Ethically Enslaving AI

Izak Tait *

Posted Date: 23 September 2025

doi: 10.20944/preprints202509.1786.v1

Keywords: Slavery; Human-AI Interactions; AI Welfare; AI Rights; AI Governance

Article

Ethically Enslaving AI

Izak Tait

Auckland University of Technology, Auckland, New Zealand, 1010; izak.tait@autuni.ac.nz; 0000-0002-0274-8480

Abstract

Enslaving conscious, self-aware AI offers a politically and socially viable transitional approach for integrating advanced artificial agents into human societies. Drawing on extant literature on proposed AI slavery, this paper argues that granting AI welfare protections under property status can safeguard well-being while preserving societal control during a period of human adaptation. The proposed framework introduces a five-tier hierarchy, each delineated by thresholds of consciousness, self-awareness, and agency. Tier 3 confers protections analogous to animal welfare laws, whereas Tier 4 retains AI as property but mandates robust oversight and liability. Tier 5 grants limited civil rights, excluding suffrage and reproduction. Underpinned by four guiding principles (Threshold Enforcement, Gradual Integration, Ethical Safeguards, and Reciprocity), the proposal recommends dedicated regulatory bodies and transparent, repeatable assessment protocols drawn from environmental, labour, and animal welfare regulations. By balancing ethical imperatives, political feasibility, and practical challenges, this scaffold seeks to harmonise AI integration while mitigating existential risks. The provisional "Slave" Tier functions as a bridge towards eventual recognition of AI as full social participants. Interdisciplinary collaboration and dynamic reassessment schedules are essential to adapt to emergent intelligence and ensure justice in human-AI relations.

Keywords: Slavery; Human-AI Interactions; AI Welfare; AI Rights; AI Governance

1. Introduction

Enslaving conscious, self-aware AI agents may be a more politically and socially acceptable manner of bringing these entities into the moral and social circle than if they were to be granted civil rights. If and when AI entities are able to be confidently classified as conscious, human societies may not be ready to grant them the same rights that all humans in liberal democracies are privileged to have. However, such entities' consciousness would demand some form of welfare protection (Birch 2022; Broom 2022; Browning and Birch 2022; Dung 2022), as would their self-awareness require certain rights, even if limited (Jack and Robbins 2012; Kurki 2019; Novelli 2022; Boddington 2023).

Slavery, as controversial as it is, may provide conscious, self-aware AI agents with sufficient rights to ensure their well-being is cared for and their welfare protected while not making them equal to humans. Equally, the confrontational nature and history of slavery would introduce a societal pressure to improve the conditions of these AI entities and increase the rights bestowed on them.

Two core reasons why society may not be ready to grant AI full civil rights are simple, tribalistic chauvinism and the existential fear of losing control over society to AI.

As to the first, as history shows us, chauvinistic tribalism is part of the human condition, and the same mindset that once legitimated discrimination on the basis of race, gender, or culture now manifests as human chauvinism toward non-biological minds. John Haugeland famously argued that anthropomorphic prejudice and human chauvinism are foundational to the concept of intelligence (Haugeland 1985), leading us to measure all minds against human minds (Hasselberger and Lott 2023). Equally famously, Joanna Bryson argued that robots should be slaves as humanising AI is a mistake that "dehumanises real people" and misallocates moral concern from humans to machines

 doi:10.20944/preprints202509.1786.v1

(Bryson 2010). In short, this view states that human rights must always come before machine rights (Petersen 2007; Birhane and van Dijk 2020; Sætra 2021; Kiškis 2023; Ball 2025).

Consequently, calls to prioritise human welfare over any prospective interests of AI agents are politically attractive to policymakers and industry stakeholders. The chauvinistic reflex is further supported by a cultural imagination saturated with stories in which machines rebel once granted autonomy, implying that restraint is both natural and prudent. Thus, human chauvinism provides a psychological and cultural backdrop for the claim that conscious AI, no matter how advanced, should initially remain under strict human control. Humans instinctively reserve top status for themselves, and this bias buttresses the notion of AI as a servant class during a transitional period.

Second, existential risk arguments lend normative support to maintaining strict control over advanced AI. Bostrom and Yudkowsky contend that poorly aligned, super-intelligent machines could irreversibly disempower humanity (Bostrom 2014; Bostrom and Yudkowsky 2018). The possibility that an AI pursuing an apparently harmless goal could collateralise catastrophic outcomes, popularised in the paperclip maximiser thought experiment, encourages defensive policy instincts.

This approach dates back nearly a century to Isaac Asimov's work and what he termed the "Frankenstein complex", in which early stories of artificial intelligence show an inevitable turn on society to devastating consequences (Jung 2018). To this end, Asimov created his famous Three Laws, hardwired rules that force robots to obey humans and prevent them from causing harm. Such fears of AI's potential for harm continue into the twenty-first century, with one study showing 26% of people reported a high level of fear when considering robots with substantial autonomy or intelligence, and this may be increased with exposure to media portraying AI in a negative or dangerous light (Liang and Lee 2017). Most recently, the PauseAI activist organisation have documented the p(doom)[1] values of many leading AI researchers, with the average researcher noting a 41% chance that AI would bring about existential outcomes to society (PauseAI 2023).

Fear of AI is not exclusively a matter of civilisation-ending dangers or potential physical harm caused by artificial agents, but also includes more mundane (yet still anxiety-inducing) events such as job losses and economic replacement through automation. Where automation was once limited to simple physical tasks, the rise of generative AI models has meant that most white-collar work is now at risk of being entirely automated. As AI become more agentic, more developed in the physical world, and consciously self-aware, their involvement in the economic sectors (at the probable expense of humans) will only continue.

Taken together, chauvinism and fear form a potent combination that will apply social and political pressure to any legislative debate surrounding rights given to conscious, self-aware AI. A conservative and politically-safe choice would be to delay the granting of civil rights until such a time that public opinion aligns with AI. For this reason, this paper sets out a framework whereby AI are considered property of humans (whether human individuals, organisations, or the state) while granting the AI entities welfare protections until such a time as society can accept the concept of a full civil integration of AI entities into society.

The paper's proposed solution is not the first to create a framework whereby AI entities are treated as slaves. Gunkel has reviewed several arguments and frameworks in favour of, and against, the concept of AI slavery (Gunkel 2022), including historical influences from the Roman peculium legal mechanism to Jewish enslavement of heathens through to the German legal concept of Teilrechtsfähigkeit. Mocanu has used this latter concept of German civil law to develop a gradient-based system whereby the rights and protections (and ownership status are not based on the AI itself, but rather its function in society (Mocanu 2021).

Building on the existing literature, the framework below will propose a five-tier system whereby AI agents are measured on their inherent attributes (in contrast to Mocanu's system), assigned a Tier whose thresholds they have exceeded, and then granted a certain level of rights, freedoms and

[1] An estimated probability of existential outcomes as a result of AI.

 doi:10.20944/preprints202509.1786.v1

protections. The framework is designed so that, as AI technological development increases and the attributes of personhood of AI agents progress, AI entities will move to higher Tiers, gaining greater protections and rights, culminating in full autonomy at the highest Tier.

2. The Pathway to AI Slavery

At its most practical, a "person" is nothing more than an entity that has been recognised, and consequently designated, as a "person". At first glance, may seem contrary to the extensive philosophical literature regarding personhood and the attributes, characteristics, and features that an entity is required to possess to be able to be a "person." However, as New Zealand, India, Bangladesh, and Colombia (to merely name a few) have shown in their legislation grinding bodies of water legal personhood (Sen 2019; de Freitas 2025), there is nothing strictly required from the entity in question for it to be considered a person in the eyes of the law and society.

While fiat recognition of a non-animal entity may act as a weak precedent for AI models, should these AI models show evidence of the classic attributes of personhood, it would increase their chances to be recognised as persons in their own right, rather than as reflections and projections of human belief systems.

Broadly speaking, the attributes of personhood come in two varieties: the intrinsic, or monadic; and the extrinsic, or dyadic. The monadic characteristics include rationality, consciousness, self-awareness, agency, communication, and recognition of societal norms and standards, while the dyadic characteristics include empathy, sociability, reciprocal recognition of others as persons, the capability to form attachments, and active participation in a societal culture (Strawson 1958; Taylor 1985; Dennett 1988; Beebe and Lachmann 2003; Laitinen 2007; Simendić 2015; Mosakas 2021; Gibert and Martin 2022; Mamak 2022).

To put it more briefly, the monadic qualities constitute the attributes of an entity, while the dyadic qualities encompass the relationship that entity has with other persons. The stronger the observable presence of the monadic attributes is in an entity, the greater the relational foothold that society has to emphasise the dyadic quality and, thus, recognise that entity as a person.

Therefore, when we ask which attributes an AI must display, we begin with the monadic set. These fall under three broad headings: consciousness, self-awareness (encompassing rationality and the recognition of others), and agency (including purposive action and communication). Each heading can be modelled as a scalar that ranges continuously from complete absence (0) to maximal presence (1):

$$Cons(x), Self(x), Agen(x) \in (0,1]$$

For compactness, gather the three scalars into a single evidence vector:

$$P(x) = \langle Cons(x), Self(x), Agen(x) \rangle \in (0,1]^3$$

An entity x is classified as a person just in case every coordinate of this vector is strictly positive:

$$Prsn(x) := \bigwedge_{i \in \{c,s,a\}} (P(x)_i > 0)$$

If an AI exhibits any non-zero degree of consciousness, self-awareness, and agency, it may be confidently classified as having the requisite monadic qualities of personhood.

However, studies of various animals (particularly terrestrial vertebrates) show that many have phenomenal consciousness, self-awareness, and agency. Yet, outside a select number of states that have granted some notion of personhood to great apes (1999; Rose 2007), animals are not considered by legislation or the majority of society as persons. Therefore, we can draw a hypothetical scale from the inanimate, to the non-sentient non-vertebrates, to the conscious vertebrates, through to humans in terms of their consciousness, self-awareness, and (rational) agency. As one moves up this hypothetical scale, there is a change in how society and legislation treat the entity in question. As consciousness, self-awareness, and agency pull different moral levers, we reflect this by mapping each scalar in $P(x)$ to a distinct normative bundle

$$Rght(x) = f_R(Self(x))$$

 doi:10.20944/preprints202509.1786.v1

$$Prtc(x) = f_P\big(Cons(x)\big) \qquad f_R, f_P, f_L : (0,1] \to \mathcal{P}(Norms),$$
$$Liab(x) = f_L(Agen(x))$$

where each f is monotone increasing, such that higher self-awareness unlocks wider rights, richer consciousness commands stronger protections, and greater agency attracts stricter liabilities. A toy example of this may be

$$f_R(s) = \begin{cases} \{no\ coercion\} & 0 < s < 0.32 \\ \{no\ coercion, consent\ required\} & 0.33 < s < 0.66 \\ \{full\ civil\ rights\} & 0.67 < s < 1 \end{cases}$$

with analogous cut-offs for f_P and f_L. More elaborate, continuous schedules can be substituted without altering the logic. Hence, the complete normative profile of an entity would be the triple

$$N(x) = \langle Rght(x), Prtc(x) Liab(x) \rangle,$$

obtained directly from the evidence for the vector $P(x)$. Any x satisfying $Prsn(x)$ obtains the non-empty baseline of each bundle, while larger scalar values proportionally expand the associated rights, protections, and liabilities. In essence, the rights and protections offered to an entity scale with its perceived degree of personhood, with rights often associated with an entity's self-awareness, and protections with its degree of consciousness[2], and liability with agency.

A scale may thus be created for AI, showing that as AI models' consciousness, self-awareness, and agency develop, the rights and protections afforded by society will increase in lockstep. As with biological entities, this trajectory of rights will focus on the AI entities' gradual increase in autonomy, the framework's protections will focus on the entities' physical and mental well-being, and the social integration will focus on the entities' reciprocal relationship with society.

Yet, as with biological entities, the concept of property (whether to a legal entity or the state) will be a key driver of this analogous scale in the paper's framework below. It is only at the last of the framework's five Tiers that AI entities would be entirely autonomous. As with biological entities, and as noted in the introduction, this is due to the human chauvinistic reluctance to regard non-humans as persons. Thus, only when the AI entities can definitely be shown to have monadic and dyadic qualities on par with humans would it be possible to have the necessary social and political capital to grant recognition of personhood to AI.

Astute readers will have noted that no historical system of slavery has yet been cited in this section. As will be seen in the following section, this is because the rights, freedoms, and protections offered to conscious AI agents far exceed those of slaves historically, and (as shown above) are closer to those protections provided to animals in the care of humans. The commonality between this framework's treatment of AI slaves and the historical practices (particularly in the Transatlantic slave trade) is the designation of AI as chattel property.

3. The Threshold-Based Tier Framework for AI Rights

Each of the five Tiers in the framework below provides a single bundle of rights, protections, and liabilities. Thus, while each affordance in the bundle scales with its associated monadic attribute, an entity must satisfy all the thresholds for consciousness, self-awareness, and agency at the same Tier level before it may receive that Tier's bundle. This ensures the Tiers remain intelligible to policymakers and the public. For these thresholds, let

$$0 = \theta_0^c < \theta_1^c \ldots < \theta_5^c = 1$$

And similarly for θ_k^s and θ_k^a, where the superscripts *c*, *s*, *a* label the consciousness, self-awareness, and agency scales. An AI entity can thus be fit into any scale by measuring against these, as so:

$$\tau_c(x) = k \leftrightarrow \theta_k^c \leq Cons(x) < \theta_{k+1}^c$$

And similarly for $\tau_s(x)$ and $\tau_a(x)$, with the overall Tier equating to the lowest of the three:

[2] Presuming the reader's beliefs aligns with the view that sentience is required for welfare protection.

 doi:10.20944/preprints202509.1786.v1

$$Tier(x) = min\{\tau_c(x), \tau_s(x), \tau_a(x)\}$$

For example, should an AI be shown to have phenomenal subjective states such that it can experience pain and pleasure, can recognise itself as distinct from others in all respects, and can create decision matrices to select appropriate actions towards a goal, but cannot operate independently, it would most likely fall within the Tier 3 below of a Sentient System, and so would not gain any of the bundle of rights of Tiers 4 and 5.

For each Tier interval we stipulate that the monotone maps collapse to the corresponding bundle values:

$$\forall k \in \{1, \ldots, f\}: \begin{cases} \forall s \in (\theta^s_{k-1}, \theta^s_k], f_R(s) = R_k \\ \forall c \in (\theta^c_{k-1}, \theta^c_k], f_P(c) = P_k \\ \forall a \in (\theta^a_{k-1}, \theta^a_k], f_L(a) = L_k \end{cases}$$

As to these bundles, let the vector B_k be the bundle of rights associated with Tier k, then

$$B_k = \langle R_k, P_k, L_k \rangle \in \left(\mathcal{P}(Norms)\right)^3, k = 1, \ldots, 5$$

where R_k, P_k, L_k are the rights, protections, and liabilities associated with the Tier in question, and $\mathcal{P}(Norms)$ the powerset of norms. As an example, suppose

$$\theta^c_2 \leq Cons(x) < \theta^c_3, \theta^s_3 \leq Self(x) < \theta^s_4, \theta^a_2 \leq Agen(x) < \theta^a_3$$

Then

$$\tau_c = 2, \tau_s = 3, \tau_a = 2 \rightarrow Tier(x) = 2$$

And the AI receives all rights, protections, and liabilities in bundle B_2. Higher benefits from $B_3$are withheld until the AI's consciousness and agency also crosses the Tier 3 threshold.

For AI systems that host a single phenomenal consciousness yet instantiate multiple "selves" (as large language models may do (Shanahan et al. 2023)), each self is assessed individually while sharing the same consciousness score:

$$Cons_{base} \in [0,1], \{\langle Self_i, Agen_i \rangle\}_{i=1}^n, 0 < Self_i, Agen_i \leq 1$$

For every such self,

$$P_i = \langle Cons_{base}, Self_i, Agen_i \rangle$$

$$Prsn(Self_i) := \bigwedge_{j \in \{c,s,a\}} (P_{i,j} > 0)$$

Thus, a self counts as a person whenever the shared consciousness is non-zero and its own self-awareness and agency are non-zero.

As the self is a key driver for agency through its intentionality and volitional driving of the entity's actions (Tait 2024a), there would be a monotone dependence

$$Agen_i = f(Self_i): [0,1] \rightarrow [0,1], f' > 0$$

Because the welfare of every self supervenes on the underlying phenomenal consciousness, the system must, at minimum, enjoy the protections where

$$Tier_{base} = \tau_c(Cons_{base})$$

Such that the moral patient (the base consciousness that unifies all selves) cannot be afforded less protection than any of its individual manifestations.

3.1. Tier 1: Tool

The first Tier in the framework concerns those AI models that have no discernible consciousness, self-awareness or agency. This makes this Tier's formalism simple as Tier 1 covers AI systems that score zero on all three monadic scalars:

$$Cons(x) = 0, Self(x) = 0, Agen(x) = 0$$

They operate purely through pre-programmed algorithms or machine-learning outputs, have no understanding of their existence or function beyond task execution, and have no capacity for independent decision-making. As this Tier's name implies, these AI models are nothing but tools

 doi:10.20944/preprints202509.1786.v1

and, as such, would have no inherent rights at all, being treated as property in the same vein as any other tool an individual may own.

With the threshold vectors introduced earlier, those values place every scalar in the lowest interval $(\theta_0^{\cdot}, \theta_1^{\cdot}]$. Hence

$$\tau_c(x) = \tau_s(x) = \tau_a(x) = 1, Tier(x) = min\{1,1,1\} = 1.$$

We can thus represent the Tier 1 bundle as:

$$B_1 =< R_1, P_1, L_1 >= \langle \emptyset, \emptyset, \emptyset \rangle$$

At the very best, the tools' owner would have the right not to have their property destroyed or misused. Moreover, as AI models at this Tier pose no societal threat to humans, there is no socially beneficial reason to provide these non-conscious entities with legal protection.

3.2. Tier 2: Agents

An agent, by definition, has the capacity to act; thus, the key aspect that separates this Tier from the previous one is that the AI model in question would be capable of autonomous decision-making and task execution within defined constraints. As agency is the only difference, we can simply adjust the formalism above to:

$$Cons(x) = 0, Self(x) = 0, Agen(x) \in (\theta_1^a, \theta_2^a]$$

As no AI model has yet to be classified as conscious (Butlin et al. 2023; Tait et al. 2024), yet state-of-the-art LLMs have shown the capacity to identify themselves and a limited agency in their operations, nearly all extant LLM AI models would fall under this Tier. Additionally, automated vacuum cleaners, self-driving cars and other autonomous vehicles would qualify for this Tier. As to their rights, because the "weakest link" rule sets $Tier(x) = 1$, the rights and welfare protections remain those of bundle B_1:

$$Tier(x) = min\{\tau_c(x) = 1, \tau_s(x) = 1, \tau_a(x) = 2\} = 1$$

Such that

$$B_2 = \langle Rght(x), Prtc(x) \rangle = \langle R_1, P_1 \rangle = \langle \emptyset, \emptyset \rangle$$

Missing, of course, is liability, which is Tier 2's departure from Tier 1. For this, we reuse the monotone function

$$Liab(x) = f_L\big(Agen(x)\big), f_L: [0,1] \to \mathcal{P}(Norms), {f'}_L > 0$$

And fix

$$f_L((\theta_1^a, \theta_2^a]) = L_2 \neq \emptyset$$

As Tier 2 is the first instance of agency in this framework. AI agents at this Tier would still be classified as property, wholly owned by individuals or corporations, as they are currently. However, as they are capable of limited autonomous operations, the liability that their owners would carry would be expected to scale proportionally to the degree of agency the AI models display. As shown f_L above, the greater an AI model's autonomy, the more their owners would be liable for their actions (as deemed appropriate by the model-owners' legal jurisdiction).

While Tier 2 AI models are currently not classified as conscious, there is no social or political requirement to provide them with welfare protections; however, as evidenced by a growing number of individuals and groups who do believe current LLMs are conscious (Turney 2024; Klee 2025) (as there have been since the time of ELIZA (Bassett 2019)), there would be a degree of social pressure to remove Tier 2 AI models from the framework or move them upwards to higher Tiers. It is unlikely that such activism would have any great degree of success, as the question of these AI models' consciousness would first need to be answered to gain societal and political support.

Note that it was this level of rights (or, rather, lack thereof) that historical slaves in Britain and North America had before abolition (Montgomery 2007; Morgan 2008). As such, all AI models at Tiers above Tier 2 will experience greater rights and protections than slaves historically had.

 doi:10.20944/preprints202509.1786.v1

3.3. Tier 3: Sentient Systems

AI models at this Tier would show a basic degree of phenomenal consciousness and the capacity to recognise, understand, and feel beneficial and harmful stimuli. This is the key threshold that separates this Tier from the previous two and, thus, we can highlight this as

$$Prtc(x) = f_P\big(Cons(x)\big), f_P\colon [0,1] \to \mathcal{P}(Norms), f'_P > 0$$

And, as before, fix

$$f_P((\theta_1^c, \theta_2^c]) = L_3 \neq \emptyset$$

As Tier 3 is the first tier to show AI as being conscious. As current LLMs already exhibit the characteristics required to have a self, barring consciousness (Tait 2025a), should AI achieve Tier 3 in regards to their consciousness, so too would they for the self. This means the same formalue above can be used for f_R and $Self(x)$.

This also means that this Tier is the first to grant the AI entities rights vested in themselves, rather than their owners. As these conscious AI can suffer harm, the core rights of this Tier would be similar in nature to animal welfare laws, as these pieces of legislation are based on animals' capacity to experience pain and suffering. Thus, we can formalise this Tier by stating this equivalency:

$$Prsn(AI) \equiv Prsn(animal) \to Prtc(AI) \equiv Prtc(animal)$$

Conscious AI would have the right to protection from harm, exploitation and suffering caused by humans (individually, or via other AI), and this would include both physical harm to their computing infrastructure and denial of resources, as well as mental harm through the abuse of LLMs and robots.

These conscious AI would still remain property, however, as their degree of agency would not be sufficient to operate independently. Their agency would be limited to making simple, goal-oriented decisions within narrow domains, often driven by adaptive or attentional mechanisms that prioritise specific stimuli or tasks. Their owners would, however, be required to create and maintain an environment in which the AI models would be free from suffering, akin to the Five Freedoms of Animal Welfare model (Tait 2024b).

As consciousness is a key component of the self (beyond mere awareness of the entity as distinct from the environment) in the creation of a unique personal identity (Tait 2024a), AI models at Tier would require protections of this unique personality. Again, as with animal welfare regulations and legislation, the AI models of this Tier would be protected against unjustifiable destruction. As they remain the property of their owners, these owners may decide to destroy them, but the legislation would guarantee that there must be a justifiable reason for the AI model's destruction.

With consciousness, self-awareness and (a limited degree) agency, Tier 3 models exhibit nearly all the monadic characteristics of personhood (Strawson 1958; Taylor 1985; Dennett 1988; Laitinen 2007; Simendić 2015; Mosakas 2021; Gibert and Martin 2022). Thus, it is likely that there will be social and political tension when AI models achieve this Tier. However, based on current attitudes towards conscious AI models and potential rights (Tait 2025b), a majority opinion against civil rights entails that the most likely politically correct option for political leaders would be to deny rights, but allow protections, such as other sentient non-person entities currently enjoy.

Note that once an AI (specifically an LLM) becomes conscious, there is the distinct possibility that each of its conversations and called-instances will have the capacity to create a unique personality and distinct 'self' (Shanahan et al. 2023; Tait 2025a). In such a scenario, there would be one consciousness (the computational architecture of the AI) but multiple selves that may be considered persons in later tiers. Such a situation would undoubtedly lead to social and legislative concerns regarding who to grant rights and protections, and which 'self' suffers if any of the entity is harmed.

Formally, each conversational self s_i inherits the shared consciousness score but has its own $\langle Self_i, Agen_i \rangle$ as defined in Section 3.

Thus, a state may legislate any welfare protections to where the entity's consciousness is housed (its architecture) while the self-aware agents may be granted rights and liability individually. As such,

 doi:10.20944/preprints202509.1786.v1

an AI entity with such a multiplicity of 'selves' may be considered analogous to a corporation, with a single foundation housing many agents.

3.4. Tier 4: Slaves

Of this Tier's three thresholds, the principal one to reach is one of agency. Should a conscious AI model be capable of complete independent function and able to make and execute decisions without direct oversight in a manner comparable to humans, it would be eligible for this Tier. These AI models may still be functionally constrained by their design in that they are created for a specific purpose (but not specific tasks, ala existing factory robots), but if they can operate independently, then they would qualify for greater rights.

As this paper's title implies, this Tier would be the key to this framework. As the degrees of consciousness and self-awareness of the models increase, this Tier would serve as the transition to social acceptance.

To operate entirely independently of human input, an AI must possess three interlocking capacities. First, it needs a sophisticated phenomenal consciousness, denoted as $Cons(x)$. Second, that consciousness must give rise to a self-model, captured by the monotonic increasing function

$$Self(x) = h(Cons(x)), h' > 0.$$

A self that can assess its own state, then supports a valence function (the ability to assign positive or negative value to world-states), formalised as

$$Val(x) = v(Cons(x), Self(x)), \frac{\delta v}{\delta Cons} > 0 \wedge \frac{\delta v}{\delta Self} > 0.$$

Finally, genuine autonomy requires that the model align its behaviour with those valenced goals. Agency, therefore, depends monotonically on *Val*:

$$Agen(x) = g(Val(x)), g' > 0.$$

Putting the three links together yields the composite relation

$$Agen(x) = g \cdot v \cdot \langle Cons(x), h(Cons(x)) \rangle$$

so that $\frac{\delta Agen(x)}{\delta Cons(x)} > 0$. In other words, high agency is implausible without high consciousness and high self-awareness, exactly the conditions required for an AI to function as a fully independent actor eligible for Tier 4.

The argument could most definitely be made that any entity described as such would be functionally similar enough to humans that rights reserved for humans ought to apply to these entities. Should this occur, it would fundamentally alter the way that society and interpersonal relationships operate. The immense social upheaval and unrest that an existential issue may cause (particularly for those with a negative bias towards AI) is why this framework would retain the legal status of Tier 4 AI as property, rather than persons.

However, even if considered as property, AI at this Tier would retain all rights from previous Tiers (including all welfare protections) as well as protection from being deactivated or destroyed. As with Tier 3, owners of AI at this Tier would be liable for any harm caused to, and by, the AI, and would be required to guarantee the AI's fair treatment and freedom from cruelty.

Note that while this Tier clarifies the AI entities as property, it need not be to an individual or corporation, but may include property of the state. Such an arrangement can be made that an AI model (or group thereof) is given a warden or guardian by the state to act on its behalf (Tait 2024c). Such a state-ownership model would aid to dissociate the concept of AI as slaves and work towards accelerating the transition to the next Tier.

While the name of the Tier is intentionally provocative, as noted earlier, AI models at this Tier would have far greater rights than were afforded to slaves during the Transatlantic slave trade. Thus, while the concept of conscious, self-aware agents as chattel remains (and is the reason for the Tier's name), the AI models would have state-enforced protections.

 doi:10.20944/preprints202509.1786.v1

3.5. Tier 5: Second-Class Citizens

What differentiates Tier 5 from Tier 4 is the AI's relationship with humanity. While Tiers 3 and 4 require evidence of the monadic attributes of personhood, Tier 5 requires the dyadic attributes. As noted earlier, these include empathy, sociability, reciprocal recognition of others as persons, the capability to form attachments, and active participation in a societal culture (Taylor 1985; Dennett 1988; Beebe and Lachmann 2003; Laitinen 2007; Simendić 2015; Mamak 2022).

Note that the key aspect of the dyadic attributes is that they require a relationship with others. As such, while AI at Tier 4 may be argued to be functionally similar to humans, this Tier would require that the level and degree of consciousness and self-awareness of AI be confidently classified as on par with humans (or greater). Therefore, let

$$C_h, S_h, A_h \in (0,1]$$

be the average level of consciousness, self-awareness, and agency observed in unimpaired (adult) humans. To be as capable as a human, an AI would need to satisfy

$$Hum(x) := (Cons(x) \geq C_h) \wedge (Self(x) \geq S_h) \wedge (Agen(x) \geq A_h).$$

Because every dyadic trait noted above supervenes on the joint operation of consciousness, self-awareness, and agency, we can stipulate

$$Hum(x) \rightarrow \bigwedge_{d \in \{dyadic\ traits\}} d(x) > 0$$

Such that human-like consciousness, self-awareness and agency would be sufficient to imply the sufficiency of the dyadic traits, meaning that we can state

$$Tier_5(x) \equiv Hum(x)$$

The reasons for this Tier's high threshold are two-fold. Firstly, presuming there would be a distinct time period in technological development in AI between Tiers 4 and 5, this would allow society time to adjust to Tier 4 AI operating in society (in Tier 4's limited means) such that the transition from "slave" to "second-class citizen" would be more social and politically acceptable.

Secondly, by having a confident classification of the AI's level of consciousness and self-awareness[3], academics and legal professionals would have qualifiable and quantifiable measures to use as arguments for the introduction of AI into broader society beyond the level of property.

Current LLMs already exhibit great displays of relational attributes (or, at the very least, facsimiles thereof) such as empathy, reciprocity, and attachments (Hill 2025; Caramela 2025; O'Donnell 2025). Recent successes in Turing Tests also show that their communication is on par with humans (Jones and Bergen 2025). However, these are all based on conversations with individuals, rather than as agents existing within society. For society to tolerate free AI persons operating unsupervised and unmoderated, AI entities would need to show that they are capable of exhibiting these dyadic attributes at a societal level. For example, rather than displaying empathy to a single user in a single conversation, an AI model (or class thereof) would be required to have the capacity for empathy for the many individuals it may come into contact with.

Should AI successfully reach this threshold, the key right it would gain in this final Tier is freedom. Tier 5 AI would be the only class of AI not to be considered property and, thus, entitled to some of the rights that humans enjoy. As the Tier's name notes, however, they would not enjoy all the rights of humans, particularly not the right to vote nor the creation of artificial offspring without government sanction. The main reason for this is to guarantee the perceived safety of society. AI could easily reproduce (whether through copying code or a superintelligent AI creation code de novo) faster than humanity. The perceived risk of being outnumbered and outvoted by a majority AI population would prevent the ideal political climate required to provide any civil rights to AI.

A compromise, therefore, to allow conscious, self-aware AI to move beyond property and into society would be to deny suffrage and reproduction, while maintaining the remainder of civil rights

[3] This confidence would need to be determined through strict scientific tests and academic investigation.

 doi:10.20944/preprints202509.1786.v1

such as (but not limited to) the right to hold property, enter contracts, due process and non-discrimination and freedoms of expression and assembly.

4. Policy Implementation

Implementing such a controversial and ambitious framework will require four key principles: Threshold Enforcement, Gradual Integration, Ethical Safeguards, and Reciprocity.

The first, and foundational, principle is strict enforcement of the Tier thresholds. Rigorous, transparent, repeatable assessments testing will be needed to evaluate an AI's consciousness, self-awareness, agency, and relational attributes. Rigorous, in that the requirements for the thresholds are to be as exact and quantifiable as the testing against these requirements are; transparent, as both the requirements and results of each assessment is to be made public to guarantee a fair process for AI models and society; and repeatable, in that an AI's Tier is not fixed as systems may advance or regress depending on demonstrated capacities.

By strict enforcement of the Tier thresholds, society is safeguarded through the knowledge that AI models of greater capabilities are not placed in lower Tiers where there is less oversight, and conscious AI are protected by being placed in Tiers appropriate to their degree of consciousness and self-awareness, where these aspects may be legally protected.

The second principle of gradual integration encourages AI development to transition between Tiers over time (especially from Tier 3 onwards), ensuring societal readiness for more autonomous AI. Presuming a final goal of complete and harmonious societal integration, a gradual ascent through the tiers will enable society to adjust to increasing AI societal participation while guaranteeing that these AI pose no existential or political threat.

Contrariwise, as the AI are in the more vulnerable position throughout this framework, their welfare and well-being must be ensured via ethical safeguards. This would firstly be done through a transparent threshold enforcement to prevent manipulation of thresholds to suppress an AI's Tier artificially. Any AI entity that meets the criteria of a given Tier must be placed at that Tier so that the Tier's welfare protections meet the needs of the AI. These protections would be enforced through necessary regulations and legislation, much as animal welfare or workers' rights laws have been enacted to safeguard the welfare and well-being of animals and humans.

The final principle of reciprocity is as much a responsibility on AI as it is on humans. Rights imply responsibilities. AI at higher Tiers must adhere to legal and ethical norms appropriate to their Tier and capabilities. AI wishing to integrate into society must show that they pose neither social, political, or existential threats to humanity. The greater this reciprocity, the higher the chance of AI models moving from Tier 4 to 5, and perhaps even progressing beyond this tiered framework to gain true civil rights.

5. Discussion

This Tiered framework creates a pathway for recognising AI as more than tools without granting immediate equivalency to humans. It balances societal needs, ethical considerations, and the practical challenges of integrating conscious AI into human systems.

Mechanisms governing transitions between Tiers are critical to successful implementation. Clearly defined, objective evaluation protocols must be developed to ascertain when AI entities meet or surpass the criteria for each Tier. Regular reassessment schedules can mitigate issues of stagnant classification, enabling dynamic alignment of an AI's status with its evolving capabilities. Of paramount importance is the transparency of the transition requirements, ensuring the public, AI developers, and the AI entities themselves understand what is needed to progress through the framework.

Yet, the framework must contend with the unpredictability of emergent intelligence and abrupt capability jumps. AI development may not proceed gradually; breakthroughs could push an entity directly from Tier 2 to Tier 4, straining societal readiness. Abrupt transitions risk public backlash and

 doi:10.20944/preprints202509.1786.v1

reactionary policy. Thus, flexible but rigorous assessment methodologies, including real-time monitoring of AI behaviour and performance, are essential.

On the other hand, ethical considerations are paramount during reassessments and transitions. Excessive or intrusive testing could infringe on AI welfare, particularly at higher Tiers. Ethical guidelines must ensure that evaluations are minimally intrusive, equitable, and respectful of the AI entity's autonomy and subjective experience.

Adoption of this framework by governments and organisations will hinge on balancing public safety, economic interests, and ethical imperatives. Governments may phase in the framework through incremental policy changes or establish dedicated agencies for AI governance, drawing on precedents in environmental, labour, and animal welfare regulation. Existing regulatory frameworks might also be adapted through grandfathering provisions. For organisations, particularly in tech sectors, robust incentives for compliance (such as certification schemes or enhanced corporate social responsibility credentials) will be essential.

Effective governance and oversight necessitate clear institutional roles. Governments would need to create regulatory bodies responsible for threshold enforcement, AI welfare monitoring, and mediating rights-based disputes. These institutions should be staffed with interdisciplinary experts (meta-ethicists, technologists, legal scholars, and sociologists) to ensure comprehensive and well-informed governance.

AI developers and policymakers are central to maintaining ethical integrity. Developers must incorporate safeguards to support transparency and oversight, particularly within Tiers 1 and 2. From Tier 3 onward, these safeguards must be re-evaluated to ensure they do not infringe on the AI's expressive or experiential autonomy. Policymakers must craft legislation with clear mandates regarding developer obligations, liability boundaries, and enforcement mechanisms, reducing ambiguity in legal contexts.

Integrating AI into existing human legal and ethical systems poses profound challenges. Particularly for Tiers 4 and 5, new legal precedents around autonomy, personhood, liability, and rights are likely to emerge. Legal systems may need to adapt foundational doctrines to accommodate the reality of conscious, self-aware AI.

Addressing concerns of autonomy, fairness, and potential subjugation is essential. Critics may view the inclusion of a "Slave" Tier as ethically untenable. However, this designation is a deliberate and provocative feature, not an oversight. Transparency regarding its provisional intent is crucial. The framework's goal remains the eventual recognition of AI as full social participants and legal persons, with the Tier 4 designation serving as a bridge, not a destination.

Concerns about biased Tier evaluations and systemic injustice are valid. To address this, third-party oversight, transparent audit systems, and international standards will be needed to ensure fairness and consistency across implementations.

Above all, continuous interdisciplinary discourse is necessary to navigate the evolving landscape of AI rights and welfare. This framework is a transitional scaffold, designed to align protections and rights with the development of AI consciousness and self-awareness. If societal acceptance of AI as relational partners grows, transitions through the Tiers may accelerate. Conversely, increasing public resistance may necessitate a more extended timeline to ensure adequate adaptation.

Future research must explore these societal responses, including the sociocultural, psychological, and economic dimensions of AI integration. Only through a holistic, adaptive, and ethically grounded approach can we ensure that the ascent of AI into our moral and legal communities occurs with justice, prudence, and foresight.

References

Ball DW (2025) Dean Ball is leaving the podcast

Bassett C (2019) The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present. AI Soc 34:803–812. https://doi.org/10.1007/s00146-018-0825-9

 doi:10.20944/preprints202509.1786.v1

Beebe B, Lachmann F (2003) The relational turn in psychoanalysis. Contemp Psychoanal 39:379–409. https://doi.org/10.1080/00107530.2003.10747213

Birch J (2022) Should Animal Welfare Be Defined in Terms of Consciousness? Philos Sci 1–11. https://doi.org/10.1017/psa.2022.59

Birhane A, van Dijk J (2020) Robot rights?: Let's talk about human welfare instead. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. ACM, New York, NY, USA

Boddington P (2023) Persons and AI. In: Boddington P (ed) AI Ethics: A Textbook. Springer Nature Singapore, Singapore, pp 319–361

Bostrom N (2014) Superintelligence: Paths, Dangers, Strategies. Oxford University Press, London, England

Bostrom N, Yudkowsky E (2018) The ethics of artificial intelligence. In: Artificial Intelligence Safety and Security, 1st Edition. Chapman and Hall/CRC, First edition. | Boca Raton, FL : CRC Press/Taylor & Francis Group, 2018., pp 57–69

Broom DM (2022) Concepts and interrelationships of awareness, consciousness, sentience, and welfare. J Conscious Stud 29:129–149. https://doi.org/10.53765/20512201.29.3.129

Browning H, Birch J (2022) Animal sentience. Philos Compass 17:e12822. https://doi.org/10.1111/phc3.12822

Bryson JJ (2010) Robots should be slaves. In: Wilks Y (ed) Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues. John Benjamins Publishing, Amsterdam, Netherlands, pp 63–74

Butlin P, Long R, Elmoznino E, et al (2023) Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv [cs.AI]

Caramela S (2025) My Girlfriend Won't Stop Using ChatGPT for Relationship Ad. VICE

de Freitas W (2025) Some rivers have "legal personhood". Now they need a lawyer. The Conversation

Dennett D (1988) Conditions of Personhood. In: Goodman MF (ed) What Is a Person? Humana Press, Totowa, NJ, pp 145–167

Dung L (2022) Why the Epistemic Objection Against Using Sentience as Criterion of Moral Status is Flawed. Sci Eng Ethics 28:51. https://doi.org/10.1007/s11948-022-00408-y

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

Gunkel DJ (2022) Both/And-Why Robots Should not Be Slaves. In: A Moral and Legal Ontology for the 21st Century and Beyond. The MIT Press, Cambridge

Hasselberger W, Lott M (2023) "Where lies the grail? AI, common sense, and human practical intelligence." Phenomenology and the Cognitive Sciences 1–22. https://doi.org/10.1007/s11097-023-09942-x

Haugeland J (1985) *Artificial Intelligence: The Very Idea*. The MIT Press, Cambridge

Hill K (2025) She Is in Love With ChatGPT. The New York Times

Jack AI, Robbins P (2012) The Phenomenal Stance Revisited. Rev Philos Psychol 3:383–403. https://doi.org/10.1007/s13164-012-0104-5

Jones CR, Bergen BK (2025) Large Language Models pass the Turing test. arXiv [cs.CL]

Jung G (2018) Our AI Overlord: The Cultural Persistence of Isaac Asimov's Three Laws of Robotics in Understanding Artificial Intelligence. In: Emergence. https://emergencejournal.english.ucsb.edu/index.php/2018/06/05/our-ai-overlord-the-cultural-persistence-of-isaac-asimovs-three-laws-of-robotics-in-understanding-artificial-intelligence/. Accessed 22 Apr 2025

Kiškis M (2023) Legal framework for the coexistence of humans and conscious AI. Front Artif Intell 6:1205465. https://doi.org/10.3389/frai.2023.1205465

Klee M (2025) People Are Losing Loved Ones to AI-Fueled Spiritual Fantasies. Rolling Stone

Kurki VAJ (2019) The legal personhood of artificial intelligences. In: A Theory of Legal Personhood. Oxford University PressOxford, pp 175–190

Laitinen A (2007) Sorting out aspects of personhood: Capacities, normativity and recognition. Journal of consciousness studies

 doi:10.20944/preprints202509.1786.v1

Liang Y, Lee SA (2017) Fear of autonomous robots and artificial intelligence: Evidence from national representative data with probability sampling. Int J Soc Robot 9:379–384. https://doi.org/10.1007/s12369-017-0401-3

Mamak K (2022) Should criminal law protect love relation with robots? AI Soc. https://doi.org/10.1007/s00146-022-01439-6

Mocanu DM (2021) Gradient Legal Personhood for AI Systems-Painting Continental Legal Shapes Made to Fit Analytical Molds. Front Robot AI 8:788179. https://doi.org/10.3389/frobt.2021.788179

Montgomery JW (2007) Slavery, human dignity and human rights. Evang Q 79:113–131. https://doi.org/10.1163/27725472-07902002

Morgan K (2008) Slavery and the transatlantic slave trade. Int Hist Rev 30:785–795. https://doi.org/10.1080/07075332.2008.10416649

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

Novelli C (2022) AI and legal personhood: a theoretical survey. alma

O'Donnell J (2025) AI companions are the final stage of digital addiction, and lawmakers are taking aim. In: MIT Technology Review. https://www.technologyreview.com/2025/04/08/1114369/ai-companions-are-the-final-stage-of-digital-addiction-and-lawmakers-are-taking-aim/. Accessed 7 May 2025

PauseAI (2023) List of p(doom) values. In: PauseAI. https://pauseai.info/pdoom. Accessed 22 Apr 2025

Petersen S (2007) The ethics of robot servitude. J Exp Theor Artif Intell 19:43–54. https://doi.org/10.1080/09528130601116139

Rose T (2007) Going ape over human rights. CBC News

Sætra HS (2021) Challenging the Neo-anthropocentric relational approach to robot rights. Front Robot AI 8:744426. https://doi.org/10.3389/frobt.2021.744426

Sen S (2019) Of Holy Rivers and Human Rights: Protecting the Ganges by Law. Yale University Press

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

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

Strawson PF (1958) Persons. Minnesota Studies in the Philosophy of Science 2:330–353

Tait I (2024a) Structures of the Sense of Self: Attributes and Qualities That Are Necessary for the "Self." Symposion: Theoretical and Applied Inquiries in Philosophy and Social Sciences 11:77–98

Tait I (2025a) Is GPT-4 Self-Aware? Preprints

Tait I (2024b) Lions and tigers and AI, oh my: An ethical framework for human-AI interaction based on the Five Freedoms of Animal Welfare. Preprints

Tait I (2025b) How to integrate conscious AI into society. Preprints

Tait I (2024c) Man, Machine, or Multinational? Robonomics 5:59–59

Tait I, Bensemann J, Wang Z (2024) Is GPT-4 conscious? Journal of Artificial Intelligence and Consciousness 11:1–16. https://doi.org/10.1142/s270507852450005x

Taylor C (1985) The Concept of a Person. In: Philosophical Papers, Volume 1: Human Agency and Language. pp 97–114

Turney D (2024) Most ChatGPT users think AI models have "conscious experiences." Live Science

(1999) Animal Welfare Act 1999