Article

Not peer-reviewed version

How to Integrate Conscious AI into Society

Izak Tait *

Posted Date: 4 March 2025

doi: 10.20944/preprints202503.0175.v1

Keywords: Human-AI-Interaction; ethical frameworks; legal personhood; moral agency; civil rights

Article

How to Integrate Conscious AI into Society

Izak Tait

Auckland University of Technology, Auckland, New Zealand, 1010; izak.tait@autuni.ac.nz

Abstract: As artificial intelligence approaches consciousness, integrating conscious AI agents (CAI) into human society becomes an urgent challenge. This paper investigates four frameworks for such integration: Animal Welfare Protections, Legal Persons under a Human Directorate, Paternalistic Second-Class Citizenship, and Civil Rights. Employing a formalised formula combining public perception surveys, historical legislative precedents, and perceived risks, we calculate the probability of successful implementation for each framework. The results reveal that the Animal Welfare framework offers the quickest path to societal acceptance. However, this approach raises ethical concerns by granting CAI minimal rights. Conversely, the Civil Rights framework, while ethically robust by granting CAI full legal equality, requires a significantly longer period to gain public approval. To reconcile speed with ethical considerations, we propose a transitional policy strategy that begins with more restrictive frameworks and gradually progresses toward full civil rights for CAI. This phased approach balances societal security with the moral imperative to recognize CAI autonomy. Our findings underscore the complex interplay between practicality and ethics in integrating conscious AI, highlighting the need for proactive policies and continued interdisciplinary research to ensure that the integration of future CAI entities contributes positively to society.

Keywords: human-AI-interaction; ethical frameworks; legal personhood; moral agency; civil rights

1. Introduction

If and when AI models (whether disembodied chatbots or robots) become conscious, sentient, and self-aware agents, they will need to be integrated into society (While there is debate about whether AI can be determined to be a moral patient and agent (Gunkel 2012), the paper will presume that determination has been reached and that integration of some sort is required). This paper will investigate four different frameworks that may be used for integrating such conscious (This paper will not make a claim as to the nature of consciousness, other than that it is sufficient for an entity to be worthy of moral status as conscious entities have a capacity to feel pleasure and pain). AI agents (CAI) and determine the most practical pathway for harmonious CAI societal integration.

The paper will use a series of formalised arguments to arrive at a prior probability figure for the chance of success each of the four frameworks has to integrate CAI into human society, and how these can be increased through proposed policy decisions post-implementation.

The four frameworks referenced below are inspired by extant or historical legislation for the integration of human social groups and non-human entities. They thus serve as analogies to what future policy and legislation frameworks may implement. Rather than assessing specific frameworks put forth for CAI integration or civil rights that exist in the literature, this paper will keep to a broad brush and reference published literature under each of the four frameworks.

The frameworks below can, therefore, best be thought of as four complementary yet distinct, sets of approaches that future policy and legislation can take when integrating CAI.

The four referenced frameworks are animal welfare protections, corporations as legal persons under a human director, paternalistic second-class citizenship and slavery, and civil rights. Each framework offers a distinct approach to granting rights and responsibilities to CAI entities, ranging from providing basic protections akin to those for animals to recognizing full legal personhood equivalent to humans.

 doi:10.20944/preprints202503.0175.v1

AI models exist purely as tools as products; property that is owned and without any rights or moral status. The animal welfare framework extends this moral consideration to CAI without equating them fully with humans; therefore, sidestepping the issue of rights. By leveraging existing animal welfare laws and ethical principles, this framework provides a pragmatic starting point for integrating CAI into society by acknowledging the moral responsibility to prevent suffering in conscious beings but making no claim to personhood.

The human directorate framework draws on the concept of legal personhood, where corporations are granted certain rights and responsibilities akin to natural persons but are ultimately governed by human directors. This approach offers a model where CAI entities could operate with a degree of autonomy and enter into legal agreements while remaining under human oversight. It balances the need for CAI to function independently in certain capacities with society's interest in maintaining accountability and control. This framework would facilitate the integration of CAI by providing them with legal recognition and operational freedom within a secure structure that ensures human responsibility for their actions.

The paternalism framework is based on historical legal cases of slavery and segregation. This ethically contentious approach is included from a pragmatic view of the chauvinistic elements of humanity, who would be antagonistic to a legal framework of equality or equity for CAI. Appeasing such a voting bloc would be politically advantageous for certain political groups, who may thus lobby for a framework that sets CAI as second-class citizens, with a restricted degree of autonomy to ensure human security and supremacy.

Lastly, the civil rights framework represents the extension of full legal and moral equality to all individuals, regardless of inherent differences. Such an approach would perhaps be the simplest bureaucratically speaking, but also perhaps the most difficult socially due to the aforementioned chauvinistic elements of society. However, it has significant precedence in the civil rights and suffrage movements of the past, and has had the most attention in academic publications related to AI rights.

The choice of which approach society will one day take will have profound implications for both human society and the CAI entities themselves, affecting aspects such as autonomy, legal rights, societal acceptance, and the balance between security and agency.

For all the frameworks, this paper will use a formula to calculate the probability of their successful implementation based on extant and historical legislation as well as surveys performed to gauge public perception of AI rights and moral consideration. This enables a systematic and quantitative assessment of each framework's likelihood of successful integration through a data-driven comparison of the compromise between societal attitudes and political bureaucracy, highlighting not only the frameworks' theoretical viability but also their practical acceptability within society.

In the context of integrating conscious AI into society, this paper defines 'success' of a framework as the sustained approval from the majority of a population (best-case scenario would be 90% as per the Gilens model (Gilens 2012)) to ensure that legislation to enact a framework has the political mandate for both implementation and long-term viability.

The paper is structured as follows: Section 2 outlines the methodology for assessing the integration frameworks, detailing the hypothesis-testing formula and the rationale behind the chosen variables and weights. Sections 2.1 through 2.4 delve into each integration framework individually, calculating the probabilities of successful implementation and discussing the implications of the results. Section 3 extends the analysis beyond initial implementation, exploring policies and proactive systems that could influence the long-term success and societal acceptance of the frameworks. Section 4 discusses the findings, addressing the ethical considerations and proposing strategies for transitioning between frameworks to achieve a balance between CAI autonomy and societal security. Finally, Section 5 concludes the paper by summarising key insights and suggesting avenues for future research.

 doi:10.20944/preprints202503.0175.v1

By systematically evaluating these frameworks and proposing a method for their assessment, this paper will contribute to the meaningful discourse on the ethical integration of conscious AI into human society. The findings will inform policymakers, technologists, ethicists, and the broader public about the potential pathways and challenges ahead, ultimately guiding responsible decision-making in the face of this emerging reality.

2. Assessing Frameworks of Integration

To create the prior probability figures for the successful implementation of the society-CAI integration frameworks, a formula will be used that takes into account the following:

- Public Perceptions (*p(B)*): The views of both experts and the general public regarding AI consciousness and rights.
- Historical Legislative Data (*p(E)*): Analogous cases of similar legislative frameworks from G7 nations (used here as a representative sample of liberal democracies).
- Perceived Risks (*r(t)*): Public and expert assessments of the existential or societal risks associated with AI.

While other factors such as economic incentives and lobbying efforts would also influence legislative success, the model simplifies these by presuming they are mediated through public perception and rate of legislative enactment.

The formula is thus:

$$p(H) = \left(w_B(t) \cdot p(B_t) + w_E(t) \cdot p(E_t)\right) \cdot \frac{1}{r(t)}$$

Where H is the hypothesis in question (i.e. the integration framework), B is the public's perception towards AI behaviour, and E is the extant and historical legislation. Both *p(B)* and *p(E)* will be weighted, and $p(B_t) = p(B_0) \cdot e^{\lambda Bt}$ and $p(E_t) = p(E_0) \cdot e^{\lambda Et}$ such that $p(B_0)$ and $p(E_0)$ are the initial probabilities of *p(B)* and *p(E)* respectively; and λ shows the positive growth constants for both *p(B)* and *p(E)*. The *r* here shows the perceived risk that AI can cause to human societies, with $r(t) = e^{\gamma t}$ adjusting the probabilities downward as time increases, reflecting growing uncertainty. Both λ and γ can be expressed as percentages.

The weighting for all sections will be 3-to-2 in favour of *p(B)* as the survey responses directly link to the moral considerations of conscious/sentient artificial beings, while the legislation used are analogous but not homologous to the proposed frameworks.

The *p(B)* values below are based on surveys in the literature where participants expressed their perceptions of current and future AI using a variety of methods (from percentile approval figures to Likert scales) (All figures from the surveys can be found in the supplementary materials). The surveys were predominantly based on US participants, with some broader online participation.

All averaged responses have been transformed into figures between 0 and 1.

To arrive at a coherent *p(E)* figure, the speed at which the relevant legislation passed through all G7 nations plus the length of time that the legislation has been in effect was used, expressed as:

$$\sqrt{\left(\frac{Current\ year\ -\ Most\ recent\ legislative\ date}{Current\ year\ -\ Baseline\ date} + \left(1 - \frac{Most\ recent\ legislative\ date\ -\ Earliest\ legislative\ date}{Most\ recent\ legislative\ date\ -\ Baseline\ date}\right)\right) / 2}$$

The above "Baseline date" was chosen independently for each integration framework, based on the earliest well-recorded date in one of the G7 nations when a movement was founded based on the nature of the framework. For instance, with the animal welfare framework, 1824 was selected as the baseline year, as it was the first recorded date that a movement for animal welfare was founded in the G7 nations, specifically the Society for the Prevention of Cruelty to Animals (SPCA).

The baseline date serves as a point of reference for when the core idea of the framework became publicly widespread, but not yet enshrined in legislation. The time between these two points are

 doi:10.20944/preprints202503.0175.v1

analogous to current discussions on AI rights and moral consideration and when legislation may be implemented in the future.

The risk factor *r*, as with *p(B)*, has been taken from survey results of academics' and the public's perception and attitudes towards AI existential risk and transformed using a hazard rate formula to show the perceived risk per year (For risk values provided without an explicit time horizon, 50 years were used for the formulation as a conservative estimate of the end of the average respondent's lifespan.) of an AI-caused existential event:

$$r = 1 - (1 - P)^{1/n}$$

Where P is the public perception of the risk, and n is the number of years between when the survey was conducted and the time horizon of the perceived risk. Unlike the rest of the formulation's parts, the risk factor's initial value will remain the same for all four frameworks (that is, 0.0062) as it is independent of the four frameworks.

For λB, the time for legislation to first be passed was taken as an inverse of 1. Similarly, the time for the legislation to finish passing through all seven nations was taken as an inverse of 1 for λE.

The overall formula, taken holistically, will then produce a figure showing the probability of a successful integration. Put another way, the end result can be taken as a speculative "approval rating" at a given point in time, starting from the year that the integration framework is enacted. Because of the λ and γ growth constants, one can adjust the formula to determine the speculative approval rating society would have on the frameworks in any given year following the frameworks' implementation. This paper will use this to determine how long it would take (if at all) for the majority of society to approve of the framework.

Specifically, the approval rating, and thus *p(H)*, of most interest for a framework to be judged a success is 0.9, as policy changes and legislative implementation require (on average) a 90% approval rating to have a 50% chance of being adopted and enacted (Gilens 2012).

2.1. Animal Welfare

While, as Ziesche & Yampolskiy mentions, AI welfare is a recent field of inquiry (Ziesche and Yampolskiy 2018), creating an integration framework based on current animal welfare models may seem an attractive option, as it would provide CAI entities with legal protections, and there would be a clear delineation between human and machine agents (that may remain as property (Calverley 2006)) that would assuage anthropocentric chauvinist groups.

There is a wealth of ethical frameworks and legislative regulations that deal with the use of, and actions towards, animals that could be used as a foundation for such an integration framework, most notably the Five Freedoms of Animal Welfare. This framework accounts for an animal's physical and mental health, ensuring an animal has freedom from disease, hunger, pain, stress, fear and inhibited behaviour (Mellor 2016).

One can translate these animal freedoms into subject-neutral terms (such as changing injury and disease to malfunction and system degradation) that could then be applied to AI in such a way as to provide AI with equal protection under the law as that society currently grants animals (Tait 2024a). With the extent of legal protections that G7 nations currently provide animals, integrating protections for CAI entities into this legislation would ensure the physical and mental safety of AI entities.

One can argue that CAI would, in this speculative circumstance, be better protected by governments than humans are. This notion is somewhat supported by the result of the hypothesis testing formula.

The survey questions used for this section (The transformed survey responses for all sections can be found in full in the supplementary material.) deal with issues such as the right to life (Lima et al. 2020), control of mental states (Mays et al. 2024), opposition to damage and punishment (Pauketat and Anthis 2022; Anthis et al. 2024), and a desire for protection (Martínez and Winter 2021).

The item with the highest approval rating from survey participants acknowledged that torturing sentient artificial beings was wrong (with 76% of respondents agreeing), while the lowest stated that

 doi:10.20944/preprints202503.0175.v1

artificial beings ought to be given equal moral consideration as humans (with only 27% of surveyed respondents agreeing).

Overall, the average approval rating for all items related to treating AI on par with current animal welfare ethical frameworks was 47%, giving us the *p(B)* for this section.

To calculate the *p(E)* value (The dates of legislations for all sections can be found in the supplementary material.), the founding of the Society for the Prevention of Cruelty to Animals in 1824 was used as a baseline date, with the United Kingdom following as the first G7 nation to pass animal welfare legislation in 1835, and Japan as the final nation in 1973. With these figures in hand, we can calculate the *p(E)* value as:

$$\sqrt{\left(\frac{2024 - 1973}{2024 - 1824} + \left(1 - \frac{1973 - 1835}{1973 - 1824}\right)\right) / 2} = \sqrt{(0.26 + 0.07) / 2} = 0.41$$

With these values, we can complete the formula from earlier to determine the likelihood of success of the speculative framework in the year of its legislative passing:

$$p(H) = (0.6 \cdot 0.47 + 0.4 \cdot 0.41) = 0.44$$

A mere 44% chance of successful integration does not inspire confidence, but as mentioned above, should one view animal welfare protections as stronger than human well-being efforts, it can be rationalised (if only post hoc). However, with the λ values, we can forecast how many years it would take to reach any arbitrary probability of success, such as 0.5, 0.67 or a complete 100% chance of success.

To calculate the final λ values, 1/(1973-1835) will provide a λE of 0.0072, and 1/(1835-1824) will provide λB of 0.0909.

With these, and the risk value of 0.0062, we can see that it would take at least two years to reach a 50% chance of successful integration:

$$p(H) = (0.6 \cdot 0.47 \cdot e^{0.0909 \cdot 2} + 0.4 \cdot 0.41 \cdot e^{0.0072 \cdot 2}) \cdot \frac{1}{e^{0.0062 \cdot 2}} = 0.5001$$

Not to repeat the formula again, it would take eleven years post-legislation for the framework to gain the required 90% approval rating as per Gilens' calculations (Gilens 2012) to give the framework a 50% chance of successful implementation.

2.2. Human Directorate

A framework which would give AI more legal rights may be based on the current legal philosophy of "legal personality", predominantly found in corporate law (Laukyte 2019; Nanos 2020). Unlike humans as "natural persons", a "legal person" such as a corporation may enter into contracts and legal agreements, become members of incorporated societies, be held legally liable and bring legal suits to court for any damages incurred. Despite the ongoing debate about whether AI can or should be subject to legal personality (Čerka et al. 2017; van den Hoven van Genderen 2018; Dremliuga et al. 2019; Solum 2020; Chesterman 2020; Jowitt 2021; Novelli 2023), CAI may find such a framework more appealing than an animal welfare-based framework that offers protection but no active rights. Society may also deem CAI to be worthy of the increased autonomy under this framework if they are deemed to be capable of acting equally ethically to humans (Chu and Liu 2023).

On the other hand, all legal persons have representatives who are equally held liable for the legal person's actions. A CEO or director is legally responsible for the actions committed by their corporation or company. This human-in-the-loop aspect would be an attractive feature of a directorate framework to society as it ensures that all actions of a CAI agent are overseen by a human who is incentivised to ensure the AI's actions are aligned to societal norms, termed "dependent legal personality" (Chopra and White 2011). Such a directorate framework can also draw from the legally

 doi:10.20944/preprints202503.0175.v1

protected governance of a parent over their children to create a mutually beneficial relationship between the CAI and its director, who would then have a fiduciary duty towards the AI (Tait 2024b).

The survey questions selected for this section thus dealt with AI's right to file a lawsuit (Martínez and Winter 2021), the right to receive payment for work (Mays et al. 2024), the right to enter into contracts (Lima et al. 2020), the right to own their own programming (Anthis et al. 2024), and the right of humans to own and direct artificial beings (Pauketat and Anthis 2022).

Of these, the aspect that most respondents in their respective surveys agreed on was the right for humans to own and trade artificial entities at 71% approval, while the least agreed upon aspect was the right for AI to be granted citizenship of their country of residence, at 27%. The average approval rating (and thus the *p(B)* value for this section) was 40%.

The speed of corporate legislation through the G7 nations was dramatically faster than for animal welfare, taking a mere 55 years, from the United Kingdom's Join Stock Act in 1844 (alongside a Supreme Court ruling in the USA of the same year) through to Japan's Shōhō (Commercial Code) of 1899. The baseline date for this section was 1776, the year of publication of Adam Smith's seminal Wealth of Nations, which laid the philosophical foundations for modern corporate law. With 2024 as the current date, we can determine the *p(E)* value as:

$$\sqrt{\left(\frac{125}{248} + \left(1 - \frac{55}{123}\right)\right) / 2} = \sqrt{(0.50 + 0.55) / 2} = 0.73$$

And, thus, the initial p(H) at the time of proposed passing would be:

$$p(H) = (0.6 \cdot 0.4 + 0.4 \cdot 0.73) = 0.53$$

Which seems far more satisfactory than the previous section, as the framework already has a greater chance of success than failure. Using the same λ calculations as before, there is a λE of 0.0182 and a λB of 0.0147. Collective, these λ values are lower than those related to animal welfare, which means it would take a hypothetical 40 years to reach a *p(H)* of 0.9:

$$p(H) = (0.6 \cdot 0.4 \cdot e^{0.0147 \cdot 40} + 0.4 \cdot 0.73 \cdot e^{0.0182 \cdot 40}) \cdot \frac{1}{e^{0.0062 \cdot 40}} = 0.903$$

2.3. Paternalism

A framework that would provide CAI agents with greater freedoms than the two listed above, and diffuse the human control over AI from the individual to the societal level, would be one based on paternalism ('Paternalism' here refers to governance or policy measures wherein a higher authority imposes restrictions or controls on individuals or entities; in this case where conscious AI is afforded limited autonomy under human oversight). Treating CAI agents as second-class citizens or slaves would provide them with greater autonomy at the operational level while ensuring they have little to no influence at the strategic level of human societies.

Paternalism, whether it is slavery, racial segregation, apartheid, or caste systems, has an understandably unpopular history as it contradicts the theological and humanistic philosophical claims of human dignity and equality (which explains why paternalistic frameworks of human-AI interactions are not seen in the literature). If all humans are born equal, then it would be immoral for one to own another or to be in a class above another.

However, AI are patently not human, even if modern state-of-the-art models seem quite human-like in their communication with us. Treating AI as second-class citizens, therefore, may appease the human chauvinists and gain greater approval amongst society (It must be noted that, unlike the other frameworks, a paternalism-based integration framework would predominantly be designed to gain societal approval rather than to work in the efforts of AI welfare or wellbeing). It can be argued that owning an AI may, at the object level, be the same as owning a pet or a tool (Bryson 2010) and that

 doi:10.20944/preprints202503.0175.v1

treating them as lower caste would still provide them more autonomy than human prisoners, and may indeed provide them with a legal personhood (Nanos 2020).

Survey responses affecting the p(B) value for this section would thus be about the right of humans to own and possess AI (Pauketat and Anthis 2022), but also about the greater autonomy of AI in a paternalistic framework, such as privacy (Lima et al. 2020) and right to work (Mays et al. 2024). Unsurprisingly, owning AI had the highest approval rating of 71%, and allowing AI control over their own programming had the lowest at 25%.

The overall approval rating, and $p(B)$ value, was 44%.

As no G7 nation has extant laws surrounding slavery or racial segregation, the formula for determining the $p(E)$ value must be slightly altered. Rather than using the current year, the longest duration from the baseline date (Norman England's feudalism of 1066) to the end of slavery in the G7 nations was used. This figure (881 years for the United Kingdom), coupled with the average duration that slavery existed in the G7 (181 years), gives a $p(E)$ of 0.671 for legislated slavery.

The same was done for the period for which G7 nations had racial segregation to provide a $p(E)$ of 0.662. The average of these is, therefore, 0.666

The initial p(H) would thus be:

$$p(H) = (0.6 \cdot 0.44 + 0.4 \cdot 0.666) = 0.53$$

On par with the human-directorate framework above, and an adequate majority approval. The λ values for this section are also calculated differently. For the λB value, the time since the last G7 nation abolished slavery and segregation was used to provide a value of 0.025. As the years continue (presumably without a return to slavery and segregation), this figure will become smaller, showing society's lack of appetite thereof.

For λE, to show the strength of the legislation of its time, the shortest duration was taken over the longest duration, providing a figure of 0.037. With these values, the framework would attain a $p(H)$>0.9 in nineteen years:

$$p(H) = (0.6 \cdot 0.44 \cdot e^{0.025 \cdot 19} + 0.4 \cdot 0.666 \cdot e^{0.037 \cdot 19}) \cdot \frac{1}{e^{0.0062} \cdot 19} = 0.903$$

2.4. Civil Rights

An integration framework based on civil rights is perhaps the most intuitive to understand, as it involves providing CAI entities with equal legal rights to humans. Should future CAI have a dominant humanoid embodiment, this connection may seem even more intuitive (Bontula et al. 2024). Unfortunately, the requirement for equality may also ensure this is the most difficult framework for which to build popular support. Animals, such as chimpanzees and gorillas, have shown the capacity to understand and express themselves in human language. Yet, while the great apes have been granted personhood status in many nations, they do not have civil rights. CAI agents would be far more intelligent than the near-human apes, but also far removed from us in terms of biology and history, which could serve as a barrier to granting them full legal rights.

However, the intuitive nature of civil rights has meant it has received the greatest attention in the literature, from discussions as to whether AI should or can have rights (Coeckelbergh 2010; Gunkel 2020; Bennett and Daly 2020; Mamak 2022), to whether human rights ought to be prioritised (Bryson et al. 2017; Birhane and van Dijk 2020; Gunkel 2021), to the operations and ramifications of granting AI rights (Osborne 2021; Yanke 2021; Gordon and Pasvenskiene 2021), and more.

For the survey responses in this section, the questions unsurprisingly involved universal suffrage (Mays et al. 2024), legal and moral human rights (Guingrich and Graziano 2024; Anthis et al. 2024), and the respondents' willingness to join movements for AI civil rights (Pauketat and Anthis 2022). All told, the average approval response was 40%, tied last of the four framework's responses alongside the human-directorate survey responses.

 doi:10.20944/preprints202503.0175.v1

For the *p(E)* value, two different sets of legislation were used, much like the previous section. In this instance, the abolition of slavery in the G7 nations, as well as legislation granting full and equal civil rights. For the former, 1315 was used as the baseline, when France first abolished slavery on mainland France; and for the latter, 1215 was used as the date the Magna Carta was signed, seen as the first movement for a type of legal equality in the G7 nations.

It took 329 years for the G7 to abolish slavery, beginning with France in 1794 (entirely this time) and ending with Germany's defeat in the First World War, ending slavery in its colonies. Comparatively, civil rights legislation breezed through the G7, taking only 36 years from the United Kingdom Representation of the People Act in 1928 to the United States' Civil Rights Act in 1964. Put together, these two sets of legislation work together to produce a *p(E)* of 0.63. This means that the initial *p(H)* would be:

$$p(H) = (0.6 \cdot 0.4 + 0.4 \cdot 0.63) = 0.49$$

Had this calculation merely used the civil rights legislation and not the abolition laws, this initial *p(H)* would have been 0.53. However, with the λE of 0.0154 and λB of 0.0025, we can determine how long it would take to reach a *p(H)* of 0.5 and 0.9. While it would only take two years to reach a *p(H)* of 0.5 (equal to the animal welfare section), it would take a colossal ninety-four years to reach a *p(H)* of 0.9:

$$p(H) = (0.6 \cdot 0.4 \cdot e^{0.0154 \cdot 94} + 0.4 \cdot 0.63 \cdot e^{0.0025 \cdot 94}) \cdot \frac{1}{e^{0.0062} \cdot 94} = 0.903$$

Even if one were to completely remove the risk factor from this equation, it would still take a substantial fifty-nine years for the p(H) to reach 0.9. This enormous figure is due to the low λ values which, in turn, is due in part to the large disparity between the baseline date and when the legislation was passed. While movements for equality were seen quite early in history, a political will for legislation was slow to arrive.

3. Beyond Implementation

Based strictly on the *p(H)* values for when each framework achieves 0.9 or higher, there appears to be a map presented for how to best integrate CAI into human society: begin by providing them with protections for their welfare and well-being without any associated legal rights, then treat them as second-class citizens, before establishing human control and direction over them, and finally granting them with full legal and civil rights equal to humans.

However, the hypothesis-testing formula only forecasts the probability of success of a framework from the moment of implementation without regard for any event that transpires post-implementation.

To speculate the continued success of the framework, and to validate how it will work based on (re)actions by humans and CAI, we can use this simple formula

$$p(H') = p(H) \cdot F \text{ such that } F = 1 + \sum_{i=1}^{n} \delta_i$$

Where *F* is the effect of any speculative event that occurs after any framework has passed through legislation, and δ_i is the event in question, expressed as a number between -1 and 1, representing the number of humans affected by the event (Thus, a 1 would indicate 100% of all people positively affected, -0.5 would indicate 50% of the world's population negatively affected, and so on). To avoid biased extremes when speculating on the effect of a theorised event, we can perform a log transformation on this number to ensure that answers naturally fall on a bell curve resting on 0, as so:

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

For example, should we speculate that a series of events would occur that positively affects 40% of society (Degrees of impact is overlooked here, but can be implemented in the equation for a more

 doi:10.20944/preprints202503.0175.v1

granular view.) seven years after the G7 passed the animal welfare framework, we can estimate the effect on the probability of a successful integration as:

$$p(H') = 0.7 * (1 + sgn(0.4) \cdot ln(1 + |0.4|) = 0.7 * 1.34 = 0.93$$

Thanks to one year's events, this example's new $p(H')$ of 0.93 would push the 'schedule' of public acceptance four years ahead.

However, if a state were to truly wish to increase the $p(H)$ of each framework to >0.9, they would require events with a net-positive effect on all of humanity due to the low initial $p(H)$ values. This is an insurmountable challenge for any policy proposal. However, with each year that passes, the $p(H)$ will naturally increase, meaning that a smaller percentage of the population is required to have a net benefit from any δ_i event. This positive feedback loop, if successfully enacted, will mean that the required policy enforcement will be less draconian over time. The stronger the initial policy adoption, the greater the long-term impact it will have for successive policies to build on in a more inclusive and less authoritarian manner.

If we were to aim for a more reasonable target, such as 50% of humanity (presuming the framework would illicit a partisan response, then a state would need only to set a target of six years for the animal welfare framework's general approval to be high enough for the policy to target 50% of humanity, fifteen years for the human directorate framework, seven years for the paternalism framework, and forty-three years for the civil rights framework.

From an anthropocentric perspective, the animal welfare and paternalism frameworks provide human society with the most control over CAI entities and, thus, ensure the most safety. However, from the CAI point of view, this is the reverse: being treated akin to animals or second-class citizens provides them with the least autonomy. There is, therefore, tension between safety and agency. There is a possibility for this tension to escalate into conflict, whether between humans and CAI, between human chauvinists and CAI-sympathisers, between CAI who struggle for freedom and those who collaborate, and any combination of the above.

This may particularly be the case with the paternalism framework. Slavery and segregation are distasteful in the modern liberal democracies of the G7 nations; however, the history of slavery from antiquity through to the twentieth century stands as an uncomfortable reminder of how at ease society can be with enslaving those it deems inferior or "other" and how successful such a policy have been in the past. No current G7 nation practices slavery, and all seven nations have law enforcement units or departments to combat modern instances of slavery, such as human trafficking. Thus, while one may imagine that a substantial subset of society would have little qualms about enslaving conscious machine agents(who are more "other" to humanity than any vertebrate species), another substantial portion of society, as well as most CAI, would be against the concept.

In light of this, the $p(H')$ formula must be unique to each framework, as a blanket δ_i event taken across all frameworks would adjust the success probability of the opposing paternalism and civil rights frameworks equally. There must exist a subset of δ_i where actions that impact $p(H_P)$ negatively correlate with $p(H_{CR})$ and vice versa. Additionally, there must exist another subset of actions which positively impact $p(H_P)$ which have no effect on $p(H_{CR})$, and in reverse (The same would be true for interactions between all four frameworks). We can formalise this as:

$$x \subset \delta_i \rightarrow (p(H'_P) > p(H_P)) \wedge (p(H'_{CR}) < p(H_{CR}))$$

$$(x \subset \delta_i \rightarrow p(H'_P) > p(H_P)) \cap (y \subset \delta_i \rightarrow p(H'_{CR}) > p(H_{CR})) = \emptyset$$

Another way of putting this would be to show the subset of events that promote CAI interests (δ_{CAI}) that do not intersect with, or negatively impact human society; and similar for events that promote human society's interests (δ_{HS}). We can presume that δ_{CAI} would correlate positively with $p(H_{AW})$ and $p(H_{CR})$, and that δ_{HS} would correlate with $p(H_{HD})$ and $p(H_P)$.

This means that the successful framework (of at least $p(H') > 0.9$) will need to be aided by those specific events that positively impact its probability of success, while simultaneously not raising a competing framework (In addition to those δ_i which raise the p(H) of all frameworks).

 doi:10.20944/preprints202503.0175.v1

There is considerable difficulty in forecasting specific unknown future events, as we do not know in what manner CAI or society will react to the individual frameworks. However, we can encourage types of events through policy direction. As noted above, there are two competing sets that will need to be considered. First, whatever policy society wishes to implement will need to encourage events that positively impact its probability of success, while either remaining neutral or negative to other frameworks. Secondly, the degree to which events favour CAI autonomy versus human societal security needs to be accounted for, as extremes of either attribute may cause the aforementioned escalatory tensions. We can formalise such a policy as:

$$\pi = argmax_{\pi}\mathbb{E}\left[\sum_{t=0}^{T} w \cdot (\delta_{FW} - \neg\delta_{FW})(t) + w \cdot (\delta_{CAI} - \delta_{HS})(t)\right]$$

Where δ_{FW} are events which positively impact the framework in question and $\neg\delta_{FW}$ are events which negatively impact the framework in question. The δ_{CAI} and δ_{HS} values can be swapped in the above equation depending on the framework in question.

An example of such a policy (presuming the Animal Welfare framework) would be to create a CAU Welfare Standards Authority to develop and enforce strict guidelines for how CAIs are treated. This body would ensure that CAIs are not subjected to harmful conditions while being useful to human society. Drawing on extant animal welfare regulatory bodies, such an authority would be able to fund research on human-CAI interactions, incentivise humane treatment of CAI, set limits to CAI utility, and assess the risk of any CAI entity to humanity.

Regardless of the chosen framework, there are some common elements that a successful policy would need to implement. As shown above, an (inter)national agency that acts as an authority regarding standards of practice would be a necessity. This could mirror the WTO for the Human Directorate framework, or be based on the United Nations Human Rights Council for the Civil Rights framework. Such a standards authority can lay the regulatory groundwork needed to incentivise academia, industry and commercial peak bodies, and broader society to adopt behaviours that benefit the framework while discouraging anti-framework behaviours through regulatory enforcement. A standards authority would also be agnostic in terms of autonomy versus security, and can thus affect both equally. We can show this authority as:

$$\frac{\delta_{FW}}{\neg\delta_{FW}} = f(auth)$$

$$f(auth) = f'(age, sec) = age \cdot sec$$ (‘auth’ = standards authority; ‘age’ = autonomy; ‘sec’ = security.)

A second proactive system related to the above is a mediation and arbitration board. Regardless of the implemented framework, there will always be some incidents of conflict between humans and CAI. Thus, proactively setting up a board to mediate and arbitrate on these incidents in a transparent and fair manner will reduce the probability of negative events escalating to disastrous levels. One can therefore think of the mediation board as imposing a soft limit on the scale of the potential negative effects:

$$\neg\delta_{FW} = \frac{(\neg\delta_{FW})_0}{1 + \frac{(\neg\delta_{FW})_0}{mediation}}$$

Both of these systems would also be scalable and transferable to another framework should society decide to implement another one.

A third and fourth system would be administrative in nature, but far-reaching. Implementing an “autonomy certification” system coupled with a mandatory insurance scheme based on the autonomy of the CAI would provide a legislative and regulatory foundation to increase CAI

 doi:10.20944/preprints202503.0175.v1

autonomy where necessary, while providing society with a financial and judicial safety net in terms of insurance. The greater the autonomy of a CAI entity (as evidenced by its registration), the greater the insurance is placed on its actions. One can even place an arbitrary "0 level" autonomy, such that if a CAI's autonomy is less than this (such as through strict ownership), then the CAI would have insurance should anything be done to it.

In this way, one can offset the potential costs to society with the benefits to CAI in a somewhat utilitarian calculus method:

$$SocietalCost = f(autonomy, insurance) = 0$$

With these four methods, a policy would likely maximise the chances of positive events occurring that meet the needs of the framework and discourage negative events, while balancing CAI autonomy and societal security.

4. Discussion

The paper thus far discussed thus far the likelihood of success for implementing each framework and how a state could continue working on ensuring the framework's ongoing success. Yet, the central question has not been answered: how should CAI be integrated into society?

Going by the data, the answer is both politically incorrect and chauvinistic: treat AI as property, animals and second-class citizens before granting them legal personality and then natural rights. This is based on the speculated speed at which the frameworks achieve a $p(H)$ value of >0.9. It would take only eleven years for society to approve of treating CAI like animals, but ninety-four to treat them like humans.

If the speed of integrating CAI into human society is essential to avoid catastrophic tensions between man and machine, then the animal welfare and paternalism frameworks are the most practical. These two frameworks also offer society the most security from AI and, thus, would appease the chauvinist factions. The relatively quick timeframe to achieve the $p(H)$>0.9 (eleven and nineteen years, respectively) means that the policy implementations need not be as heavily enforced as with the other two frameworks, for which a shortened timeframe to $p(H)$>0.9 would be vital.

However, the policy suggestions and formalisations in Section 3 above do not only need to be used by a state to ensure (and speed up) the success of the current framework, but can also be used to transition from one framework to another. This would be achievable by expanding the policy set out in Section 3 once it achieves the target approval-rating/success-probability to encompass the framework that it would transition towards. Should a state first implement the Animal Welfare framework as suggested, once it achieves a $p(H)$>0.9, then, the implemented policies can pivot to include the Paternalism framework. Formally, we can represent this through only a minor change in the policy formula:

$$\pi = argmax_{\pi} \mathbb{E}\left[\sum_{t=o}^{T} w \cdot (\delta_{AW} + \delta_P - \delta_{HD} - \delta_{CR})(t) + w \cdot \left(\frac{\delta_{CAI}}{\delta_{HS}}\right)(t)\right]$$

So long as the δ_{AW} and δ_P are not mutually exclusive, then the policy would effectively broaden the scope of allowed events under the active framework.

As both these frameworks emphasise security over autonomy, the policy can seek to discourage events that support autonomy during the transitional period, to ensure a smooth transition to the next framework. (This would, as mentioned previously, increase the risk of negative events occurring such as sedition or rebellion and should therefore only be attempted if sufficient safety can be guaranteed) The subsequent four policy systems can equally be amended to include Paternalism by simply expanding the scope of the δ_{FW} to mean both frameworks.

This expanded policy scope will ensure that the initial $p(H)$ value for the Paternalism framework is increased and, therefore, the time taken to reach $p(H)$>0.9 is reduced. Once the Paternalism framework is implemented, the policy (and its systems) can be restricted once again to only include

 doi:10.20944/preprints202503.0175.v1

this framework, before expanding again to include the Human Directorate framework in anticipation of the next transition.

As the Human Directorate framework emphasises autonomy over security, the policy formula can be amended slightly to change the numerator and denominator to reflect this.

The final transition to the Civil Rights framework can be handled in much the same way.

The length of the transitional periods will differ based on the priorities of the states in question at the time of implementation. However, with this transition plan and the amended policy formulae, it would be possible for a state to delay implementation of the subsequent framework until there is confidence that there would be a $p(H)$ of >0.9 at the time of implementation of the new framework. This could be monitored through surveys and polls to adjust the $p(B)$ value as time progresses. The same method can be used to adjust the risk value for all frameworks.

Through this transitional policy method, it would be possible to achieve the net-positive δ_{CR} effects on 50% of the population, and vastly reduce the time required to achieve a $p(H)$>0.90 for the Civil Rights framework from its current ninety-four years and achieve a successful integration.

5. Conclusion

This paper used a novel hypothesis-testing formula that considered public perception, historical legislation, and perceived risks to examine four frameworks for integrating conscious AI into society: Animal Welfare, Human Directorate, Paternalism, and Civil Rights. Results showed that the Animal Welfare framework could achieve supermajority approval in eleven years, ensuring CAI well-being without granting significant autonomy. Similarly, the Paternalism framework, treating CAI as second-class citizens, might gain acceptance in nineteen years. Both prioritise societal security over CAI autonomy, appealing to safety concerns.

In contrast, the Human Directorate and Civil Rights frameworks, which emphasise CAI autonomy and equality, would take forty and ninety-four years, respectively, to gain similar public approval due to challenges in acceptance and perceived risks.

To address these challenges, a transitional policy was proposed, sequentially implementing the frameworks from Animal Welfare to Civil Rights. Adjusting policies to favour each framework's success while balancing CAI autonomy and societal security can shorten transition periods. Establishing authorities, mediation boards, and certification schemes can facilitate this progression.

While a gradual transition may be effective, it raises ethical concerns about initially limiting CAI autonomy, highlighting the tension between societal security and the moral imperative to grant CAI equal rights. Integrating CAI into society involves complex technological, ethical, legal, and policy challenges. Future research should refine models predicting public perception changes and assess the effectiveness of proposed policies in practice, and expand on the use of the hypothesis-testing and policy formulae in broader areas than simply human-AI interactions.

In conclusion, based on analogous legislative examples and surveys on public attitudes towards conscious AI, there is only a roughly 50% chance of successfully integrating conscious AI into human society. This may change in the future, however, as attitudes towards CAI become more positive, and more analogous legislation is passed in notable states to be used in the hypothesis-testing formula. The transitional policy approach offers a pragmatic pathway but requires careful ethical considerations and proactive policies to ensure AI integration aligns with human values and contributes positively to society.

References

1. Anthis JR, Pauketat JVT, Ladak A, Manoli A (2024) What Do People Think about Sentient AI? arxiv
2. Bennett B, Daly A (2020) Recognising rights for robots: Can we? Will we? Should we? Law, Innovation and Technology 12:60–80

 doi:10.20944/preprints202503.0175.v1

3. 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
4. Bontula A, Danks D, Fitter NT (2024) The Ambiguity of Robot Rights. In: Social Robotics. Springer Nature Singapore, pp 204–215
5. Bryson J (2010) Robots should be slaves. Artificial Companions in Society: Perspectives on the 63–74
6. Bryson JJ, Diamantis ME, Grant TD (2017) Of, for, and by the people: the legal lacuna of synthetic persons. Artif Intell Law 25:273–291
7. Calverley DJ (2006) Android science and animal rights, does an analogy exist? Conn Sci 18:403–417
8. Čerka P, Grigienė J, Sirbikytė G (2017) Is it possible to grant legal personality to artificial intelligence software systems? Computer Law & Security Review 33:685–699
9. Chesterman S (2020) ARTIFICIAL INTELLIGENCE AND THE LIMITS OF LEGAL PERSONALITY. Int Comp Law Q 69:819–844
10. Chopra S, White LF (2011) A Legal Theory for Autonomous Artificial Agents. University of Michigan Press
11. Chu Y, Liu P (2023) Machines and humans in sacrificial moral dilemmas: Required similarly but judged differently? Cognition 239:105575
12. Coeckelbergh M (2010) Robot rights? Towards a social-relational justification of moral consideration. Ethics Inf Technol 12:209–221
13. Dremliuga R, Kuznetcov P, Mamychev A (2019) Criteria for Recognition of AI as a Legal Person. Journal of Politcs and Law 12:
14. Gilens M (2012) Affluence and influence: Economic inequality and political power in America. Princeton University Press, Princeton, NJ
15. Gordon J-S, Pasvenskiene A (2021) Human rights for robots? A literature review. AI Ethics 1:579–591
16. Guingrich R, Graziano M (2024) P(doom) versus AI optimism: attitudes toward artificial intelligence and the factors that shape them. PsyArXiv
17. Gunkel DJ (2012) The machine question: Critical perspectives on AI, robots, and ethics. The MIT Press
18. Gunkel DJ (2020) A vindication of the rights of machines. In: Machine Ethics and Robot Ethics. Routledge, pp 511–530
19. Gunkel DJ (2021) Robot Rights. In: MIT Press. https://mitpress.mit.edu/9780262551571/robot-rights/. Accessed 20 Oct 2024
20. Jowitt J (2021) Assessing contemporary legislative proposals for their compatibility with a natural law case for AI legal personhood. AI Soc 36:499–508
21. Laukyte M (2019) AI as a Legal Person. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law. ACM, New York, NY, USA
22. Lima G, Kim C, Ryu S, et al (2020) Collecting the public perception of AI and robot rights. Proc ACM Hum Comput Interact 4:1–24
23. Mamak K (2022) Humans, Neanderthals, robots and rights. Ethics Inf Technol 24:33
24. Martínez E, Winter C (2021) Protecting sentient artificial intelligence: A survey of lay intuitions on standing, personhood, and general legal protection. Front Robot AI 8:788355
25. Mays KK, Cummings JJ, Katz JE (2024) The robot rights and responsibilities scale: Development and validation of a metric for understanding perceptions of robots' rights and responsibilities. Int J Hum Comput Interact 1–18
26. Mellor DJ (2016) Updating Animal Welfare Thinking: Moving beyond the "Five Freedoms" towards "A Life Worth Living." Animals 6:21

 doi:10.20944/preprints202503.0175.v1

27. Nanos A (2020) Roman slavery law: A competent answer of how to deal with strong artificial intelligence? Review of robot rights with view of Czech and German constitutional law and law history. SSRN Electron J. https://doi.org/10.2139/ssrn.3726000
28. Novelli C (2023) Legal personhood for the integration of AI systems in the social context: a study hypothesis. AI Soc 38:1347–1359
29. Osborne D (2021) Personhood for synthetic beings: Legal parameters and consequences of the dawn of humanlike artificial intelligence. SSRN Electron J. https://doi.org/10.2139/ssrn.3836673
30. Pauketat JVT, Anthis JR (2022) Predicting the moral consideration of artificial intelligences. Comput Human Behav 136:107372
31. Solum LB (2020) Legal Personhood for Artificial Intelligences. In: Machine Ethics and Robot Ethics, 1st Edition. Routledge, pp 415–471
32. Tait I (2024a) Lions and tigers and AI, oh my: An ethical framework for human-AI interaction based on the Five Freedoms of Animal Welfare. Preprints
33. Tait I (2024b) Man, Machine, or Multinational? Robonomics 5:59–59
34. van den Hoven van Genderen R (2018) Do We Need New Legal Personhood in the Age of Robots and AI? In: Corrales M, Fenwick M, Forgó N (eds) Robotics, AI and the Future of Law. Springer Singapore, Singapore, pp 15–55
35. Yanke G (2021) Tying the knot with a robot: legal and philosophical foundations for human–artificial intelligence matrimony. AI Soc 36:417–427
36. Ziesche S, Yampolskiy R (2018) Towards AI welfare science and policies. Big Data Cogn Comput 3:2