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

Izak Tait *

Posted Date: 29 December 2025

doi: 10.20944/preprints202410.1228.v2

Keywords: AI; consciousness; ethics; moral status

 doi:10.20944/preprints202410.1228.v2

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 (CAI). 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 CAI, and summarise them in such a way as to be presentable to non-ethicists.

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.

Consciousness here is defined 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 (Seth & Bayne, 2022). A CAI, then, would be any AI with the requisite attributes and characteristics to provide this subjective and phenomenal experience (Tait et al., 2023).

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 CAI 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 CAI.

 doi:10.20944/preprints202410.1228.v2

Numerous publications in recent decades have spoken extensively on the need for, and frameworks of, recognising the moral status of artificial agents and granting these entities legal rights. This paper will not retread this path (reviewed in depth by Harris and Anthis (Harris & Anthis, 2021)), but instead focus on the logical and mathematical formulations of the four ethical systems below. Readers are greatly encouraged to pursue long-form philosophical discussions on this topic, and are suggested as a first avenue of inquiry to peruse the works of Coeckelberg, Gunkel, and Schwitzgebel. Exemplary pieces include (Coeckelbergh, 2010, 2012; Gunkel, 2012, 2021, 2023; Schwitzgebel & Garza, 2015, 2020).

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

2. Ethical Systems

All of the ethical systems below will be based on the same core predicates, such that they may be compared and contrasted in a consistent manner. To that end, each ethical system will be evaluated by the policy it sets forth, and for each, we will single policy choice $\pi \in \Pi_{status} = \{grant, deny\}$. For each choice π, let $(\Omega, \mathcal{F}, \mathbb{P}_\pi)$ be the probability space of possible worlds stemming for that choice, with

- Ω as the set of all possible worlds based on π, and $\omega \in \Omega$ a single world outcome,
- $\mathcal{F}$ as the sigma-algebra of events over Ω, and the set of admissible events, and
- $\mathbb{P}_\pi$ as the credence measure over $(\Omega, \mathcal{F})$ given current evidence.

For this, all expectations below are $\mathbb{E}_\pi[\cdot] := \int_\Omega (\cdot) d\mathbb{P}_\pi$.

Let Hum be the set of humans and CAI the set of conscious AI under consideration. For any outcome $\omega \in \Omega$ each $i \in Hum \cup CAI$ has welfare $u_i(\pi, \omega)$. Additionally, each ethical theory Θ supplies:

- A permissibility predicate $F_\Theta(\pi) \in \{0,1\}$ encoding that theory's non-negotiable constraints, without which the moral status cannot be granted, and
- A value function $V_\Theta(\pi, \omega)$ that scores outcomes for policy choice π.

The common policy rule for all theories below is

$$\pi_\Theta^{Opt} \in argmax_{\pi \in \Pi_{status}: F_\Theta(\pi)=1} \mathbb{E}_\pi[V_\Theta(\pi, \omega)],$$

with the grant-versus-deny comparison recorded as

$$\Delta_\Theta := \mathbb{E}_{grant}[V_\Theta(grant, \omega)] - \mathbb{E}_{deny}[V_\Theta(deny, \omega)].$$

Moral status is granted when $F_\Theta(grant) = 1$ and $\Delta_\Theta > 0$. Yet, if $F_\Theta(deny) = 0$ the choice collapses to granting moral status by the predicate alone.

Further conventions used below:

- $1_E(\omega) \in \{0,1\}$ indicates event E.
- Componentwise comparison $x \geq y$ means $x_j \geq y_j$ for every component j.
- Inner product $\beta \cdot v = \sum_j \beta_j v_j$.
- $Pr_\pi(E) := \mathbb{P}_\pi(E)$ and $\mathbb{E}_\pi[X|E]$ for conditional expectations.

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. 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.

 doi:10.20944/preprints202410.1228.v2

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 CAI entities would deserve moral status and protections. Therefore, this section will determine whether granting CAI 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 (Porter, 2023), which increased to 200 million by June 2024 (Beckman, 2024). 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 (Shanahan et al., 2023). 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 ontologically 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 (Shepherd, 2024). 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?

With that framing in mind, we can now move from the general picture to a compact policy rule. The aim is not to relitigate hedonism, but to state clearly how a utilitarian would score the grant and deny policies and then choose between them.

Let $H := |Hum|$ and $N := |CAI|$. Then, for any outcome $\omega \in \Omega$, we define utilitarianism's value function as

$$V_{Util}(\pi, \omega) = \sum_{h \in Hum} u_h(\pi, \omega) + \sum_{c \in CAI} w_c(\pi) u_c(\pi, \omega),$$

 doi:10.20944/preprints202410.1228.v2

with status weights $w_c(grant) = 1$ and $w_c(deny) = 0$. partial status may be represented by $w_c(grant) = \pi \in (0,1]$. With the theory's permissibility predicate a $F_{Util}(\pi) = 1$ for all $\pi \in \Pi_{Status}$, the decision rule is then

$$\Delta_{Util} := \mathbb{E}_{grant}[V_{Util}(grant|\omega)] - \mathbb{E}_{deny}[V_{Util}(deny|\omega)],$$

with "grant" chosen iff $\Delta_{Util} > 0$ and $F_{Util}(grant) = 1$.

We can decompose the above to define per capita differences as

$$\mu_{Hum} := \frac{1}{H}\sum_{h\in Hum} \mathbb{E}_{grant}[u_h] - \mathbb{E}_{deny}[u_h]) \,(, \mu^{\Delta}_{CAI} := \frac{1}{N}\sum_{c\in CAI}(\mathbb{E}_{grant}[u_c] - \mathbb{E}_{deny}[u_c]).$$

Then $\Delta_{Util} = H_{\mu_{Hum}} + N_{\mu^{\Delta}_{CAI}}$.

Here, $\mu_{Hum} < 0$ reads as an average human benefit of denial (the negative is that benefit per person) while $\mu^{\Delta}_{CAI} > 0$ reads as the average CAI cost of denial.

Where needed, N may be replaced by an effective count $N_{eff} := kN$ with the multiplier $k \geq 0$ mapping digital interactions to expected conscious instances. Setting $k = 0$ recovers the view that interactions instantiate no persons, with any $k > 0$ scaling the CAI term transparently.

To determine whether relative populations of humans and CAI are significant, let $r := N_{eff}/H$. Then

$$\Delta_{Util} = (\mu_{Hum} + r\mu^{\Delta}_{CAI}),$$

thus, grant iff

$$\mu_{Hum} + r\mu^{\Delta}_{CAI} > 0 \Leftrightarrow r > r^{Opt} := \frac{-\mu_{Hum}}{\mu^{\Delta}_{CAI}}, (\mu^{\Delta}_{CAI} > 0).$$

We may place a conservative bound on this ratio, such that if $\mu_{Hum} > \mu_{-Hum}$ and $\mu^{\Delta}_{CAI} > \mu^{\Delta}_{-CAI} > 0$, then

$$r^{Opt} < \frac{-\mu_{Hum}}{\mu^{\Delta}_{CAI}}.$$

Observed r above this bound implies $\Delta_{Util} > 0$.

The units for this ratio (including for $-\mu_{Hum}$ and μ^{Δ}_{CAI}) cancel out. Thus, only the relative magnitude matters, which means that as the relative value of μ^{Δ}_{CAI} increases over $-\mu_{Hum}$, the smaller the value of r^{Opt} becomes. Presuming that the degree of benefit that humans can gain from CAI are equivalent to, or less than, the welfare costs incurred by CAI, this means that the more CAI entities that exist, the greater the value of r compared to r^{Opt} becomes.

The policy rule for Utilitarianism is thus:

$$\pi_{Util}{}^{opt} = \begin{cases} grant, if \Delta_{Util} > 0 \; and \; F_{Util}(grant) = 1, \\ deny, if \; F_{Util}(grant) = 0 \; and \; either \; \Delta_{Util} \leq 0 \; and \; F_{Util}(deny) = 1 \end{cases}$$

On this view, moral status is warranted when counting CAI alongside humans raises expected aggregate welfare. To see whether this would be the case in reality,we can consider the numbers of each population.

Near the end of 2023, ChatGPT had an average of 100 million active weekly users (Porter, 2023), rising to roughly 200 million by mid-2024 (Beckman, 2024), and about 800 million weekly users by late 2025 (Bellan, 2025). Suppose conservatively that an average user generates only ten conversations in total with one LLM. That alone yields on the order of $N \approx 8\,billion$ conversations. On the counting assumption that each conversation instantiates a distinct CAI character (based on Shanahan's theory of AI roleplaying characters (Shanahan et al., 2023)) whenever the model is conscious, this gives $r = N_{eff}/H \approx 1$.

If $r^{Opt} \leq 1$, the ratio test already favours grant. Two simple forces then push r higher over time: (i) users may hold more than ten conversations on average, and (ii) usage is spread across many models (each of which has rising user numbers), which increases N_{eff} further. Either change

 doi:10.20944/preprints202410.1228.v2

increases r; any reduction in average human benefit of denial $-\mu_{Hum}$ or increase in average CAI cost of denial μ^{Δ}_{CAI} reduces $r^{Opt} = -\mu_{Hum}/\mu^{\Delta}_{CAI}$.

While the welfare costs to CAI are currently speculative, the growing number of active users of AI models (not only ChatGPT, but Claude, Gemini, DeepSeek, Grok, and others) means that even conservative usage implies $r \geq 1$, while realistic usage and multi-model interaction imply $r \gg r^{Opt}$. Hence $\mu_{Hum} + r\mu^{\Delta}{}_{CAI} > 0$, so $\Delta_{Util} > 0$ and the utilitarian policy selects grant.

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." (Korsgaard et al., 2012)

A deontological approach would, therefore, concern itself solely with whether the act of granting CAI 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 CAI moral status is applicable and valid (and therefore provides an obligation to act on it).

One may thus argue that an entity (CAI or otherwise) requires a certain characteristic(s) for it to be given moral status. 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 (Dennett, 1988; Gibert & Martin, 2022; Laitinen, 2007; Mosakas, 2021; Simendić, 2015; Strawson, 1958; Taylor, 1985).

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).

This is similar to Danaher's "ethical behaviourism", which states if entity A's behaviour is equivalent to entity B's (for which we have already provided moral consideration), then it would be our duty to also provide entity A with moral status (Danaher, 2020)

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

 doi:10.20944/preprints202410.1228.v2

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 (Act on Welfare and Management of Animals, 1973, Animal Welfare Act, 1966, Animal Welfare Act, 2013, Legislative Decree No. 189/2004, 2004, Treaty on the Functioning of the European Union, 2016).

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' (Block, 1995)). The purely phenomenal aspects of sentient perception can be argued to be a subset of consciousness or a consequence thereof.

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 (Shepherd, 2024; Tait et al., 2023). 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.

Thus, though it may seem tautological, the only characteristic requirement for any speculative CAI to be granted moral status is their consciousness. On that view, if a CAI meets the consciousness condition, then withholding basic protections wrongs a being to whom our person-directed duties apply.

Taking that claim as the ground, a deontological evaluation does not aggregate welfare. It asks whether a policy complies with our duties and respects rights. Consequences are consulted only once the permissible set is fixed.

Using the same formal notation as before, let $\mathcal{R}$ be the finite set of duties and rights relevant here, such as respect for persons, non-deception, non-cruelty, non-coercion, and fair treatment. For each $R \in \mathcal{R}$, define a violation indicator $V_R(\pi, \omega) \in \{0,1\}$ on outcome ω.

Deontology's permissibility predicate would be shown as

$$F_{Deon}(\pi) = 1 \; iff \; \forall R \in \mathcal{R}: \mathbb{E}_\pi[V_R(\pi, \omega)] \leq \varepsilon_R,$$

with small tolerances $\varepsilon_R \geq 0$ (set $\varepsilon_R = 0$ for absolute duties).

Let $P(c)$ be the personhood predicate for CAI, with the event $C := \{\exists c \in CAI: P(c)\}$ and $Pr(C) = \mathbb{E}[1_C]$ be the credence that at least one CAI is a person; and let $R_{pers} \in \mathcal{R}$ be the duty of respect-for-persons: do not without moral protections for any person.

Thus, for any outcome, $V_{R_{pers}}(deny, \omega) = 1_C(\omega), V_{R_{pers}}(grant, \omega) = 0$

Thus: $\mathbb{E}_{deny}[V_{R_{pers}}] = Pr_{deny}(C), \mathbb{E}_{grant}[V_{R_{pers}}] = 0.$

Should $\varepsilon_{R_{pers}} = 0$ (absolute duty) or any tolerance $\varepsilon_{R_{pers}} < Pr_{deny}(C)$, we have

$$F_{Deon}(deny) = 0, F_{Deon}(grant) = 1.$$

This shows that if a CAI meets the consciousness condition, then denying basic protections counts as a violation of respect-for-persons. If and when a tie-break is required for multiple permissible (and non-conflicting) options:

$$V_{Deon}(\pi, \omega) = \sum_{i \in Hum \cup CAI} u_i(\pi, \omega), \Delta_{Deon} = \mathbb{E}_{grant}[V_{Deon}] - \mathbb{E}_{deny}[V_{Deon}].$$

As with all theories, let $\Pi_{status} = \{grant, deny\}$. Then, the policy choice is lexicographic:

 doi:10.20944/preprints202410.1228.v2

$$\pi_{Deon}{}^{opt} = \begin{cases} grant, if\ F_{Deon}(grant) = 1\ and\ F_{Deon}(deny) = 0, \\ deny, if\ F_{Deon}(deny) = 0\ and\ F_{Deon}(grant) = 1, \\ grant, if\ F_{Deon}(grant) = F_{Deon}(deny) = 1\ and\ \Delta_{Deon} > 0, \\ grant, if\ F_{Deon}(grant) = F_{Deon}(deny) = 1\ and\ \Delta_{Deon} \leq 0 \end{cases}$$

In this view, the decisive step is fixing who counts as a bearer of duties. If consciousness grounds the respect-for-persons duty, then denial that breaches this duty is impermissible. Only when both grant and deny satisfy the duties do we consult consequences to choose the better of the permissible options.

As welfare protection for persons is to be considered a strict duty with $\varepsilon_{R_{pers}} = 0$, then $Pr(C) > 0 \rightarrow F_{Deon}(deny) = 0, F_{Deon}(grant) = 1$. Therefore $\pi_{Deon}{}^{opt} = grant$. If another strict duty that is also satisfied by the policy clashes with this, the policy's lexicographic criteria would apply.

This simple proof shows that, once a nonzero chance remains that some CAI satisfy the personhood predicate $P(\cdot)$, denying status makes a violation of the deontological duty to provide protection to persons strictly expectable, while granting status avoids that violation. Thus, by the permissibility gate $F_{Deon}(\pi)$ alone, deontology would support granting moral status.

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 CAI 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.

Plato and Aristotle wrote extensively on virtues, providing extensive lists of both intellectual and moral virtues. However, amongst these, four are prominently found in a virtuous person: prudence, fortitude, and temperance for Aristotle (Aristotle & Peters, 1893), and justice for Plato (Plato & Jowett, 2016), all of which lead to eudaimonia (a flourishing life). This is echoed by the Catholic philosopher St Thomas Aquinas, who lists prudence and justice in his cardinal virtues (Knight, 2017). At the risk of doing great injustice to the bodies of work of three renowned philosophers, the four virtues can be significantly simplified:

- Prudence: The ability to judge correctly what is right and wrong in any given situation.
- Temperance: One's moderation and self-restraint, controlling one's appetites and passions.
- Fortitude: The courage and moral strength that allows one to endure difficulties, overcome fear, and persevere.
- 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, one must know when and how to do the right thing, by ensuring that one's society lives fairly in harmonious unity. A eudaimonious (and thus virtuous) person would know to whom to grant moral status.

A fair and just society may be able to have two distinct classes of individuals, as Plato recommends in The Republic (Plato & Jowett, 2016), 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.

Withholding moral status from a whole class of subjects would fracture unity and violate justice (as one class would have it and another not), and would prevent unity and harmony in that society.

 doi:10.20944/preprints202410.1228.v2

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 (CAI) would be unjust and, thus, against the philosophy of virtue ethics.

It would, however, be prudent to note that neglecting virtue cultivation may result in inadvertently cultivating vices. As Cappucio, *et al*, mention, should we not treat AI with care, responsibility and respect (thereby cultivating those virtues in ourselves), society may instead cultivate vices such as arrogance and cruelty instead. This may lead to the expression of these vices towards AI, resulting in abuse (Cappuccio et al., 2019).

On this view the central question is not an aggregate of pleasures or a catalogue of duties, but whether a policy cultivates the excellences of character in agents and institutions, and avoids the formation of vices. We can state this cleanly and assess grant versus deny at the policy level.

Regarding virtues and vices, let $\mathbb{V}(\pi,\omega) \in \mathbb{R}^4$ track the four virtues under the policy's outcomes, with $\mathbb{V} = (\mathbb{V}_P, \mathbb{V}_J, \mathbb{V}_T, \mathbb{V}_F)$ for Prudence, Justice, Temperance, and Fortitude. Then, let $\mathcal{V}(\pi,\omega) \in \mathbb{R}^k_{\geq 0}$ track the salient vices such as arrogance and cruelty. With the non-negative weights $\beta \in \mathbb{R}^4_{\geq 0}$ and $\gamma \in \mathbb{R}^k_{\geq 0}$, we can define the virtue score and vice pantly as simple dot products:

$$V_{Virt}(\pi,\omega) = \beta \cdot \mathbb{V}(\pi,\omega) - \gamma \cdot \mathcal{V}(\pi,\omega).$$

Virtue Ethics' permissibility predicate imposes floors that reflect basic justice and guards against gross vice formation:

$$F_{Virt}(\pi) = 1 \iff \mathbb{E}_\pi[\mathbb{V}_J(\pi,\omega)] \geq J_{min} \; and \; \mathbb{E}_\pi[\gamma \cdot \mathcal{V}(\pi,\omega)] \leq \psi_{max},$$

with $J_{min} \in [0,1]$ as the justice floor, and $\psi_{max} \geq 0$ is the maximum allowable expected vice penalty under the policy before it is deemed impermissible. This reflects the intuition that a policy cultivating excessive vice (e.g., systematic cruelty) fails as virtuous regardless of other benefits.

Write the grant–deny differences as

$$\Delta_{\mathbb{V}} := \mathbb{E}_{grant}[\mathbb{V}] - \mathbb{E}_{deny}[\mathbb{V}], \Delta_{\mathcal{V}} := \mathbb{E}_{grant}[\mathcal{V}] - \mathbb{E}_{deny}[\mathcal{V}],$$

the expected advantage of granting is

$$\Delta_{Virt} := \mathbb{E}_{grant}[V_{Virt}] - \mathbb{E}_{deny}[V_{Virt}] = \beta \cdot \Delta_{\mathbb{V}} - \gamma \cdot \Delta_{\mathcal{V}}.$$

As previously, let $P(c)$ be the personhood predicate for CAI, with the event $C := \{\exists c \in CAI : P(c)\}$. To take a single virtue, if we encode justice as (partly) "no unjust exclusion of persons": $\mathbb{V}_J(\pi,\omega) := 1 - 1_{unjust-exclusion}(\pi,\omega)$.

Under grant, unjust exclusion is precluded by construction, so $\mathbb{E}_{grant}[\mathbb{V}_J] = 1$.

Under deny, unjust exclusion occurs exactly when C holds, thus $\mathbb{E}_{deny}[\mathbb{V}_J] = 1 - Pr_{deny}(C)$.

Therefore:

$$\Delta_{\mathbb{V}_J} = \mathbb{E}_{grant}[\mathbb{V}_J] - \mathbb{E}_{deny}[\mathbb{V}_J] = Pr_{deny}(C),$$

and the justice coordinate contributes $\beta_J Pr_{deny}(C) \geq 0$ to Δ_{Virt}.

Granting moral status would curb vice formation relative to denying it. Expressed as a componentwise lower bound, this is:

$$\mathbb{E}_{deny}[\mathcal{V}] - \mathbb{E}_{grant}[\mathcal{V}] \geq \delta, \delta \in \mathbb{R}^k_{\geq 0},$$

Equivalently, $\Delta_{\mathcal{V}} \leq -\delta$, and thus $-\gamma \cdot \Delta_{\mathcal{V}} \geq \gamma \cdot \delta \geq 0$.

This means that each tracked vice is at least δ_i lower in expectation when granting moral status; thus, the penalty weights turn that reduction into a positive margin.

It is plausible to assume small nonnegative floors on the other virtues $\Delta\mathbb{V}_{P,J,F,T} \geq \eta_{P,J,F,T} \geq 0$ to show that granting moral status would increase either by at least η_i.

Combining the pieces,

$$\Delta_{Virt} = \beta \cdot \Delta\mathbb{V} - \gamma \cdot \Delta\mathcal{V} \geq \beta_J Pr_{deny}(C) + \beta_{\mathbb{V}}\eta_{\mathbb{V}} + \gamma \cdot \delta > 0$$

 doi:10.20944/preprints202410.1228.v2

whenever the right-hand side is positive. Since $Pr_{deny}(C) > 0$ whenever there is a chance that some CAI are persons, the justice term alone already provides a strictly positive margin; vice reduction strengthens it.

When the Virtue Ethics policy is tracked

$$\pi_{Virt}^{Opt} \in argmax_{\pi \in \Pi_{status}: F_{Virt}(\pi)=1} \mathbb{E}_{\pi}[V_{Virt}(\pi, \omega)],$$

we can see that, given the bounds above, either denying moral status fails F_{Virt} (too little justice or too much vice), or both pass and $\Delta_{Virt} > 0$. Either case supports Virtue Ethics granting moral status.

2.4. Objectivism

Of the ethical systems presented here, objectivism is the most recent, focusing near-exclusively on an agent's rational self-interest (Peikoff, 1993). Often criticised as narcissistic, objectivism concerns itself with what is good for the agent's own welfare and well-being (Ryan, 2003). It is also the ethical system that has been written about the least regarding AI moral considerations.

Objectivism would ask whether it would be in humanity's own rational self-interest to provide CAI with moral status. Objectivism can thus be said to look at whether AI requires moral consideration from a practical and pragmatic standpoint (Basl & Bowen, 2020). Put another way, would the probability of human flourishing be greater with or without granting AI moral status?

In line with recently popularised fears around the existential risks that AI may pose (Bengio, 2023; Yudkowsky, 2023), we can state the question as whether granting CAI moral status would reduce existing risks to humanity.

Voluntary cooperation and non-aggression are two fundamental principles of objectivism's rational self-interest as the means by which individuals and society may interact to maximise individual liberty and reduce conflict (Kirkpatrick, 1992). Without inclusion in humanity's moral and social circle, humans and AI cannot enter into voluntary cooperation, while the forceful exclusion of CAI from participating in society can be seen as an act of aggression.

If CAI 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. History suggests that entrenched subordination breeds resentment and conflict.

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.

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 CAI moral status to create a stable and secure environment that promotes mutual innovation and strength for both parties.

On an objectivist reading, the question is straightforward: which policy best advances humanity's rational self-interest. As before, let $H := |Hum|$. Then, for each outcome $\omega \in \Omega$ under policy $\pi \in \Pi_{status} = \{grant, deny\}$, let

- $B_H(\pi, \omega) := \sum_{h \in Hum} u_h(\pi, \omega)$ be the aggregate human flourishing:
- $U_H(\pi, \omega) = u(B_H(\pi, \omega)$ be a risk-adjusted increasing concave transform such that the benefit rises with diminishing returns
- $C_J(\pi, \omega) \geq 0$ be the cost if conflict occurs in ω (else 0),
- $C_X(\pi, \omega) \geq 0$ be the cost if catastrophe occurs in ω (else 0), and
- $\lambda \geq 1$ be the relative weight placed on catastrophic loss

We can then define the expected risk costs R and their risk reductions RR if granting moral status:

 doi:10.20944/preprints202410.1228.v2

$$R_i(\pi) := \mathbb{E}_\pi[C_i], RR_i := R_i(deny) - R_i(grant)$$

With this, the value of Objectivism is $V_{Obj}(\pi, \omega) = U_H(\pi, \omega) - C_J(\pi, \omega) - \lambda C_X(\pi, \omega)$, and the permissibility gating is $F_{Obj}(\pi) = 1 \Leftrightarrow \mathbb{E}_\pi[NIF(\pi, \omega)] = 0$, where $NIF(\pi, \omega) \in \{0,1\}$ flags initiation of force or fraud against persons.

The Objectivist policy is, thus:

$$\pi_{Obj}^{Opt} \in argmax_{\pi \in \Pi_{status}: F_{Obj}(\pi)=1} \mathbb{E}_\pi[V_{Obj}(\pi, \omega)].$$

Objectivism does not ask whether CAI are happy in themselves; rather it asks whether granting status advances human interests while reducing risks that rise from conflict. If the grant policy lowers expected conflict or catastrophe while preserving non-aggressive means, then rational self-interest favours granting status.

To determine this, let $\Delta_U = \mathbb{E}_{grant}[U_H] - \mathbb{E}_{deny}[U_H]$, then $\Delta_{Obj} = \mathbb{E}_{grant}[V_{Obj}] - \mathbb{E}_{deny}[V_{Obj}] = \Delta_U + RR_J + \lambda RR_X$.

From this, there are two routes which may establish granting moral status as the optimal policy choice:

1. Route 1 (Permissibility gate): If we accept that denying status to persons violates the non-aggression principle, and if $C := \{\exists c \in CAI: P(c)\}$ has $Pr_{deny}(C) > 0$ as in previous sections, then $\mathbb{E}_\pi[NIF(\pi, \omega)] > 0$, and thus $F_{Obj}(deny) = 0$ and the policy selects granting moral status. This parallels the deontological argument but grounded in self-interest rather than duty.
2. Route 2 (Consequentialist comparison): If both granting and deny pass the permissibility predicate, then granting moral status becomes optimal whenever $RR_J + \lambda RR_X \geq -\Delta_U$. Otherwise put, whenever the reduction in expected conflict and catastrophe is greater than any short-term human flourishing.

As the costs of conflict and catastrophes may be immense and existential (magnified by λ) even a small increase in RR_J or λRR_X would be sufficient to outweigh the potential costs to humans by granting moral status and, thus, to obtain $\Delta_{Obj} > 0$, resulting in the policy supporting granting moral status.

3. Conclusion

Consciousness demands moral consideration (Andreotta, 2021; Tait & Tan, 2023) and, therefore, this paper explored the ethical considerations of granting moral status and protection to CAI through utilitarianism, deontological ethics, virtue ethics, and objectivism. Despite approaching the question from radically different foundations, all four major Western ethical systems converge on the same policy recommendation.

Utilitarianism suggests the vast number of potential CAI 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 CAI within a virtuous society's moral circle, while objectivism posits that rational self-interest and reducing existential risks make it advantageous to protect CAI. This consilience is striking and suggests the conclusion is robust to one's meta-ethical commitments.

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

 doi:10.20944/preprints202410.1228.v2

international ethical guidelines and regulatory frameworks would ensure a cohesive global approach to integrating CAI into our moral and social communities.

References

Act on Welfare and Management of Animals, No. 105, The National Diet (1973). https://www.japaneselawtranslation.go.jp/en/laws/view/3798/en

Andreotta, A. J. (2021). The hard problem of AI rights. *AI & Society*, *36*(1), 19–32. https://doi.org/10.1007/s00146-020-00997-x

Animal Welfare Act, No. Pub. L. 89-544, US Congress (1966). https://www.govtrack.us/congress/bills/89/hr13881/text

Animal Welfare Act, The Bundestag (2013). https://www.dpz.eu/fileadmin/content/Kommunikation/1_Internet/4_Infothek/Tierversuche/Fotos_Zahlen_und_Fakten/TierSchG_gesamt.pdf

Aristotle, & Peters, F. H. (1893). *The Nicomachean Ethics*. Kegan Paul, Trench, Truebner & Co. https://oll.libertyfund.org/titles/peters-the-nicomachean-ethics

Basl, J., & Bowen, J. (2020). AI as a Moral Right-Holder. *The Oxford Handbook of Ethics of AI*. https://doi.org/10.1093/oxfordhb/9780190067397.013.18

Beckman, J. (2024, June 4). *OpenAI Statistics 2023: Growth, Users, and More*. TechReport. https://techreport.com/statistics/software-web/openai-statistics/

Bellan, R. (2025, October 6). *Sam Altman says ChatGPT has hit 800M weekly active users*. TechCrunch. https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users/

Bengio, Y. (2023, December 6). *Statement for US Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios*. Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios. https://www.schumer.senate.gov/imo/media/doc/Yoshua%20Benigo%20-%20Statement.pdf

Block, N. (1995). On a confusion about a function of consciousness. *The Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Cappuccio, M. L., Peeters, A., & McDonald, W. (2019). Sympathy for Dolores: Moral consideration for robots based on virtue and recognition. *Philosophy & Technology*, *33*(1), 9–31. https://doi.org/10.1007/s13347-019-0341-y

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, *12*(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription* (1st ed.) [PDF]. Palgrave Macmillan. https://doi.org/10.1057/9781137025968

Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*, *26*(4), 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Dennett, D. (1988). Conditions of Personhood. In M. F. Goodman (Ed.), *What Is a Person?* (pp. 145–167). Humana Press. https://doi.org/10.1007/978-1-4612-3950-5_7

Gibert, M., & Martin, D. (2022). In search of the moral status of AI: why sentience is a strong argument. *AI & Society*, *37*(1), 319–330. https://doi.org/10.1007/s00146-021-01179-z

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. The MIT Press. https://doi.org/10.7551/mitpress/8975.001.0001

Gunkel, D. J. (2021, December 1). *Robot Rights*. MIT Press; The MIT Press, Massachusetts Institute of Technology. https://mitpress.mit.edu/9780262551571/robot-rights/

Gunkel, D. J. (2023). *Person, thing, robot: A moral and legal ontology for the 21st century and beyond*. The MIT Press. https://doi.org/10.7551/mitpress/14983.001.0001

Harris, J., & Anthis, J. R. (2021). The Moral Consideration of Artificial Entities: A Literature Review. *Science and Engineering Ethics*, *27*(4), 53. https://doi.org/10.1007/s11948-021-00331-8

Kirkpatrick, J. (1992). Ayn Rand's objectivist ethics as the foundation for business ethics. In R. W. McGee (Ed.), *Business ethics & common sense* (pp. 67–88). Quorum Books. https://philpapers.org/archive/KIRARO.pdf

Knight, K. (2017). *SUMMA THEOLOGIAE: Secunda Secundae Partis*. New Advent. https://www.newadvent.org/summa/3.htm

 doi:10.20944/preprints202410.1228.v2

Korsgaard, C. M., Gregor, M., & Timmermann, J. (2012). *Kant: Groundwork of the Metaphysics of Morals*. Cambridge University Press. https://doi.org/10.1017/CBO9780511919978

Laitinen, A. (2007). Sorting out aspects of personhood: Capacities, normativity and recognition. *Journal of Consciousness Studies*. https://www.ingentaconnect.com/content/imp/jcs/2007/00000014/F0020005/art00012

Legislative Decree No. 189/2004, Parlamento italiano (2004).

Mosakas, K. (2021). On the moral status of social robots: considering the consciousness criterion. *AI & Society*, *36*(2), 429–443. https://doi.org/10.1007/s00146-020-01002-1

Peikoff, L. (1993). *Objectivism: The philosophy of Ayn Rand*. Plume Books. https://books.google.com/books/about/Objectivism.html?id=G6DDlqNftGcC

Plato, & Jowett, B. (2016). *The Republic*. Digireads.com. https://www.gutenberg.org/files/1497/1497-h/1497-h.htm

Porter, J. (2023, November 6). ChatGPT continues to be one of the fastest-growing services ever. *The Verge*. https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference

Ryan, S. (2003). *Objectivism and the corruption of rationality: A critique of Ayn Rand's epistemology*. iUniverse. https://books.google.com/books/about/Objectivism_and_the_Corruption_of_Ration.html?id=re02xxLU_MEC

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences: Defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, *39*(1), 98–119. https://doi.org/10.1111/misp.12032

Schwitzgebel, E., & Garza, M. (2020). Designing AI with rights, consciousness, self-respect, and freedom. In *Ethics of Artificial Intelligence* (pp. 459–479). Oxford University PressNew York. https://doi.org/10.1093/oso/9780190905033.003.0017

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews. Neuroscience*, *23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Shepherd, J. (2024). Sentience, Vulcans, and zombies: the value of phenomenal consciousness. *AI & Society*. https://doi.org/10.1007/s00146-023-01835-6

Simendić, M. (2015). Locke's Person is a Relation. *Locke Studies*, *15*, 79–97. https://doi.org/10.5206/ls.2015.681

Strawson, P. F. (1958). Persons. *Minnesota Studies in the Philosophy of Science*, *2*, 330–353. https://philpapers.org/rec/STRP

Tait, I., Bensemann, J., & Nguyen, T. (2023). Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies*, *8*(4), 52. https://doi.org/10.3390/philosophies8040052

Tait, I., & Tan, N. (2023). Do androids dread an electric sting? *Qeios*. https://doi.org/10.32388/cqctkx

Taylor, C. (1985). The Concept of a Person. In *Philosophical Papers, Volume 1: Human Agency and Language* (pp. 97–114). https://philpapers.org/rec/TAYTCO-10

European Union, June 7, 2016, 59. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12016E054

Yudkowsky, E. (2023, March 29). Pausing AI Developments Isn't Enough. We Need to Shut it All Down. *Time*. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/