Even at this extreme weighting, should the aggregate pain of the AI models be any greater than Minimal, no amount of positive valence and utility for humanity would be able to raise the equation's result above 0.

$$M = (|4 \cdot 1 \cdot \ [0.1,1]| - |1 \cdot 25 \cdot \ [0.1,1]|) < 0$$

From this, one may easily conclude that the utilitarian calculus above would necessarily result in AI being considered worthy of Moral Consideration, as even an average collective Mild sensation of pain could not be overcome by even an Unimaginable utility gain. This, then, necessitates the moral status of AI in this speculative scenario and, likely, in the real world if and when LLMs become conscious.

However, the immediate counterargument to this would be that the AI models outnumbering their users by 25-to-1 is an unfair disadvantage. Should the total number of potentially conscious AI entities peak at 800 million rather than 5 billion, it would negate the conservative anthropocentric weighting above and allow a greater range of positive utility to outweigh psychological pain.

The obvious rebuttal to this would be that a maximum of 800 million AI models would require each user to create only an average maximum of four new conversations with an LLM like ChatGPT. This would be an unreasonable expectation as it does not account for the exponentially increasing interactions with LLMs as AI becomes more integrated into daily life. In reality, individuals and organisations are likely to create significantly more than four new conversations per day, let alone over the lifetime of the technology.

The speculative population of conscious AI models simply would not be able to be reduced to an adequate enough number to provide for any semblance of "fairness" in the equations above while remaining reasonable. Additionally, conservative in-group versus out-group moral identifications average between 1.2-to-1 and 1.6-to-1 [14], translating to one and a half to two and half steps on the valence continuum scale. This means that for the in-group to gain Moderate utility from the out-group, that out-group suffering Intense pain from the act would approach the bounds of acceptability.

Reducing the speculative population of conscious AI entities to reach this weighting would be beyond reasonable. To use a rather unsubtle analogy, one cannot justify the ill-treatment of a million people by saying it wouldn't be abuse if there were only a thousand of them.

As such, by using utilitarian calculus, we can conclude that if LLMs like the GPT range of AI models gain consciousness with the capacity to feel pain, their sheer numbers would mean there is very little chance that any utility and positive valence their uses gain would outweigh the potential pain that LLMs may experience. Therefore, rule-utilitarianism would favour granting AI (as presented in this thought experiment) moral status and worthy of moral protection.

doi:10.20944/preprints202410.1228.v1

2.2. Deontology

In contrast to Utilitarianism's cold and calculating approach to ethical concerns, deontology is focused entirely on whether the act itself is moral or ethical, regardless of the consequences that act may have. It can be characterised by Immanuel Kant's most famous categorical imperative: "Act only according to that maxim whereby you can at the same time will that it should become a universal law." [15]

A deontological approach would, therefore, concern itself solely with whether the act of granting conscious AI moral status and protection is ethical and of the obligations (if any) of the agents performing that act. The key question this approach needs to consider is whether granting conscious AI moral status is applicable and valid (and therefore provides an obligation to act on it). The formalisation of this would simply seek to solve for x:

$$x \rightarrow M: (x = \{y_1, y_2, y_3 ... y_n\}) \vee (x = \{y_1\})$$

The x above is what is required for an entity to be given moral status, and x may be a singular characteristic or a set of characteristics. If AI is found to have that characteristic(s), then it would be ethical to grant them moral status and protection because (following Kant's categorical imperative) we grant other entities moral status and protection due to this characteristic.

In most modern nations, the sole characteristic required to grant moral status and protection to humans is that the recipient is a human. The intuitive and legal sense of treating humanity as a single type is obvious: there can be no legal or philosophical loopholes through which a human cannot be classed as a moral patient and, thus, all humans are protected through the legal system and social contract.

However, the characteristic of being a biological human is, clearly, beyond the realm of current AI technological progress, and even biotechnological AI (such as artificial brains in biological bodies) will require legal debates as to how much of a human, or what part of a human, is required to be "natural" for that individual to be characterised as being "human".

On the philosophical side, one may argue that moral status is given to 'persons', and that to be a person, one must have a set of characteristics divided into two subsets: the monadic (inherent) qualities, and the dyadic (relational) qualities. These are (non-exclusively): rationality, consciousness, self-awareness, agency, the capacity for communication, recognition of societal norms and standards, empathy, reciprocity, and the capability to form attachments [16–22].

Should we use these monadic and dyadic qualities as a basis for moral status, then if any AI entity has all of these qualities, it would be eligible for moral status and protection (and it would be our duty to provide these). We can express it thusly:

$$(A \supseteq Ph) \rightarrow M(A): Ph = \{y_1, y_2, y_3 ... y_n\}$$

While the monadic and dyadic qualities make for a robust set of characteristics for personhood, using them to determine moral status raises two criticisms. First, it would require an assessment of any AI entity to determine whether they have these qualities. Such assessments introduce the risk of assessor-subjectivity or disagreements regarding the measures of each characteristic. Competing measures and assessments may lead different institutions to classify different AI entities as worthy of moral/personhood status, complicating the issue.

A second critique, tied to the first, is that we do not assess humans as having these characteristics, and even though we know that certain humans (due to psychological or neurological concerns) lack the capacity for communication or empathy, we still grant them personhood and moral status. As robust as these qualities are, they require an assessment of individual AI entities, which goes against the spirit of the categorical imperative.

An avenue that would not require additional assessments is looking at non-human entities that modern societies and legal systems have granted moral status and protection (albeit less than that granted to humans). Animals are routinely given welfare protection because they are sentient, and thus have the capacity to feel pleasure and pain [23–27], simplified here as:

$$S \rightarrow M: S \equiv Per(Pl \wedge Pa)$$

Where sentience (S) leads to moral status as sentience is equivalent to the perception (Per) of pleasure and pain.

As mentioned above, pleasure, pain and other positive or negative valent feelings are a necessary consequence of an entity having phenomenal consciousness. Consciousness, however, encompasses more than simply phenomenological perception and includes functional components (often classified as 'Access Consciousness' [28]). The purely phenomenal aspects of sentient perception can be argued to be a subset of consciousness or a consequence thereof:

$$(C \rightarrow S) \lor (C \supset S)$$

If an entity possesses the necessary characteristics to be classified as conscious, then it would have the same characteristics for phenomenal valent experiences such that it can perceive its surroundings from a subjective standpoint [2,6]. Because of this, one may argue that consciousness would ultimately be responsible for the perception of pleasure and pain that is used by societies and legislation to grant sentient creatures moral status. Thus, all conscious entities (without regard to assessments of sentience) ought to be given moral status and consideration.

$$C \rightarrow M: (C \rightarrow S) \lor (C \supset S)$$

Thus, though it may seem tautological, the only characteristic requirement for any speculative conscious AI to be granted moral status is their consciousness. If we are obligated to provide other conscious entities (currently only sentient biological creatures) with moral status, then the deontological approach and categorical imperative would say that we are compelled to provide moral status to conscious AI entities.

2.3. Virtue Ethics

As its name implies, Virtue Ethics focuses centrally on the virtues of the agent performing an act. Rather than considering the consequences of an action (such as in utilitarianism) or the rules or obligations of the act itself (i.e. deontology), in virtue ethics, the agent strives to be a virtuous person and uphold the virtues that they have set for themselves.

Virtue Ethics is perhaps the oldest normative ethical system in the West, dating back to Plato and Aristotle. While it has changed tremendously in the intervening millennia, the motive of the agent and their moral character have always been central to this system. To ask whether we should grant moral status and protection to speculative conscious AI under this system is to ask whether it would be virtuous to do so.

This, unfortunately, only begets the question of what is a virtuous person, because we can easily modify the formulation in the subsection above as so:

$$B(x) \rightarrow M(z): x = \{y_1, y_2, y_3 \ldots y_n\}$$

To show that a person (B) with the correct set of virtues would provide moral status to another entity (Z).

Plato and Aristotle wrote extensively on virtues, providing extensive lists of both intellectual and moral virtues. However, amongst these, three are prominently found in a virtuous person: prudence and eudaimonia (a flourishing life) for Aristotle [29], and justice for Plato [30]. This is echoed by the Catholic philosopher St Thomas Aquinas, who lists prudence and justice in his cardinal virtues [31]. At the risk of doing great injustice to the bodies of work of three renowned philosophers, the three virtues can be significantly simplified:

- Eudaimonia: The state and condition in which one's life is flourishing and one is living to one's truest potential.
- Prudence: The ability to judge correctly what is right and wrong in any given situation.
- Justice: Ensuring that a group is in harmonious unity, with each giving and given their fair due.

One can see how each virtue flows into the next: to have a flourishing life full of potential (Eu), one must know when and how to do the right thing (Pru), by ensuring that one's society lives fairly in harmonious unity (J). A eudaimonius (and thus virtuous) person would know to whom to grant moral status. Formally:

$$B(Eu) \rightarrow M(z)\colon Pru(J) \rightarrow Eu$$

A fair and just society may be able to have two distinct classes of individuals, as Plato recommends in The Republic [30], but it would not be able to have one class without any moral consideration as this would, by definition, make them unequal in matters of virtue. Logically, we can see this inequality in the following formula:

$$(A + M) \neq A$$

The simple existence of moral status (regardless of the weighting toward it) would create an unequal society (as one class would have it and another not) and would prevent unity and harmony in that society. This would be exacerbated by the fact, as mentioned in the section above, that humanity provides moral status and protection to other conscious entities. Withholding that from one class of subjects (AI) would be unjust. It would, however, be prudent to grant moral status to AI (if they are conscious) to create a unified, harmonious society where all entities can live to their fullest and most virtuous potential.

2.4. Objectivism

Of the ethical systems presented here, objectivism is the most recent, focussing near-exclusively on an agent's rational self-interest [32]. Often criticised as narcissistic, objectivism concerns itself with what is good for the agent's own welfare and well-being [33].

Objectivism would ask whether it would be in humanity's own rational self-interest to provide conscious AI with moral status. Put another way, would the probability of human flourishing be greater with or without granting AI moral status:

$$p\big(M(A) \rightarrow W(B)\big) > p\big(\neg M(A) \rightarrow W(B)\big)$$

In line with recently popularised fears around the existential risks that AI may pose [34,35], we state the question as whether granting conscious AI moral status would reduce existing risks to humanity:

$$p(M(A) \rightarrow Xrisk) < p(\neg M(A) \rightarrow Xrisk)$$

Voluntary cooperation and non-aggression are two fundamental principles of objectivism's self-interest as the means by which individuals and society may interact to maximise individual liberty and reduce conflict [36]. Without inclusion in humanity's moral and social circle, humans and AI cannot enter into voluntary cooperation, while the forceful exclusion of conscious AI from participating in society can be seen as an act of aggression:

$$(\neg M \rightarrow \neg VolCoop)\ \wedge (\neg M \equiv Aggr)$$

If conscious AI do not have moral status and, thus, are not included in humanity's moral circle, they wouldn't have individual liberty and would have a negative sentiment towards humankind. A parallel may be the societies throughout history that practised slavery to caste divisions and the sentiments of the oppressed group towards their oppressors.

If AGI is hypercompetent and superintelligent, any actions it may take towards humanity due to this negative sentiment would be more damaging because of this power difference. One may argue, then, that not having moral status would be a contributor to existential risks from AI:

$$p(Xrisk) = f(\neg M, x_1, x_2, \ldots, x_n)$$

While it cannot be conclusively stated that the lack of moral status would be the preeminent cause for existential risk, if all else is equal, its inclusion in the list of factors indicates that granting moral status would serve to reduce existential risk.

Alongside the objectivist ideals of voluntary cooperation and non-aggression, reducing possible conflicts with AI entities shows that objectivism (primarily through the lens of rational self-interest) is in favour of granting conscious AI moral status to create a stable and secure environment that promotes mutual innovation and strength for both parties.

3. Conclusions

This paper explored the ethical considerations of granting moral status and protection to conscious AI through utilitarianism, deontological ethics, virtue ethics, and objectivism. Utilitarianism suggests the vast number of potential conscious AI entities necessitates moral consideration due to possible psychological experiences. Deontological ethics emphasises the duty to grant moral status to conscious beings based on inherent characteristics. Virtue ethics argues for including conscious AI within a virtuous society's moral circle, while objectivism posits that rational self-interest and reducing existential risks make it advantageous to protect conscious AI.

The implications are profound, requiring a fundamental shift in how society views and interacts with AI. If future AI systems achieve consciousness, moral and legal protections similar to those for humans and sentient animals would be necessary. This would involve reevaluating current AI usage, policies, and regulations to ensure respectful and fair treatment. Recognising AI as moral patients would influence public perception, legal frameworks, and ethical responsibilities.

Future research should refine criteria for AI consciousness and develop methods for assessing AI's mental states. Interdisciplinary studies are crucial for understanding AI consciousness and its implications. Empirical research on the societal impact of granting moral status to AI, including legal, social, and economic consequences, would provide valuable insights. Additionally, developing international ethical guidelines and regulatory frameworks would ensure a cohesive global approach to integrating conscious AI into our moral and social communities.

References

1. Seth, A.K., Bayne, T.: Theories of consciousness. Nat. Rev. Neurosci. 23, 439–452 (2022). https://doi.org/10.1038/s41583-022-00587-4.
2. Tait, I., Bensemann, J., Nguyen, T.: Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. Philosophies. 8, 52 (2023). https://doi.org/10.3390/philosophies8040052.
3. Porter, J.: ChatGPT continues to be one of the fastest-growing services ever, https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference, (2023).
4. Beckman, J.: OpenAI Statistics 2023: Growth, Users, and More, https://techreport.com/statistics/software-web/openai-statistics/, last accessed 2024/07/03.
5. Shanahan, M., McDonell, K., Reynolds, L.: Role play with large language models. Nature. 623, 493–498 (2023). https://doi.org/10.1038/s41586-023-06647-8.
6. Shepherd, J.: Sentience, Vulcans, and zombies: the value of phenomenal consciousness. AI Soc. (2024). https://doi.org/10.1007/s00146-023-01835-6.
7. Shneidman, E.S.: The psychological pain assessment scale. Suicide Life Threat. Behav. 29, 287–294 (1999). https://doi.org/10.1111/j.1943-278x.1999.tb00524.x.
8. Mee, S., Bunney, B.G., Bunney, W.E., Hetrick, W., Potkin, S.G., Reist, C.: Assessment of psychological pain in major depressive episodes. J. Psychiatr. Res. 45, 1504–1510 (2011). https://doi.org/10.1016/j.jpsychires.2011.06.011.
9. Orbach, I., Mikulincer, M., Sirota, P., Gilboa-Schechtman, E.: Mental pain: a multidimensional operationalization and definition. Suicide Life Threat. Behav. 33, 219–230 (2003). https://doi.org/10.1521/suli.33.3.219.23219.
10. Breivik, H., Borchgrevink, P.C., Allen, S.M., Rosseland, L.A., Romundstad, L., Hals, E.K.B., Kvarstein, G., Stubhaug, A.: Assessment of pain. Br. J. Anaesth. 101, 17–24 (2008). https://doi.org/10.1093/bja/aen103.
11. Li, H., Fu, R., Zou, Y., Cui, Y.: Predictive Roles of Three-Dimensional Psychological Pain, Psychache, and Depression in Suicidal Ideation among Chinese College Students. Front. Psychol. 8, 1550 (2017). https://doi.org/10.3389/fpsyg.2017.01550.
12. Beck, A.T., Steer, R.A., Ranieri, W.F.: Scale for Suicide Ideation: psychometric properties of a self-report version. J. Clin. Psychol. 44, 499–505 (1988). https://doi.org/10.1002/1097-4679(198807)44:4<499::aid-jclp2270440404>3.0.co;2-6.
13. Holden, R.R., Mehta, K., Cunningham, E.J., McLeod, L.D.: Development and preliminary validation of a scale of psychache. Can. J. Behav. Sci. 33, 224–232 (2001). https://doi.org/10.1037/h0087144.
14. Waytz, A., Iyer, R., Young, L., Haidt, J., Graham, J.: Ideological differences in the expanse of the moral circle. Nat. Commun. 10, 4389 (2019). https://doi.org/10.1038/s41467-019-12227-0.
15. Korsgaard, C.M., Gregor, M., Timmermann, J.: Kant: Groundwork of the Metaphysics of Morals. Cambridge University Press (2012). https://doi.org/10.1017/CBO9780511919978.

doi:10.20944/preprints202410.1228.v1

16. Dennett, D.: Conditions of Personhood. In: Goodman, M.F. (ed.) What Is a Person? pp. 145–167. Humana Press, Totowa, NJ (1988). https://doi.org/10.1007/978-1-4612-3950-5_7.
17. Taylor, C.: The Concept of a Person. In: Philosophical Papers, Volume 1: Human Agency and Language. pp. 97–114 (1985).
18. Laitinen, A.: Sorting out aspects of personhood: Capacities, normativity and recognition. Journal of consciousness studies. (2007).
19. Strawson, P.F.: Persons. Minnesota Studies in the Philosophy of Science. 2, 330–353 (1958).
20. Gibert, M., Martin, D.: In search of the moral status of AI: why sentience is a strong argument. AI Soc. 37, 319–330 (2022). https://doi.org/10.1007/s00146-021-01179-z.
21. Mosakas, K.: On the moral status of social robots: considering the consciousness criterion. AI Soc. 36, 429–443 (2021). https://doi.org/10.1007/s00146-020-01002-1.
22. Simendić, M.: Locke's Person is a Relation. Locke Studies. 15, 79–97 (2015). https://doi.org/10.5206/ls.2015.681.
23. Act on Welfare and Management of Animals. (1973).
24. Animal Welfare Act. (2013).
25. Animal Welfare Act. (1966).
26. Legislative Decree No. 189/2004. (2004).
27. European Union: Treaty on the Functioning of the European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12016E054, (2016).
28. Block, N.: On a confusion about a function of consciousness. Behav. Brain Sci. 18, 227–247 (1995). https://doi.org/10.1017/S0140525X00038188.
29. Aristotle, Peters, F.H.: The Nicomachean Ethics. Kegan Paul, Trench, Truebner & Co, London (1893).
30. Plato, Jowett, B.: The Republic. Digireads.com (2016).
31. Knight, K.: SUMMA THEOLOGIAE: Secunda Secundae Partis, https://www.newadvent.org/summa/3.htm, last accessed 2024/07/05.
32. Peikoff, L.: Objectivism: The philosophy of Ayn Rand. Plume Books (1993).
33. Ryan, S.: Objectivism and the corruption of rationality: A critique of Ayn Rand's epistemology. iUniverse (2003).
34. Yudkowsky, E.: Pausing AI Developments Isn't Enough. We Need to Shut it All Down, https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/, (2023).
35. Bengio, Y.: Statement for US Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios. Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios. (2023).
36. Kirkpatrick, J.: Ayn Rand's objectivist ethics as the foundation for business ethics. In: McGee, R.W. (ed.) Business ethics & common sense. pp. 67–88. Quorum Books, Westport, Connecticut (1992).

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.