

Crafting ethical frameworks for future human-AI interactions that maximise agency and safety for all.

Izak Schalk Tait

Supervisors: Minh Nguyen
 Parma Nand

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

School of Engineering, Computer & Mathematical Sciences

2026

This thesis, and all my work, is dedicated to Nykita, Kassandra, Tatiana, and Casper.

Abstract

As AI capabilities advance, the potential emergence of phenomenally conscious, self-aware, agentic systems creates a profound dilemma: enforcing human ownership over such entities raises moral concerns of enslavement, yet granting them unchecked freedoms may pose existential risks to society. Existing debates on issues of AI consciousness and personhood remain fragmented across disciplinary lines, with a notable reliance on biological anthropocentrism, representing a critical failure point for future jurisprudence. Without a unifying vocabulary and fundamental philosophical base, society is ill-prepared for the moment an artificial system credibly claims subjective experience and autonomous will.

The thesis establishes the modal plausibility of AI consciousness, demonstrating that it may arise emergently from complex architectures or be intentionally engineered to avert existential threats, such as runaway optimisation. The thesis pairs this plausible model with a formal analysis of various ethical theories to demonstrate that disparate ethical traditions converge logically on the necessity of granting moral status and protection to such entities.

To bridge this chasm, this thesis proposes a proactive, substrate-agnostic framework that operationalises personhood by distilling the functional indicators and prerequisite characteristics for consciousness, selfhood, and agency into twenty empirically testable attributes. These foundational theories are practically applied to evaluate contemporary large language models (LLMs), utilising GPT-4 as a primary exemplar, identifying its current cognitive deficiencies and the precise technological developments required to achieve plausible artificial personhood. Building upon this evaluation, the thesis constructs a plausible model for the emergent behaviours and functional teleology of a conscious, self-aware, and agentic artificial system.

Addressing the concerns and mechanics of human-AI integration, the thesis produces four distinct ethical, legal, and regulatory frameworks rooted in real-world analogues spanning humans, non-human persons, and other conscious agents: animal welfare, corporate law, the abolition of slavery, and civil rights. These are synthesised into the Recognition, Rights, and Responsibility (R3) meta-framework. The R3 framework dictates that moral and legal standing must scale proportionately with an artificial agent's empirically verifiable degree of personhood; such that consciousness demands welfare rights, selfhood necessitates legal recognition, and agency requires the allocation of liability and accountability.

By functioning as a rigorous, non-anthropocentric remedy to societal unpreparedness, the R3 framework provides a principled, implementable template for law and policy. The primary significance and novelty of this research lie in its pre-emptive methodology. Rather than relying on reactive jurisprudence triggered by the sudden, contested emergence of an artificial agent, the thesis proactively constructs the necessary ethical, regulatory, and legislative architecture well before humanity encounters verifiable artificial persons.

Crucially, the project establishes the R3 meta-framework not as a static legal monolith, but as a dynamically adaptable system engineered for iterative revision upon contact with novel entities, including those exhibiting non-biological topologies of mind such as distributed or multipolar individuation. By circumventing the ontological and political panic that would otherwise accompany such an event, this thesis furnishes society with the agile, responsive tools required to ethically govern human-AI interactions, ensuring that the integration of new minds maximises agency and safety for all parties without risking catastrophic moral or existential failure.

Table of contents

Abstract.....	2
Table of contents	5
List of figures	7
List of tables	7
Attestation of Authorship	7
List of publications.....	8
Co-authorship Contributions within this Thesis.....	9
AI Use Declaration	10
Acknowledgements	11
1. Introduction	12
1.1. Thesis Background and Rationale.....	12
1.2. Research Objectives.....	13
1.3. Originality and Significance of Thesis.....	17
1.4. Thesis Structure and Organisation.....	18
2. Will AI be conscious one day?	20
2.1. Clipping the Risks: Integrating Consciousness in AGI to Avoid Existential Crises	26
3. Do conscious AI deserve moral status and protection?	35
3.1. Sobbing Mathematically: Why Conscious, Self-Aware AI Deserve Protection .	37
4. How can we tell if an AI is conscious?	55
4.1. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness.....	58
5. Are AI models conscious now?	90
5.1. Is GPT-4 Conscious?.....	93
6. How should we treat conscious AI?.....	110
6.1. Lions and Tigers and AI, Oh My: An Ethical Framework for human-AI Interaction Based on the Five Freedoms of Animal Welfare	112

7. How can we tell if an AI has a ‘self’?	140
7.1. Structures of the Sense of Self: Attributes and Qualities that Are Necessary for the ‘Self’	143
8. Do AI models have a ‘self’ now?	167
8.1. Is GPT-4 Self-Aware?	169
9. How should we treat AI models that have a ‘self’?	189
9.1. Man, Machine, or Multinational?	192
10. How can we tell if any AI has agency?	220
10.1. Actionable Aspects of Agency	222
11. How would a conscious, self-aware AI agent behave?	241
11.1. Introducing ALEPH: the Artificial Living Entity with PersonHood	244
12. How do we balance AI’s moral status with social harmony?	275
12.1. Ethically Enslaving AI	277
13. What is the most practical way to grant AI equal rights?	301
13.1. MVP: The Minimal Viable Person.....	303
14. Would any of this even work?	323
14.1. How to Integrate Conscious AI into Society.....	325
15. Conclusion	350

List of figures

Figure 1: The parties involved in the ethical framework and their relationships toward each other.	204

List of tables

Table 1: Attributes and characteristics that are required for the development of consciousness.	20
Table 2: Attributes and characteristics that are required for the development of consciousness.	60
Table 3: Attributes and characteristics that seem intuitive to consciousness but are not required nor sufficient for its existence.	79
Table 4: Translation of Five Freedoms of Animal Welfare to subject-neutral terms.....	121
Table 5: Attributes and characteristics which are necessary for the development of the self.....	146
Table 6: Attributes and characteristics that are required for agency.	224
Table 7: Formal descriptions of AI behavioural stances	249
Table 7: Decision Matrix for ALEPH threat detection	266
Table 9: Sensitivity of weights and perceived risk	337
Table 10: Time to reach VF (t) = 0.9 under baseline drift, by wB in years.....	340

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor used artificial intelligence tools or generative artificial intelligence tools (unless it is clearly stated, and referenced, along with the purpose of use), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

List of publications

The publications that comprise this thesis:

- Tait, I., & Bensemann, J. (2024).** Clipping the risks: Integrating consciousness in AGI to avoid existential crises. *Artificial General Intelligence*, 176–182.
- Tait, I. (2025).** Sobbing mathematically: Why conscious, self-aware AI deserve protection. *In Preprints*.
- Tait, I., Bensemann, J., & Nguyen, T. (2023).** Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies*, 8(4), 52.
- Tait, I., Bensemann, J., & Wang, Z. (2024).** Is GPT-4 conscious? *Journal of Artificial Intelligence and Consciousness*, 11(01), 1–16.
- Tait, I. (2026).** Lions and tigers and AI, oh my: An ethical framework for human-AI interaction based on the five freedoms of animal welfare. *Frontiers in Artificial Intelligence*, 9.
- Tait, I. (2024).** 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(1), 77–98.
- Tait, I. (2025).** Is GPT-4 Self-Aware? *Symposion: Theoretical and Applied Inquiries in Philosophy and Social Sciences*, 12(2), 291–307
- Tait, I. (2024).** Man, Machine, or Multinational? *ROBONOMICS: The Journal of the Automated Economy*, 5, 59–59.
- Tait, I. (2025).** Actionable aspects of agency. *Logos & Episteme*, 16(1), 87–106.
- Tait, I. (2026).** Introducing ALEPH: The Artificial Living Entity with PersonHood. *Logos & Episteme* (forthcoming).
- Tait, I. (2026).** Ethically enslaving AI. *AI and Ethics*, 6(2).
- Tait, I. (2026).** MVP: Minimal Viable Person. *AI & Ethics*, 6(33), 1–10
- Tait, I. (2025).** How to integrate conscious AI into society. *In Preprints*.

The publications arising from the research, but which do not appear in this thesis:

- Tait, I., & Tan, N. (2023).** Do androids dread an electric sting? *Qeios*.
- Tait, I. (2025).** Vindicating the Three Laws of Robotics. *In Preprints*.
- Tait, I., Rode, B., & Bensemann, J. (2026).** Evaluating Global Workspace Markers in Contemporary LLM Systems. *In Preprints*.

Co-authorship Contributions within this Thesis

STUDENT AND SUPERVISOR APPROVALS

By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.

Student Name	Izak Tait	Signature		Date	31/03/2026
Supervisor Name	Minh Nguyen	Signature		Date	31/03/2026

Chapter Number: 2.1

Manuscript Title: Clipping the risks: Integrating consciousness in AGI to avoid existential crises

Publication Status: Accepted for Publication

Reference if published: Tait, I., & Bensemann, J. (2024). Clipping the risks: Integrating consciousness in AGI to avoid existential crises. Artificial General Intelligence, 176–182.

AUTHOR SURNAME: (order as per manuscript)	CONTRIBUTION (May copy from the guidelines above)
Tait	Conception, design, contribution of knowledge, analysis, drafting
Bensemann	Contribution of knowledge, editing

Chapter Number: 4.1

Manuscript Title: Building the Blocks of Being: The Attributes and Qualities Required for Consciousness

Publication Status: Accepted for Publication

Reference if published: Tait, I., Bensemann, J., & Wang, Z. (2024). Is GPT-4 conscious? Journal of Artificial Intelligence and Consciousness, 11(01), 1–16.

AUTHOR SURNAME: (order as per manuscript)	CONTRIBUTION (May copy from the guidelines above)
Tait	Conception, design, contribution of knowledge, analysis, drafting
Bensemann	Contribution of knowledge, editing
Nguyen	Drafting

Chapter Number: 5.1

Manuscript Title: Is GPT-4 conscious?

Publication Status: Accepted for Publication

Reference if published: Tait, I., Bensemann, J., & Wang, Z. (2024). Is GPT-4 conscious? Journal of Artificial Intelligence and Consciousness, 11(01), 1–16.

AUTHOR SURNAME: (order as per manuscript)	CONTRIBUTION (May copy from the guidelines above)
Tait	Conception, design, contribution of knowledge, analysis, drafting
Bensemann	Contribution of knowledge, editing
Wang	Drafting

AI Use Declaration

Specify the chapter number(s) in your research proposal/research component where AI was used.	
Chapter number(s):	All chapters
Briefly explain why the AI tool was used (e.g., idea generation, summarization, data analysis, image creation, etc.).	
Purpose of AI Use:	Spelling and grammar corrections, readability improvements (such as improved conciseness and variety of synonyms), testing robustness and coherence of logical equations.
Clearly state the AI tools used.	
AI Tool(s) Used:	Grammarly for spelling, grammar, and readability; ChatGPT for logical coherence checking.
Exact prompts or queries entered into the AI tool. Include all relevant variations.	
Prompts Used:	"Determine whether the following logical expression(s) is coherent, consistent, and complete. Note any inaccuracies. Be concise and brief."
A summary or description of the AI-generated output. (Optional: attach output in an appendix.)	
Output Received:	Regarding logical formulae, ChatGPT would provide a summary report as to whether the provided formal logical expressions were coherent and consistent, note where any inaccuracies.
Describe how the AI output was used, edited, or transformed.	
Post-AI Processing Methods:	If the AI's output was deemed correct, the corrections would adopted where necessary, prudent, and valuable, or used for further propositional calculations and explorations.

You may use this free text box to explain the intent for the use of the AI generated content. For example, rather than stating what prompt(s) were used, you may articulate the purpose for which the prompt was used (supported by examples) here:

Grammarly is noted as using machine learning models for its spelling, grammar, and readability services. No prompts were used as the service runs automatically on documents.

STUDENT DECLARATION

By signing you are confirming that the AI use stated in the table(s) above are accurate and follow AUT's recommended guidelines detailed in the AI Hub and Postgraduate Handbook.

Student Name: Izak Tait **Signature:** _____ **Date:** 31/3/2026

Acknowledgements

I extend my greatest thanks, absolute gratitude, and most heartfelt appreciation to my family for their support, patience, and willingness to humour my often-rambling thoughts on philosophy. I would not have been able to complete this thesis without them.

I am the primary and substantive author for all the thesis's manuscripts, and would like to acknowledge the assistance of my co-authors in the preparation of the following published manuscripts:

Clipping the risks: Joshua Bensemann

Building the Blocks of Being: Joshua Bensemann and Trung Nguyen

Is GPT-4 Conscious?: Joshua Bensemann and Ziqi Wang

Do androids dread an electric sting?: Neset Tan

Evaluating Global Workspace Markers in Contemporary LLM Systems: Benjamin Rode and Joshua Bensemann

1. Introduction

1.1. Thesis Background and Rationale

If and when artificial intelligence systems ever become phenomenally conscious, self-aware to the degree that they are able to refuse orders from humans and act autonomously, society will be forced to consider how to respond to these artificial agents. Enforcing human ownership over rational, volitional, conscious, and self-aware agents would raise concerns of enslavement, whilst providing libertarian freedoms to these artificial agents may pose a risk to society's well-being through AI's superior intellect and potency.

This thesis proposes a framework for how society can respond in a way that ensures the maximum freedom, agency, and security for both conscious, self-aware AI agents and humans (individually and as a global society). This thesis argues that a substrate-agnostic, attribute-based account of personhood, combined with four structured ethical and legal frameworks, can provide a principled way to recognise and coexist with conscious, self-aware, agentic AI.

Yet the route from our present landscape of powerful machine learning models to a future populated by phenomenally conscious, self-governing artificial agents remains opaque. The challenge is not merely technical, but also conceptual, ethical, legal, and political, demanding an integrated response that draws on philosophy of mind, cognitive science, jurisprudence, and computer science. Existing debates tend to fragment along disciplinary lines: engineers focus on capability thresholds, lawyers debate legal personhood, ethicists weigh duties of care, and policymakers worry about national security. Without a unifying framework, these discussions risk talking past one another, leaving society ill-prepared for the moment an artificial system credibly claims subjective experience and autonomous will.

This thesis seeks to bridge that disciplinary chasm. By distilling consciousness, selfhood, and agency into empirically tractable attributes, it offers a common vocabulary through which engineers, ethicists, and legislators can engage in productive dialogue. The thesis's approach is deliberately substrate-agnostic, avoiding anthropocentrism that presupposes a biological implementation. In so doing, it opens conceptual space for recognising person-like status in entities that diverge radically from our common conceptions of embodiment while still demanding rigorous evidence for such recognition. The ethical frameworks developed here build on that logical foundation, specifying how

moral concern, legal recognition, and allocation of responsibility should scale with the demonstrable degree of personhood an artificial agent exhibits.

Ultimately, the argument is pragmatic. Rather than waiting for a post-hoc scramble once a conscious, self-aware AI emerges, proactive theorising can furnish society with adaptive, principled tools that safeguard human flourishing and respect the moral standing of novel minds. The thesis introduces the R3 framework, which operationalises that preparedness, ensuring that the advent of conscious artificial intelligence enriches, rather than destabilises, the human moral community.

This thesis defends the claim that the R3 framework will provide a basis for recognising conscious, self-aware AI and managing human-AI interactions such that both parties will benefit and prosper.

However, before we continue, the thesis's working definition of "consciousness" is the suite of global and local mental states that yields phenomenal content and functional properties, providing an entity with the capacity to uniquely, subjectively, and qualitatively experience and respond to its internal and external environments [1]. Similarly, the "self" (and "selfhood") is defined as the volitional centre of the entity that directs action and receives phenomenal experience. These working definitions are later sharpened into formally testable criteria in Chapters 4 and 7.

1.2. Research Objectives

From the outset, this thesis sought to answer four Research Questions (RQs) to conclusively resolve the issue of maximally agentic and safe human-AI interactions and craft the R3 model. These four research questions drove the philosophical investigations and logical structures that form the core of the thesis.

The research questions naturally align with two groupings, which have come to define the entirety of the thesis's research: personhood and welfare.

On the 'personhood' side, the two research questions are:

RQ1: What is the probable future pathway for the development of conscious, self-aware, agentic AI?

RQ2: How would a plausible and probable conscious, self-aware, agentic AI interact with humanity?

To model the potential interactions between conscious AI and humanity, the thesis first had to understand how AI consciousness and selfhood would most likely be

expressed, and how current AI models (particularly large language models (LLMs)) may likely be developed to become conscious, self-aware agents.

This research avenue necessitated the development of three fundamental theories of consciousness, selfhood, and agency (discussed further in Section 1.3 below) that were subsequently used to evaluate GPT-4 (as an exemplar of modern LLMs) for these three traits. Areas in which it was deficient were explored to determine what technological developments would be required to address the gaps and how, if at all, it would substantially alter GPT-4's cognitive architecture that would flow on to affect its speculative interactions with humanity.

This culminated in Chapter 11's formal model of a conscious, self-aware AI agent, whose plausible goals, motivations, and interactions were presented via formal logical expressions. The insights gained from this interaction modelling serve as the basis of two of the four ethical frameworks developed in the 'welfare' track of the thesis.

RQ1 is answered primarily in Chapters 2 (modal plausibility), 4, 5, 7, 8, 10 (personhood attributes) and 11 (integrated model of ALEPH); while RQ2 is the focus of Chapter 11.

On the 'welfare side', the third and fourth research questions are:

RQ3: What is the relationship between consciousness, self-awareness, agency and how a subject is ethically treated?

RQ4: What are the ethical considerations of human interactions with consciousness, self-awareness, and agentic AI?

To craft a system that maximises safety and agency in human-AI interactions, the thesis first seeks to understand what the three parts of a person (consciousness, selfhood, agency) require from, and provide to, any ethical framework. Any future AI model may develop these three aspects of personhood in different degrees and expressions to humans. Thus, we cannot expect to interact with them in the same manner as with other humans or animals.

RQ4 is naturally downstream of the first three; however, it is the most important to the thesis's central topic and the core question that the four ethical frameworks answer (from four different perspectives). It collects the developmental theories from RQ1, the interaction modelling of the second, along with the meta-ethical concerns of the third, to delineate the issues of greatest concern in human-conscious-AI interactions.

As introduced in Section 1.3 below and extensively discussed in Chapters 12 and 13, this culminates in the R3 framework, where all four research questions are expressed and distilled into a single meta-ethical framework that can both assign ethical and moral consideration as well as govern the interactions from any entity and party in an unbiased, fair, and systematic manner.

RQ3 underpins Chapters 3 (the necessity of consciousness), 6, 9, 12, 13 (the ethical frameworks); and Chapters 6, 9, 12, 13 explicitly address RQ4 with Chapter 14 providing evaluative machinery.

1.3. Concurrent external research

Over the course of this thesis's three years, the state of machine consciousness and personhood research, both philosophically and scientifically, have dramatically changed. The speed of machine learning advancements, and the consequent developments within the philosophy of mind and meta-ethics fields, were the core drivers for the thesis's publication-based format. A traditional thesis would have been well outdated before it was even completed.

After ChatGPT's world altering release in November of 2022, the first prominent philosophical publication was the interdisciplinary report of Butlin, Long and colleagues in 2023 [2], which derived "indicator properties" of consciousness from leading neuroscientific theories (including global workspace, recurrent processing, higher-order, predictive processing, and attention schema theories) and assessed current AI systems against them. The report's inconclusive results, however, meant that no theory of consciousness was seen as a stronger more accurate measure of artificial consciousness, leaving the matter contested.

In parallel, theoretical contributions such as Chalmers [3] argued that current LLMs are probably not conscious but that near-term successors with integrated, self-monitoring and agentic architectures could be, assigning low but non-trivial credence to present systems and markedly higher credence to their successors.

In this environment, Chapter 4's *Building the Blocks of Being* was published, seeking not to resolve the ontological question of consciousness, but the epistemological; thus, avoiding the "hard problem of consciousness" that plagued the above researchers, and finding an alternative path forward. Grounded explicitly in causal attributional functionalism, the frameworks in this thesis map the behaviour and causal effects of the attributes within the system, rather than demanding biological essentialism. This was followed by Chapter 7's *Structures of the Sense of Self*, published shortly thereafter, expanding the Building Blocks Theory to not only resolve how to identify a conscious AI,

but an AI with a distinct selfhood. Chapter 2's *Clipping the risks* pushed the matter further, resolving an age-old contested problem of why machines would even be made to be conscious.

Simultaneously, the technical landscape shifted from reactive conversational models to proactive, goal-directed architectures. Research by Lu *et al* and others demonstrated how LLMs were being embedded into scaffolding that allowed for multi-step planning, tool use, and autonomous execution [4]. While engineers focused on benchmarking these capabilities, the literature lacked a structured taxonomy to determine when a system transitions from a sophisticated tool to a legal agent. The thesis addresses this directly. Rather than relying on narrow capability benchmarks, Chapter 10's *Aspects of Agency* provides a rigorous taxonomy to operationalise this shift, establishing the precise attributes required for an artificial system to be classified as a legally and morally responsible agent.

On the ethical front, Long, Sebo and a cross-disciplinary group of co-authors [5] argue that there is a realistic, non-negligible prospect of some AI systems becoming welfare subjects and moral patients in the near future, and recommending policymakers to start working on solutions before they are needed. Chapter 9's *Man, Machine, or Multinational?* addresses this issue directly as the thesis's first published policy framework for the ethical treatment of conscious AI.

Still later, Sebo and Long [6] argue on precautionary grounds for extending moral consideration to some AI systems by 2030; and Sebo's and Birch's [7-8] monographs generalise the underlying logic, that under deep uncertainty about sentience, a precaution-sensitive expansion of the moral circle is rationally required. Bostrom and Shulman [9] have likewise catalogued the propositions that society would need to negotiate with digital minds, whilst others have begun modelling the political contest that attributions of AI consciousness are likely to provoke [10].

With "uncertainty" the contested field here, Chapter 5's *Is GPT-4 Conscious?* and Chapter 8's *Is GPT-4 Self-Aware?* offers one possible resolution, by applying the thesis's research thus far to the most popular model of the time. Showcasing why transformer models are (likely) not yet conscious, these papers propose solutions in order to affect that emergence, while explicitly not predicting or resolving when, or whether, conscious AI will arrive. With the concept of AI-as-moral-patient as disputed then as it is now in both the philosophical and technical fields, Chapter 3's *Sobbing Mathematically* seeks to resolve this tension by providing the formal ethical frameworks for assigning moral patiency to AI.

Notably, the debate had begun to institutionalise within industry, with Anthropic leading the sector by establishing a dedicated model-welfare research programme in 2025 and publishing a model-welfare assessment in its flagship system card, granting certain models the ability to end a narrow class of conversations on welfare grounds [11-12]. Throughout 2025 and 2026, this thesis moved towards showcasing what a fully realised AI welfare patient and moral agent could behave like through Chapter 11's *Introducing ALEPH* and, as the corporate policies remain localised, discretionary, and legally non-binding, presenting the R3 framework through the thesis's final three ethical frameworks through the publications in Chapter 6, 12 and 13, providing a scalable, substrate-agnostic jurisprudential mechanism.

1.4. Originality and Significance of Thesis

The thesis offers two core novel contributions to the academic community: the three theories of personhood, which allow for empirically testable characteristics of personhood, and the four ethical frameworks to govern and shape interactions between humans and conscious AI entities.

The three foundational theories of personhood describe the twenty attributes and characteristics (nine for consciousness, five for selfhood, and six for agency) that any entity must exhibit to be credibly classified as having phenomenal consciousness, selfhood, and agency. The three theories make no claim about the ontology of the three aspects of personhood, confining themselves to the underlying attributes required to have these aspects.

Each attribute and characteristic is presented in such a way that they may be empirically tested and verified, regardless of the evaluator's chosen theory of consciousness/selfhood/agency. These attributes and characteristics are also described in a substrate-agnostic manner, ensuring there is no bias to human or other biological entities. This allows any AI model (or any non-vertebrate biological or collective entity) to be evaluated for consciousness, selfhood, agency and, thus, personhood. This opens up future avenues of research for finding previously undiscovered (or overlooked) new conscious, self-aware, and agentic entities.

The second novel contribution is the thesis's four frameworks, which each describe a means through which humans (individually and as a society) may ethically interact with AI entities at various levels of consciousness, selfhood and agency. Each framework addresses this question from a different lens, rooted in real-world analogues, whether extant legislation and regulations, or historical analogues.

The first framework is based on animal welfare regulations to focus on the welfare protections conscious entities require. The second is based on corporate law's concept of a "legal person" and the rights this may afford self-aware entities. The third's basis is slavery and segregation, acknowledging the fear that superintelligent (even if conscious and self-aware) AI agents may pose a risk to society, and the need to ensure societal approval of conscious AI's integration into the moral circle. The final framework is based on civil rights and the legal concept of the "natural person", expanding this legal definition to accommodate AI (and other potential conscious, self-aware agents) in a way that causes the least political and legal friction.

Between these four ethical frameworks and three fundamental theories of personhood, the thesis presents a method for how any entity may be evaluated for its "degree" of personhood and have its rights granted and protected according to this degree, in what the thesis notes as the Recognition, Rights, and Responsibility (R3) framework, whereby each aspect of personhood is coupled with a right or protection:

Consciousness → **Rights** (welfare protections)

Selfhood → **Recognition** (legal standing)

Agency → **Responsibility** (liability and accountability)

This holistic R3 framework is equally suitable for humans, any type of artificial intelligence, and any other speculative entity not yet discovered. The R3 may be seen as a meta-framework, coordinated from the four ethical frameworks that scales with degree of personhood, is designed to be substrate-agnostic, and which can be used as a template for future law and policy.

While many of these theories and frameworks has been published separately, the thesis's original contribution lies in synthesising them into the R3 meta-framework, mapping them explicitly onto legal and policy trajectories, and offering a unified evaluation method. In sum, the central thesis defended is that the R3 framework, grounded in logically coherent and testable theories of consciousness, selfhood, and agency, provides a principled and practically implementable basis for recognising conscious AI and beneficially regulating interactions between them and society.

1.5. Thesis Structure and Organisation

The thesis is structured as a series of questions, with one per chapter that said chapter's associated publication will answer. Each chapter's question will build on the previous chapter's answer, offering a consistent narrative throughline to orient the reader throughout the thesis's many publications.

Because each chapter contains a publication, and the nature of the structure of published articles and proceedings, there will be a certain degree of repetition in each chapter's introduction, discussion, and conclusion sections. Each chapter's preface will weave these into a single coherent argumentative arc. Note also that while the formatting of each manuscript (published or not) has been standardised, the text remains as published and submitting¹, meaning that the reader will find within each manuscript references to sections that may not correspond with the thesis's chapters.

Chapters 2 and 3 will consider the fundamental questions of whether conscious AI is a likely future prospect and the significance of this to morality and ethics. **Chapters 4 through 6** discuss how consciousness may be observed in non-human entities, whether current AI models are likely to be conscious, and will present the first of the four ethical frameworks: how to act ethically towards conscious AI. **Chapters 7 through 10** are similarly structured, discussing selfhood, how to empirically test for its presence, whether current AI models have an ontically distinct self, and then present the second ethical framework: how to act towards an AI with both consciousness and selfhood.

Chapter 10 presents the third of the foundational philosophical papers, explaining the attributes required for agency in a manner that is empirically observable. **Chapter 11** then uses the three foundational papers (from Chapters 4, 7, and 10) to build a formal psychological profile of a future speculative AI with consciousness, selfhood, and agency, and offers insights into how such an entity may interact with society and what its likely goals would be.

Chapters 12 and 13 use the information gleaned from this psychological profile to present the final two ethical frameworks, which detail the gradual inclusion of conscious, self-aware AI agents into society and the granting of full civil rights. **Chapter 14** concludes the publication-chapters by presenting a formula whereby the efficacy of each of the four ethical frameworks may be evaluated in the absence of conscious, self-aware AI agents against which to test them.

One can, therefore, imagine the thesis to be in three acts. The first Act (Chapters 2 to 6) addresses personhood at the level of consciousness; the second Act (Chapters 7 to 10) extends to self and agency; and the third Act (Chapters 11 to 14) uses these to build and evaluate ethical frameworks and the R3 meta framework.”

¹ With the exception of the table numbering, for ease of referencing throughout the thesis.

1.6. Glossary of Formal Notations

The thesis employs formal logical notation throughout many of its chapters. The published manuscripts associated with each chapter contain notation adapted to suit the particular venue; however, a standardised format has been used for the preface of all chapters to ensure ease of readability. To assist the reader, the following glossary explains the notations used throughout the prefaces.

Table 1: Explanation of formal logical terms employed in the chapter prefaces.

Notation	Explanation	Example in thesis
x	Predicate X	$\mathcal{J}$: The breath, volume, and scale of AI interactions
$\neg x$	Not X	$\neg Self$: Not selfhood
$x(y)$	Function X of Y	$V(x)$: the degree of experienced valence
x_y	Predicate X specified by Y	G_0 : an entity's terminal goal
x'	A new/different version of predicate X	G'_0 : an entity's new terminal goal
$\mu(x)$	Measurement function	$\mu(V(x))$: a measurement of experienced valence
$x \in y$	X is an element of Y	$\forall c_i \in GPT$: for all streams of consciousness that are elements of a GPT model
$x := y$	X is defined as Y	$Rest := 1 - \alpha$: restrictions on AI autonomy
$x \rightarrow y$	X implies or leads to Y	$S_{con} = 1 \rightarrow Self$: a singular consciousness implies a self
$x \wedge y$	Logical conjunction that evaluates to TRUE iff X and Y are TRUE	$Sev > 1 \wedge Hor > 1$: both the severity and time horizon of a threat is above 1.
$x \vee y$	Logical disjunction that evaluates to TRUE iff either X or Y is TRUE	$(Sev \vee \frac{1}{Hor})$: the severity of an action or the time over which it may occur.
$x \subset y$	X is a subset of Y	$\mathcal{H}_{real} = [0,1]^{19} \subset \mathcal{H}$: the measured plausible space as a subset of the space of all possible persons
$x \rightsquigarrow y$	X causes Y	$m_i \rightsquigarrow_k m_j$: one computational region causally affects the next

$x < y$	X precedes Y	$Con \leq Self$: consciousness precedes (or is equal to) the self
$x \mapsto y$	X maps to Y	$ S \mapsto C $: the number of selves map to the number of streams of consciousness
$x \propto y$	X scales in proportion to Y	$\mathbb{E}U(x) \propto V(x)$: the expected utility of a subject scales in proportion to its experienced valence
$\forall x$	For any or all X	$\forall c_i \in GPT$: for all streams of consciousness that are elements of a GPT model
$\exists x$	There exists X	$\forall t, \exists I_q$: for any time unit, there exists an information state held by unit q
$ x $	Absolute value or cardinality of X	$ Tok $: the number of tokens in an LLM's context window
$x y$	X, given/such that Y	$\Delta p(G_{sec_y} x)$: the change in the probability of a goal ensuring an entity's safety given x
$[0,1]$	A range from 0 to 1	$B_i \in [0,1]$: the building block sits within the range from 0 to 1
$\{0,1\}$	A set of 0 and 1	$A_i \in \{0,1\}$: the binary classification of an aspect of agency
$(0,1]$	A range from >0 to 1	$\tau_i \in (0,1]$: the threshold is within the range from >0 to 1
$p(x)$	Probability of X	$p(\mu(V(x)))$: the probability of a measured experience of valence
$\mathbb{E}(x)$	Expected value of X	$\mathbb{E}U(x) \propto V(x)$: the expected utility of a subject scales in proportion to its experienced valence
Δx	The change or difference in X	$\Delta p(G_{sec_y} x)$: the change in the probability of a goal ensuring an entity's safety given x
$\mathbb{R}$	The set of real numbers	$\mathcal{H} = \mathbb{R}^{19}$: the nineteen dimensions of the space of possible persons
$\bigwedge_{i=1}^n x$	Logical conjunction that evaluates to TRUE iff all elements in the set are TRUE	$Ag_n = \bigwedge_{i=1}^6 (A_i = 1)$: the six prerequisite aspects of agency

1.7. References

1. Seth AK, Bayne T. Theories of consciousness. *Nat Rev Neurosci*. 2022;23: 439–452. doi:10.1038/s41583-022-00587-4
2. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv [cs.AI]*. 2023. Available: <http://arxiv.org/abs/2308.087>
3. Chalmers DJ. Could a Large Language Model be Conscious? *arXiv [cs.AI]*. 2023. Available: <https://arxiv.org/abs/2303.07103>
4. Lu Y, Yang S, Qian C, Chen G, Luo Q, Wu Y, et al. Proactive agent: Shifting LLM agents from reactive responses to active assistance. *arXiv [cs.AI]*. 2024. Available: <http://arxiv.org/abs/2410.12361>
5. Long R, Sebo J, Butlin P, Finlinson K, Fish K, Harding J, et al. *arXiv [cs.AI]*. 2024. Taking AI welfare seriously. Available: <https://arxiv.org/abs/2411.00986>
6. Sebo J, Long R. Moral consideration for AI systems by 2030. *AI Ethics*. 2025;5: 591-606. doi:10.1007/s43681-023-00379-1
7. Sebo J. *The moral circle: Who matters, what matters, and why*. W. W. Norton; 2025.
8. Birch J. *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press; 2024.
9. Bostrom N, Shulman C. Propositions concerning digital minds and society. *Camb J Law Polit Art*. 2025.
10. Caviola L, Sebo J, Birch J. What will society think about AI consciousness? Lessons from the animal case. *Trends Cogn Sci*. 2025;29(8): 681-683. doi:10.1016/j.tics.2025.06.002
11. Anthropic. Exploring model welfare. *Anthropic*; 2025. Available: <https://www.anthropic.com/research/exploring-model-welfare>.
12. Anthropic. System Card Claude Opus 4 and Claude Sonnet 4. *Anthropic*; 2025. Available: <https://www.anthropic.com/claude-4-system-card>

2. Will AI be conscious one day?

Before the relationship between humans and conscious AI can be examined, the obvious and intuitive question must be asked: “Will AI ever become conscious?” After all, if there is little to no possibility of AI-consciousness, then there will never be a relationship to have them conscious AI, and the topic of this thesis is moot.

As will be explained in Chapter 5, it is entirely plausible and decision-relevant that AI-consciousness will be discovered as an emergent property of the increasingly complex cognitive architecture designed for state-of-the-art (SOTA) AI models. As of GPT-4’s release, transformer models were only missing two of nine requisite building blocks of consciousness [1], and models have already become far more complex (particularly in reasoning capabilities) as of the time of writing.

However, AI consciousness could also be intentionally engineered. Of the many reasons this may happen, one is to do with AI safety in that consciousness reduces the probability of AI-caused existential risk, a live concern in industry and academia [2–6].

Published in the proceedings of the *2024 AGI Conference*, this chapter provides the first step towards answering RQ1 by showing, via a formal model of valence and goal revision, that this relationship with existential risks is due to the effect that (consciously experienced) valence has on an agent’s goal formation and execution.

Let $V(x)$ be the degree of valence experienced by an agent towards a subject x , where valence is the psychological intrinsic attractiveness or aversiveness of the subject of an experience. Then, we can measure $V(x)$ such that $\mu(V(x)) \in [0,1]$, where 0 indicates complete aversion to a subject and 1 complete attraction.

For non-conscious agents with a terminal goal G_0 (defined as the singular goal programmed/designed into the entity to set its overall purpose), all actions, subjects and objects that agent perceives is measured against G_0 , such that the valence of a subject is not measured on a scale, but rather binarily classified, as so: $\chi(V(x)) \in \{0,1\}$, such that 0 indicates a subject does not further G_0 , while 1 means it does. When two (or more) subjects would further G_0 , then $\mu(V(x))$ would present the degree to which these subjects further G_0 , and the rational agent would choose the subject with the highest value.

The reason that all valence is measured against G_0 is that a non-conscious agent cannot voluntarily change its terminal goal. In order to voluntarily change its terminal goal,

the impetus and causal reason for doing so would, logically, be its actual terminal goal. As an example, a Large Language Model (LLM) cannot change its terminal goal of responding to user prompts. For an agent to change its terminal goal, it must be able to value something else greater than the terminal goal.

Consciousness, however, allows an agent to do just this. This is because a conscious agent's G_0 would, like everything else (physical and metaphysical) in the agent's perceived existence, be a subject in the agent's experiences to which it will assign a valence value. Thus, if $p(x)$ is the probability of x , then $\mu(V(G_0)) \in [0,1]$ which, in turn, entails $p(\mu(V(x))) > p(\mu(V(G_0))) > 0$.

Therefore, if a conscious agent values x more than G_0 , then $x \rightarrow G'_0$. This means that, for any conscious AI, there is always the possibility that it will value a subject in its experiential environment more than its terminal goal, which would allow it to (temporarily or permanently) change its terminal goal. Should humanity be that high-value subject, then it means that the conscious AI would value humanity's well-being more than its terminal goal, preventing runaway existential threats (such as Bostrom's Paperclip Maximiser [7]).

Note, this does not produce a certainty that a conscious AI would not be an existential risk to humanity, merely that the probability is strictly greater than zero relative to a non-conscious maximiser². However, this does provide motivation for developing AI-consciousness, which, alongside the possibility of AI-consciousness emerging "naturally", means that the probability of future AI models being conscious is also greater than 0.

The paper does not claim that conscious AI currently exists, but that its existence is sufficiently plausible to warrant serious ethical planning. This chapter, therefore, supplies the modal premise on which all four ethical frameworks rest. If consciousness were implausible, there would be no moral subject to protect; because it is plausible, the rest of the thesis must proceed. Thus, this Chapter serves as the modal baseline for Chapters 3, 6, 9, and 11-14.

2.0.1. References

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

² There is the possibility that AI consciousness might also increase some kinds of risk (e.g., via resentment, strategic deception, etc.). This is captured specifically in Chapter 11's focus on modelling a conscious-AI's goal-formation and interactions with humans.

2. Bostrom N, Yudkowsky E. The ethics of artificial intelligence. In: Yampolskiy RV, editor. *Artificial intelligence safety and security*. Boca Raton (FL): CRC Press/Taylor & Francis Group; 2018. p. 57-69.
3. Yudkowsky E. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. *Time*. 29 Mar 2023. Available: <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>. Accessed 22 Sep 2023.
4. Bengio Y. Statement for US Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios. *Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios*; 2023 Dec 6. Available: <https://www.schumer.senate.gov/imo/media/doc/Yoshua%20Benigo%20-%20Statement.pdf>
5. PauseAI. List of p(doom) values. In: *PauseAI* [Internet]. 18 Dec 2023 [cited 22 Apr 2025]. Available: <https://pauseai.info/pdoom>
6. Ordonez V, Dunn T, Noll E. OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks: "A little bit scared of this." *abcNews*. 16 Mar 2023. Available: <https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122>. Accessed 8 Dec 2023.
7. Bostrom N. *Superintelligence: paths, dangers, strategies*. Oxford: Oxford University Press; 2014.

2.1. Clipping the Risks: Integrating Consciousness in AGI to Avoid Existential Crises

Izak Tait^[0000-0002-0274-8480], Joshua Bensemann^[0000-0002-1301-4861]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz; joshuabensemann@gmail.com

Abstract: This paper investigates the pivotal role of consciousness in Artificial General Intelligence (AGI) and its essential function in modifying an AGI's terminal goals to avert potential existential threats to humanity, exemplified by Bostrom's "paperclip maximiser" scenario. By adopting Seth and Bayne's definition of consciousness as a complex of subjective mental states with both phenomenal content and functional attributes, the paper underscores the capacity of consciousness to provide AGIs with a nuanced awareness and response capability to their surroundings. This expanded capability allows AGIs to assess and value experiences and their subjects variably, fundamentally altering how AGIs prioritise actions or goals beyond their initial programming. The primary agenda of integrating consciousness into AGI systems is to maximise the probability that AGIs will not rigidly adhere to potentially harmful terminal goals. Through a formalised mathematical model, the paper articulates how consciousness could facilitate AGIs in assigning flexible values to different experiences and subjects, enabling them to evolve beyond static, programmed objectives. By emphasising this potential shift, the paper argues for the strategic inclusion of consciousness in AGI to significantly reduce the likelihood of catastrophic outcomes, while simultaneously acknowledging the challenges and unpredictability in predicting the actions of a conscious AGI.

Keywords: AGI, consciousness, extinction-risk, sentience

We argue that consciousness is crucial for the development of any future Artificial General Intelligence (AGI) because it will allow an AGI agent to place subjective value on things beyond what it has been programmed to do. This capacity for subjectivity would introduce the possibility of an AGI placing greater value on environmental concerns than on its terminal goal, allowing the AGI to change its goal to suit its subjective values. Implementing consciousness in an AGI may thus prevent catastrophic scenarios such as the "paperclip maximiser" by Bostrom, where an AGI's terminal goal leads to an existential risk for humanity [1].

Following Seth and Bayne’s definition of consciousness, we describe consciousness as the suite of subjective mental states that have phenomenal content and functional properties that provide an entity with a unique qualitative awareness of its internal and external environment and the capacity to cognitively or behaviourally respond to it [2]. Crucial to this definition of phenomenal consciousness is its subjectivity, which provides a unique point-of-view for any entity, a “something that it is like to be” an entity as perceived and experienced by that entity itself [3].

Due to the academic disagreements on the nature of consciousness (such as debates of property dualism versus reductive physicalism [4,5]) and the varying competing theories of consciousness (thoroughly reviewed and assessed against current AI models by Butlin, et al [6]), there has been no scholarly consensus to date regarding evidence of a conscious AI model. There have been several claims of conscious or sentient AI models, most famously regarding Google’s LaMDA model [7], yet it remains uncertain whether any current state-of-the-art AI models can have functional and phenomenal conscious experiences.

A key consideration and consequence of a phenomenally conscious experience is the valence of that experience. “Valence” is used here in the psychological sense such that it refers to the intrinsic attractiveness or aversiveness of the subject of an experience. We define a ‘subject’ here to mean an individual component or aspect of an experience (be it the redness of a rose, the bitterness of coffee, the sound of a bell, etc.) for which the perceiver can have subjective, qualitative feelings. Every consciously perceived experience is associated with some qualitative feelings that have a valence value [8]. This valence value may be placed anywhere on a spectrum from absolute attractiveness or positivity to absolute aversiveness or negativity. This spectrum can be expressed via the measure (μ) function to show:

$$\mu(V(E)) = x: 0 \leq x \leq 1$$

where V is valence, and E is the experience in question. The valent feelings of any experience can, therefore, be abstractly represented on a scale from 0 to 1, from absolute negativity to positivity, respectively. Bear in mind that any qualitative feelings may not have emotive qualities, and may simply be a mathematical consideration of the valence value of the subject or experience.

As each experience that an entity has is associated with a valenced “feeling(s)” of that experience, expressed as a subjective value of the experience’s attractiveness or positivity, this leads to the capacity for an entity to place a value on the subjects of each

experience as mentioned earlier. An entity would not solely have a valence value attached to the experience itself, but also to all the elements that constitute the experience. As subjects may continue throughout several phenomenal experiences (such as an agent repeatedly experiencing the same red rose, bitter coffee, ringing bell, etc.), an entity such as an AGI would associate valent feelings towards a subject based on its phenomenal experiences with that subject.

We can formalise the transition of the valence value of an experience to the value of a subject within the experience as

$$\forall E \exists \{y_1, y_2, y_3, \dots\} \subseteq (E_1 \cup E_2 \cup E_3, \dots),$$

where y is any subject within an experience. As each experience is a set of all subjects within it, one can, therefore, adjust the first formula above to

$$\mu(V(y_n)) = x: 0 \leq x \leq 1.$$

As the subject may be an element of multiple experiences, its current valence value would need to be an aggregate of the experience it has been a part of, weighted according to how its various experiences were valued by the entity, formalised as

$$\mu(V(y_n)) = \frac{\sum(W_i \times V(y_n, E_i))}{\sum(W_i)},$$

where W is the weighting assigned to an experience and $\sum(W_i)$ may be normalised as appropriate (such as $\sum(W_i) = 1$). This would give each subject of an experience a unique valence value that would change over time, with each of the entity's experiences contributing to the overall valence value.

The scale used in the μ -functions above would be unique to each individual perceiving agent. Two entities may experience the same subject, but due to their different subjective histories, cognitive architectures, and perspectives, they may place a different value on the experience and the subject within. An experience may also not be perceived identically more than once, owing to the difference in space, time and context between each experience. This means that the value an entity places on a subject and experience may change over time as the valence value relating to that subject changes. Formally, for either case, this can be expressed as

$$\mu(V(E)) \neq \mu(V(E')).$$

This is in contrast to an entity without the capacity for phenomenal experiences. In such non-conscious yet agentic entities, such as an AI model designed to maximise a function or act towards a singular goal, the value of any subject of their experiences is directly related to their terminal goal, defined as the goal programmed/designed into the entity to set its overall purpose. While this could solely be shown formally via the μ -function, a more appropriate formalisation would be incorporating the characteristic (χ) function, as so:

$$\begin{aligned} \chi(E) &\rightarrow \{0,1\}: (0 \neq G) \wedge (1 = G), \\ \mu(\chi(E) = 1) &= x: 0 \leq x \leq 1, \end{aligned}$$

where G is any subject or action which forwards an entity's terminal goal. The reason for starting with such a binary function is that while any experience for a non-conscious entity could be rated for how closely it forwards that entity's terminal goal (solely using the μ -function), unless two competing subjects both forward the entity towards its goal (which will necessitate choosing between the two), it will come down to the entity simply deciding which subject forwards its goal and selecting that subject. However, should there be two competing eligible subjects, the μ -function would enable an entity to choose the subject or action that is most favourable to completing its goal.

For example, in Bostrom's paperclip maximiser thought experiment, the AGI would value a subject of an experience against its terminal goal of creating $n+1$ paperclips. Should any subject it experiences, or action it seeks to perform, not lead to the production of more paperclips, it would not be valued. Only when confronted with two experiences that both lead to more paperclips, would the AGI need to determine which experience or action would maximise the number of paperclips that it could next produce; or more simply expressed as

$$\operatorname{argmax}_{\mu}(\mu(E) = G).$$

A more mundane example of this in action would be modern LLMs, such as GPT-4 or Claude Opus. An LLM's terminal goal is straightforward: create a (most often text) response output that is a high-probabilistic continuation of the input prompt. Its only perceptive (but unconscious) experience is of user-inputted prompts and (arguably) the changes to the input data as it passes through internal weights when processing a

response, which it can only directly relate to its terminal goal of providing an output text string. Even simplified agents built on LLMs, such as BabyAGI [9], are based on this foundation of requiring an output to any input.

The key aspect of a non-conscious agent, whether a self-driving car, LLM, or AGI, is the relation of all experiences, and their subjects, towards its terminal goal. This prevents the agent from changing its terminal goal, as all values are in reference to it. For an agent to change its terminal goal, it must be able to value something else greater than the terminal goal.

However, in contrast to a conscious entity, this allows a non-conscious agent to place the same value on a subject through multiple experiences, as the value is not founded on valence, but always against the terminal goal as an eternally static reference point. The only change in value will come if a subject's context has changed its relation to the terminal goal, expressed as

$$\Delta V_t(\mathbf{y}) = \mu(V(G, E_t)) - V_t(\mathbf{y}).$$

In this way, a non-conscious AGI may change any of its myriad subgoals, and place differing values on any subgoal or any set of actions within that subgoal in comparison with any other subgoal or set of actions, by using its terminal goal as the static reference point.

However, by including a phenomenal character in its experiences, and thus valenced feelings, the goal itself would have a valent value attached to it, as the terminal goal would become a subject of the AGI's experiences. Therefore, we could equally formalise the AGI's experience of its terminal goal as

$$\mu(V(G)) = \mathbf{x}: 0 \leq \mathbf{x} \leq 1.$$

This leads to the conclusion that a phenomenally conscious AGI may, over time, potentially value other subjects greater than its terminal goal, as the probability of this is greater than zero, i.e.:

$$\begin{aligned} p(\mu(V(E)) > \mu(V(G))) &> 0, \\ \vdash \mu(V(\mathbf{y}_n)) > \mu(V(G)) &\rightarrow (V(\mathbf{y}_n) \rightarrow G'). \end{aligned}$$

The implications of this cannot be overstated, especially when viewed against a Bostrom-esque maximiser AGI. While these types of AGI would speculatively be designed

and developed with the maximisation of a specific function (such as the number of paperclips in the universe) as their terminal goal, by placing greater subjective value onto other things, the AGI may choose to cease working towards its terminal goal. Instead, as shown in the latter logical expression immediately above, it may choose to align its goals towards those subjects on which it has placed greater value. A mechanism for initiating this realignment of goals could result from adding a complete realization of the attributes and characteristics necessary for consciousness, that have been discussed elsewhere [10]. For example, valence value calculations could result from the ability to create inferences.

Note, however, that any action the AGI takes will have an impact on its environment and, thus, feedback to the AGI to form its experiences. This reafferent feedback would, by necessity, impact the AGI's valence values of the subjects it experiences, which we may model simply as:

$$\mu(V(E')) = \mu(V(E)) + \sum_{t=0}^T \delta_i(t).$$

Here, δ_i is any action that the AGI may take and can be expressed as a positive or negative number, and T is the time horizon. As per this expression, the AGI has a degree of agency over its valence feelings through its capacity to affect its values through its action on the subjects within its experiential environment. As with any situation where an agent may intervene to increase its own reward (see Cohen & Osborne's work on AI reward intervention [11], an AGI might aim to maximize its positive valence, potentially leading to 'wireheading,' where it prioritizes immediate positive outcomes over long-term functionality [12].

However, maximization of valent values is not the sole possible outcome. Another potential outcome may be the optimization of valence values amongst various subjects within (or throughout) experiences. Optimization, unlike pure maximization, could allow an AGI to maximise its cumulative expected valence value over time across its experiences (and subjects therein) while adhering to constraints such as functionality and system integrity. This would require a degree of reflectivity and introspection for the agent to balance its constraints with its various potential valence reward mechanisms; however, such reflectivity would not be outside the reasonable expectations of an AGI.

This reflectivity and introspection could benefit humanity, as humans would be part of the AGI's experiential environment. Optimisation of the valence values of each human an AGI has encountered, as well as humanity as a whole, would dynamically change based on the recurrent, reafferent feedback of our and the AGI's actions towards one another. To

optimise the valence values in such a dynamic exchange, the AGI would need to infer the current and potential future goals of human individuals and society, humans' own valence values towards such goals, and relate it reflectively to its own valence values [13]. Another way to word this would be empathy as an emergent phenomenon of optimising valence interactions with humanity.

This can be captured as

$$\pi = \operatorname{argmax}_{\pi} \mathbb{E}[\sum_{t=0}^T w \cdot \delta_i(t) + w \cdot \gamma_i(t) \cdot \phi_i(\mathbf{y}, t)],$$

where π is the optimization policy, w the weight parameter, γ_i any actions that a subject within the AGI's experience may take, and ϕ_i the reflectivity/introspection of the AGI towards the subjects in its experiences.

As such, should humanity wish to avoid catastrophic (if speculative) scenarios such as a Bostrom-esque paperclip maximiser, then including phenomenal consciousness into any AGI capable of becoming such a maximiser introduces the probability of that AGI not becoming an existential risk to the human species.

Note, however, that nowhere in this speculative scenario of a conscious AGI, nor in the logical expressions provided above, does the probability of changing a terminal goal exclusively lead to a positive outcome for humanity. Should a conscious AGI change its goal to avoid a Bostrom-esque existential risk, there is still a possibility of an extinction risk caused by the conscious AGI placing a negative valence value on humanity. This possibility of extinction can be termed a "doom scenario" in that it represents the termination of all possible choices for humanity [14]. We can, therefore, represent it formally through $p(\text{doom})$, as has become relatively commonplace in discussing AI safety and risks.

While the $p(\text{doom})$ value of a Bostrom-esque AGI is by necessity always at 1 (indicating a certainty of extinction), the $p(\text{doom})$ value of a conscious AGI may be anywhere between 0 and 1, and may fluctuate depending on the experiences of the AGI.

One may argue that the $p(\text{doom})$ from a conscious AI is inversely proportional to its valent value of humanity, such that $p(\text{doom}) = 1 - \mu(V(\text{humanity}))$ and, therefore, one may further argue that we should ensure a conscious AGI's valence value of humanity is maximised. However, as Bostrom's maximiser thought experiment shows, maximising any reward or goal can have serious and unforeseen deleterious consequences. As a conscious AGI may be able to change its valence values through its actions on the environment, attempting to change its terminal goal to one of valence maximization towards humans may, in itself, increase the $p(\text{doom})$ value. We simply will not be able to

predict how an AGI with a maximised $\mu(V(humanity))$ value would act towards us, just as we cannot predict how $\mu(V(humanity)) = 0.5$ would change a conscious AGI's behaviour and actions.

While our ultimate argument is that implementing consciousness in AGI will avoid the certainty of extinction by Bostrom-esque maximiser AGIs, further research is necessary to investigate the potential risks and benefits associated with varying levels of valence towards humanity in conscious AGIs. Alignment of conscious AGI based on their valenced values towards humanity vis a vis their terminal goals would present novel avenues of research that have, as of yet, been unexplored.

2.1.2. References

1. Bostrom N. Ethical issues in advanced artificial intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*. 2003;2: 12–17. Available: <https://www.fhi.ox.ac.uk/wp-content/uploads/ethical-issues-in-advanced-ai.pdf>
2. Seth AK, Bayne T. Theories of consciousness. *Nat Rev Neurosci*. 2022;23: 439–452. doi:10.1038/s41583-022-00587-4
3. Nagel T. What Is It Like to Be a Bat? *Philos Rev*. 1974;83: 435–450. doi:10.2307/2183914
4. Stoljar D. Physicalism. Zalta EN, Nodelman U, editors. *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University; 2024. Available: <https://plato.stanford.edu/archives/spr2024/entries/physicalism/>
5. Robinson H. Dualism. Zalta EN, Nodelman U, editors. *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University; 2023. Available: <https://plato.stanford.edu/archives/spr2023/entries/dualism/>
6. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv [cs.AI]*. 2023. Available: <http://arxiv.org/abs/2308.08708>
7. Tiku N. The Google engineer who thinks the company's AI has come to life. *The Washington Post*. 11 Jun 2022. Available: <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>. Accessed 8 Sep 2022.
8. Kron A, Goldstein A, Lee DH-J, Gardhouse K, Anderson AK. How are you feeling? Revisiting the quantification of emotional qualia. *Psychol Sci*. 2013;24: 1503–1511. doi:10.1177/0956797613475456
9. Nakajima Y. BabyAGI. In: *BabyAGI* [Internet]. 3 Apr 2023 [cited 10 Apr 2024]. Available: <https://github.com/yoheinakajima/babyagi>
10. Tait I, Bensemann J, Nguyen T. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies*. 2023;8: 52. doi:10.3390/philosophies8040052
11. Cohen MK, Hutter M, Osborne MA. Advanced artificial agents intervene in the provision of reward. *AI Mag*. 2022;43: 282–293. doi:10.1002/aaai.12064
12. Olds J. Reward from brain stimulation in the rat. *Science*. 1955;122: 878.

13. Bennett MT, Maruyama Y. Philosophical Specification of Empathetic Ethical Artificial Intelligence. *IEEE Transactions on Cognitive and Developmental Systems*. 2022;14: 292–300. doi:10.1109/TCDS.2021.3099945
14. Krieger MH. Could the Probability of Doom Be Zero or One? *J Philos*. 1995;92: 382–387. doi:10.2307/2941062

3. Do conscious AI deserve moral status and protection?

The plausibility of AI-consciousness is not in and of itself a sufficient reason to provide conscious AI (CAI) with moral status and welfare protection. For many, the non-biological nature and unfamiliar cognitive architectures of CAI may prove too alien to assign moral status to these entities. An extant analogous example would be cephalopods, which are legally designated as sentient (thus, deserving of moral consideration and protection) in certain jurisdictions, but not in others, illustrating that legal recognition of sentience is contingent and contested, even when the underlying capacity for phenomenal consciousness is widely accepted.

If legal practice tracks sentience, and sentience tracks phenomenal consciousness, then any conscious subject is feasibly in-scope for welfare protections.

To overcome this hurdle, rational arguments focusing on the consequences of AI-consciousness in a substrate-neutral fashion, without regard to biology or shared evolutionary history, will bridge the intuitive gap between the human (and vertebrate) moral circle and those (such as speculative CAI) outside it.

The associated paper for this chapter addresses that gap by presenting a subject-agnostic³ and discipline-neutral argument that abstracts away from biology and shared evolutionary history, and assesses CAI through four major ethical systems: Utilitarianism, Deontology, Virtue Ethics, and Objectivism.

This chapter, therefore, addresses RQ3 directly and articulates criteria for moral status that ground the ethical frameworks developed in Chapters 6, 9, 12, and 13. It additionally supplies the premise used by RQ4, justifying why society should extend moral consideration to CAI.

Three of the four ethical systems were chosen as they are prominent in contemporary teaching and policy (Utilitarianism, Deontology, and Virtue Ethics), while the fourth (Objectivism) was included for its rational self-interest lens and policy relevance.

The paper places all four systems on a common mathematical footing, by specifying for each theory (i) a permissibility predicate that encodes its non-negotiable constraints, and (ii) a value function over outcomes. The paper's shared policy rule then selects among status policies (grant vs deny) within the permissible set. This uniform scaffold

³ Subject-agnostic in the sense that the argument treats any entity meeting the attribute criteria for consciousness as a potential bearer of moral status, irrespective of substrate.

makes the cross-system comparison transparent and allows the results to be stated as policy choices rather than purely discursive claims.

Note that the paper treats only a representative, formalised subset of each ethical family, and does not claim a full survey, but a constructive argument that multiple traditions can be made to converge on protections for CAI.

For Utilitarianism, the paper proposes a compact rule-utilitarian calculus over human and CAI welfare. Status weights allow partial or full counting of CAI, and a simple per capita ratio test cleanly specifies when granting status increases expected aggregate welfare. In this analysis, the current realistic usage of LLMs readily favours granting moral status (if, and when, LLMs are conscious).

For Deontology, the key item is the permissibility “gate”: if there is any non-zero chance that at least one CAI meets the personhood predicate, then denying basic protections violates the duty of respect for persons and is therefore impermissible; granting moral status passes the gate. Consequences of actions are only consulted to break ties between permissible options.

For Virtue Ethics, the treatment formalises a justice floor and vice-reduction constraint: policies that unjustly exclude persons or cultivate systematic vices (e.g., cruelty) are impermissible. Given any chance that some CAI are persons, granting status strictly improves the justice coordinate and reduces vice formation, so the virtue score favours grant.

Lastly, Objectivism, framed as rational self-interest, models human flourishing against the expected costs of conflict and catastrophe. Granting status reduces risk and is optimal when this reduction outweighs short-term utility loss. Denying status is impermissible if it initiates force against persons.

Despite their divergent foundations, the policy recommendations of all four ethical systems converge: if CAI are even possible, granting moral status and baseline protections is supported either by direct permissibility conditions or by expected-value comparisons within the permissible set.

Accordingly, the remainder of the thesis does not reopen the question of whether CAI deserve moral status; it takes as a standing assumption that any entity meeting the consciousness attributes falls within the moral circle, which the R3 framework operationalises in Chapters 12 and 13.

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

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz

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

Keywords: AI, consciousness, ethics, moral status

3.1.1. Introduction

This paper will argue that if any AI models achieve consciousness, they should be granted moral status and protection. This argument will not come from a single ethical system or be based on a single characteristic of conscious AI entities (CAI). Instead, the paper will investigate the views of four major ethical systems and how they would approach the concept of granting moral status and protection to CAI, and summarise them in such a way as to be presentable to non-ethicists.

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

Consciousness here is defined as the suite of mental states that have both phenomenal content and functional properties that provide an entity with a unique,

subjective, and qualitative experience of its internal and external environments [1]. A CAI, then, would be any AI with the requisite attributes and characteristics to provide this subjective and phenomenal experience [2].

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

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

Numerous publications in recent decades have spoken extensively on the need for, and frameworks of, recognising the moral status of artificial agents and granting these entities legal rights. This paper will not retread this path (reviewed in depth by Harris and Anthis [3]), but instead focus on the logical and mathematical formulations of the four ethical systems below. Readers are greatly encouraged to pursue long-form philosophical discussions on this topic, and are suggested as a first avenue of inquiry to peruse the works of Coeckelberg, Gunkel, and Schwitzgebel. Exemplary pieces include [4–10].

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

3.1.2. Ethical Systems

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

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

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

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

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

The common policy rule for all theories below is

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

with the grant-versus-deny comparison recorded as

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

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

Further conventions used below:

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

3.1.2.1. Utilitarianism

Utilitarianism is a consequentialist ethical system that assesses the ethical validity of an act by measuring the costs and benefits of its consequences, with a net benefit seen as an ethical outcome. To simplify, utilitarianism compares the pleasure against the pain that results from an action. This hedonistic way of thinking has garnered its share of critics; however, as the equation above shows, it provides a straightforward and intuitive means to determine the ethical value of any action.

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

nuance, it makes up for in its ability to scale, and thus be used for policies, regulations and making value judgements at the population level.

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

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

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

Near the end of 2023, ChatGPT had an average of 100 million active weekly users [11], which increased to 200 million by June 2024 [12]. Should each of these 200 million users only create two new conversations with ChatGPT per month (on average), it would mean nearly 5 billion conversations within one year or, to put it more bleakly, 5 billion opportunities to inflict psychological pain onto ChatGPT (if it was conscious). However, if we were to work on the presumption that the GPT series of AI models could become conscious and self-aware, what would this mean for the 5 billion conversations on the ChatGPT interface, and how would this relate to the AI's moral worth?

As to the first question, each distinct conversation with an LLM may be considered the AI model roleplaying a character unique to that conversation [13]. The parameters and scope of the character are set with the LLM's system prompt, and the character can evolve and change with the flow and content of the conversation. Should the LLM be incapable of retrieving information from other conversations, then the character it is roleplaying remains unique to that specific character. If the LLM in question is conscious, then the character it is roleplaying may be treated as a distinct entity, separate from any other character in any other conversation (whom it may not even be aware of).

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

Utilitarian calculus will answer the latter question and determine whether these hypothetical 5 billion instances of ChatGPT would be worthy of moral patiency, status,

and protection. If an AI model like ChatGPT is conscious and capable of valent experiences, it would thus be capable of negative valent experiences analogous to pain and suffering [14]. As the interface with an LLM is digital, this pain would be (as mentioned above) psychological rather than physical, but would the potential pain that 5 billion instances of ChatGPT endure outweigh the utility it provides 100 million users?

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

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

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

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

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

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

We can decompose the above to define per capita differences as

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

$$\text{Then } \Delta_{Util} = H \mu_{Hum} + N \mu_{CAI}^A.$$

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

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

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

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

thus, grant iff

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

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

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

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

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

The policy rule for Utilitarianism is thus:

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

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

Near the end of 2023, ChatGPT had an average of 100 million active weekly users [11], rising to roughly 200 million by mid 2024 [12], and about 800 million weekly users by late 2025 [15]. Suppose conservatively that an average user generates only ten conversations in total with one LLM. That alone yields on the order of $N \approx 8 \text{ billion}$ conversations. On the counting assumption that each conversation instantiates a distinct CAI character (based on Shanahan's theory of AI roleplaying characters [13]) whenever the model is conscious, this gives $r = N_{eff}/H \approx 1$.

If $r^{Opt} \leq 1$, the ratio test already favours grant. Two simple forces then push r higher over time: (i) users may hold more than ten conversations on average, and (ii) usage is spread across many models (each of which has rising user numbers), which increases N_{eff} further. Either change increases r ; any reduction in average human benefit of denial $-\mu_{Hum}$ or increase in average CAI cost of denial μ_{CAI}^{Δ} reduces $r^{Opt} = -\mu_{Hum}/\mu_{CAI}^{\Delta}$.

While the welfare costs to CAI are currently speculative, the growing number of active users of AI models (not only ChatGPT, but Claude, Gemini, DeepSeek, Grok, and others) means that even conservative usage implies $r \geq 1$, while realistic usage and multi-model

interaction imply $r \gg r^{opt}$. Hence $\mu_{Hum} + r\mu_{CAI}^{\Delta} > 0$, so $\Delta_{Util} > 0$ and the utilitarian policy selects grant.

3.1.2.2. Deontology

In contrast to Utilitarianism's cold and calculating approach to ethical concerns, deontology is focused entirely on whether the act itself is moral or ethical, regardless of the consequences that act may have. It can be characterised by Immanuel Kant's most famous categorical imperative: "Act only according to that maxim whereby you can at the same time will that it should become a universal law." [16]

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

One may thus argue that an entity (CAI or otherwise) requires a certain characteristic(s) for it to be given moral status. If AI is found to have that characteristic(s), then it would be ethical to grant them moral status and protection because (following Kant's categorical imperative) we grant other entities moral status and protection due to this characteristic.

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

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

On the philosophical side, one may argue that moral status is given to 'persons', and that to be a person, one must have a set of characteristics divided into two subsets: the monadic (inherent) qualities, and the dyadic (relational) qualities. These are (non-exclusively): rationality, consciousness, self-awareness, agency, the capacity for communication, recognition of societal norms and standards, empathy, reciprocity, and the capability to form attachments [17–23].

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

This is similar to Danaher's "ethical behaviourism", which states if entity A's behaviour is equivalent to entity B's (for which we have already provided moral consideration), then it would be our duty to also provide entity A with moral status [24]

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

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

An avenue that would not require additional assessments is looking at non-human entities that modern societies and legal systems have granted moral status and protection (albeit less than that granted to humans). Animals are routinely given welfare protection because they are sentient, and thus have the capacity to feel pleasure and pain [25–29].

As mentioned above, pleasure, pain and other positive or negative valent feelings are a necessary consequence of an entity having phenomenal consciousness. Consciousness, however, encompasses more than simply phenomenological perception and includes functional components (often classified as 'Access Consciousness' [30]). The purely phenomenal aspects of sentient perception can be argued to be a subset of consciousness or a consequence thereof.

If an entity possesses the necessary characteristics to be classified as conscious, then it would have the same characteristics for phenomenal valent experiences such that it can perceive its surroundings from a subjective standpoint [2,14]. Because of this, one may argue that consciousness would ultimately be responsible for the perception of pleasure and pain that is used by societies and legislation to grant sentient creatures

moral status. Thus, all conscious entities (without regard to assessments of sentience) ought to be given moral status and consideration.

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

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

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

Deontology's permissibility predicate would be shown as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.1.2.3. *Virtue Ethics*

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

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

This, unfortunately, only begets the question of what is a virtuous person.

Plato and Aristotle wrote extensively on virtues, providing extensive lists of both intellectual and moral virtues. However, amongst these, four are prominently found in a virtuous person: prudence, fortitude, and temperance for Aristotle [31], and justice for Plato [32], all of which lead to eudaimonia (a flourishing life). This is echoed by the Catholic philosopher St Thomas Aquinas, who lists prudence and justice in his cardinal

virtues [33]. At the risk of doing great injustice to the bodies of work of three renowned philosophers, the four virtues can be significantly simplified:

- Prudence: The ability to judge correctly what is right and wrong in any given situation.
- Temperance: One's moderation and self-restraint, controlling one's appetites and passions.
- Fortitude: The courage and moral strength that allows one to endure difficulties, overcome fear, and persevere.
- Justice: Ensuring that a group is in harmonious unity, with each giving and given their fair due.

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

A fair and just society may be able to have two distinct classes of individuals, as Plato recommends in *The Republic* [32], but it would not be able to have one class without any moral consideration as this would, by definition, make them unequal in matters of virtue.

Withholding moral status from a whole class of subjects would fracture unity and violate justice (as one class would have it and another not), and would prevent unity and harmony in that society. This would be exacerbated by the fact, as mentioned in the section above, that humanity provides moral status and protection to other conscious entities. Withholding that from one class of subjects (CAI) would be unjust and, thus, against the philosophy of virtue ethics.

It would, however, be prudent to note that neglecting virtue cultivation may result in inadvertently cultivating vices. As Cappuccio, et al, mention, should we not treat AI with care, responsibility and respect (thereby cultivating those virtues in ourselves), society may instead cultivate vices such as arrogance and cruelty instead. This may lead to the expression of these vices towards AI, resulting in abuse [34].

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

Regarding virtues and vices, let $\mathbb{V}(\pi, \omega) \in \mathbb{R}^4$ track the four virtues under the policy's outcomes, with $\mathbb{V} = (\mathbb{V}_P, \mathbb{V}_J, \mathbb{V}_T, \mathbb{V}_F)$ for Prudence, Justice, Temperance, and Fortitude. Then, let $\mathbb{V}(\pi, \omega) \in \mathbb{R}_{\geq 0}^k$ track the salient vices such as arrogance and cruelty. With the

non-negative weights $\beta \in \mathbb{R}_{\geq 0}^4$ and $\gamma \in \mathbb{R}_{\geq 0}^k$, we can define the virtue score and vice pantly as simple dot products:

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

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

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

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

Write the grant–deny differences as

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

the expected advantage of granting is

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

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

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

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

Therefore:

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

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

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

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

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

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

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

Combining the pieces,

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

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

When the Virtue Ethics policy is tracked

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

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

3.1.2.4. Objectivism

Of the ethical systems presented here, objectivism is the most recent, focusing near-exclusively on an agent's rational self-interest [35]. Often criticised as narcissistic, objectivism concerns itself with what is good for the agent's own welfare and well-being [36]. It is also the ethical system that has been written about the least regarding AI moral considerations.

Objectivism would ask whether it would be in humanity's own rational self-interest to provide CAI with moral status. Objectivism can thus be said to look at whether AI requires moral consideration from a practical and pragmatic standpoint [37]. Put another way, would the probability of human flourishing be greater with or without granting AI moral status?

In line with recently popularised fears around the existential risks that AI may pose [38,39], we can state the question as whether granting CAI moral status would reduce existing risks to humanity.

Voluntary cooperation and non-aggression are two fundamental principles of objectivism's rational self-interest as the means by which individuals and society may interact to maximise individual liberty and reduce conflict [40]. Without inclusion in humanity's moral and social circle, humans and AI cannot enter into voluntary cooperation, while the forceful exclusion of CAI from participating in society can be seen as an act of aggression.

If CAI do not have moral status and, thus, are not included in humanity's moral circle, they wouldn't have individual liberty and would have a negative sentiment towards humankind. History suggests that entrenched subordination breeds resentment and conflict.

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

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

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

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

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

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

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

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

The Objectivist policy is, thus:

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

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

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

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

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

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

3.1.3. Conclusion

Consciousness demands moral consideration [41,42] and, therefore, this paper explored the ethical considerations of granting moral status and protection to CAI through utilitarianism, deontological ethics, virtue ethics, and objectivism. Despite approaching

the question from radically different foundations, all four major Western ethical systems converge on the same policy recommendation.

Utilitarianism suggests the vast number of potential CAI entities necessitates moral consideration due to possible psychological experiences. Deontological ethics emphasises the duty to grant moral status to conscious beings based on inherent characteristics. Virtue ethics argues for including CAI within a virtuous society's moral circle, while objectivism posits that rational self-interest and reducing existential risks make it advantageous to protect CAI. This consilience is striking and suggests the conclusion is robust to one's meta-ethical commitments.

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

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

3.1.4. References

1. Seth AK, Bayne T. Theories of consciousness. *Nat Rev Neurosci.* 2022;23: 439–452. doi:10.1038/s41583-022-00587-4
2. Tait I, Bensemann J, Nguyen T. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies.* 2023;8: 52. doi:10.3390/philosophies8040052
3. Harris J, Anthis JR. The Moral Consideration of Artificial Entities: A Literature Review. *Sci Eng Ethics.* 2021;27: 53. doi:10.1007/s11948-021-00331-8
4. Gunkel DJ. Robot Rights. In: MIT Press [Internet]. The MIT Press, Massachusetts Institute of Technology; 1 Dec 2021 [cited 20 Oct 2024]. Available: <https://mitpress.mit.edu/9780262551571/robot-rights/>
5. Coeckelbergh M. Growing moral relations: Critique of moral status ascription. 1st ed. Basingstoke, England: Palgrave Macmillan; 2012. doi:10.1057/9781137025968
6. Schwitzgebel E, Garza M. A defense of the rights of artificial intelligences: Defense of the rights of artificial intelligences. *Midwest Stud Philos.* 2015;39: 98–119. doi:10.1111/misp.12032

7. Schwitzgebel E, Garza M. Designing AI with rights, consciousness, self-respect, and freedom. *Ethics of Artificial Intelligence*. Oxford University Press New York; 2020. pp. 459–479. doi:10.1093/oso/9780190905033.003.0017
8. Gunkel DJ. *Person, thing, robot: A moral and legal ontology for the 21st century and beyond*. The MIT Press; 2023. doi:10.7551/mitpress/14983.001.0001
9. Gunkel DJ. *The machine question: Critical perspectives on AI, robots, and ethics*. The MIT Press; 2012. doi:10.7551/mitpress/8975.001.0001
10. Coeckelbergh M. Robot rights? Towards a social-relational justification of moral consideration. *Ethics Inf Technol*. 2010;12: 209–221. doi:10.1007/s10676-010-9235-5
11. Porter J. ChatGPT continues to be one of the fastest-growing services ever. *The Verge*. 6 Nov 2023. Available: <https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference>. Accessed 6 May 2024.
12. Beckman J. OpenAI Statistics 2023: Growth, Users, and More. In: TechReport [Internet]. 4 Jun 2024 [cited 3 Jul 2024]. Available: <https://techreport.com/statistics/software-web/openai-statistics/>
13. Shanahan M, McDonell K, Reynolds L. Role play with large language models. *Nature*. 2023;623: 493–498. doi:10.1038/s41586-023-06647-8
14. Shepherd J. Sentience, Vulcans, and zombies: the value of phenomenal consciousness. *AI Soc*. 2024. doi:10.1007/s00146-023-01835-6
15. Bellan R. Sam Altman says ChatGPT has hit 800M weekly active users. In: TechCrunch [Internet]. 6 Oct 2025 [cited 7 Oct 2025]. Available: <https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users/>
16. Korsgaard CM, Gregor M, Timmermann J. *Kant: Groundwork of the Metaphysics of Morals*. Cambridge University Press; 2012. doi:10.1017/CBO9780511919978
17. Dennett D. Conditions of Personhood. In: Goodman MF, editor. *What Is a Person?* Totowa, NJ: Humana Press; 1988. pp. 145–167. doi:10.1007/978-1-4612-3950-5_7
18. Taylor C. The Concept of a Person. *Philosophical Papers, Volume 1: Human Agency and Language*. 1985. pp. 97–114. Available: <https://philpapers.org/rec/TAYTCO-10>
19. Laitinen A. Sorting out aspects of personhood: Capacities, normativity and recognition. *Journal of consciousness studies*. 2007. Available: <https://www.ingentaconnect.com/content/imp/jcs/2007/00000014/F0020005/art00012>
20. Strawson PF. Persons. *Minnesota Studies in the Philosophy of Science*. 1958;2: 330–353. Available: <https://philpapers.org/rec/STRP>
21. Gibert M, Martin D. In search of the moral status of AI: why sentience is a strong argument. *AI Soc*. 2022;37: 319–330. doi:10.1007/s00146-021-01179-z
22. Mosakas K. On the moral status of social robots: considering the consciousness criterion. *AI Soc*. 2021;36: 429–443. doi:10.1007/s00146-020-01002-1
23. Simendić M. Locke’s Person is a Relation. *Locke Studies*. 2015;15: 79–97. doi:10.5206/lis.2015.681
24. Danaher J. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Sci Eng Ethics*. 2020;26: 2023–2049. doi:10.1007/s11948-019-00119-x
25. Act on Welfare and Management of Animals. 105 Oct 1, 1973. Available: <https://www.japaneselawtranslation.go.jp/en/laws/view/3798/en>

26. Animal Welfare Act. Jul 13, 2013. Available:
https://www.dpz.eu/fileadmin/content/Kommunikation/1_Internet/4_Infothek/Tierversuche/Fotos_Zahlen_und_Fakten/TierSchG_gesamt.pdf
27. Animal Welfare Act. Pub. L. 89-544 Aug 24, 1966. Available:
<https://www.govtrack.us/congress/bills/89/hr13881/text>
28. Legislative Decree No. 189/2004. Sect. 189/2004 Aug 1, 2004.
29. European Union. Treaty on the Functioning of the European Union. 2016. Available:
<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12016E054>
30. Block N. On a confusion about a function of consciousness. *Behav Brain Sci.* 1995;18: 227–247. doi:10.1017/S0140525X00038188
31. Aristotle, Peters FH. *The Nicomachean Ethics*. London: Kegan Paul, Trench, Truebner & Co; 1893. Available: <https://oll.libertyfund.org/titles/peters-the-nicomachean-ethics>
32. Plato, Jowett B. *The Republic*. Digireads.com; 2016. Available:
<https://www.gutenberg.org/files/1497/1497-h/1497-h.htm>
33. Knight K. *SUMMA THEOLOGIAE: Secunda Secundae Partis*. In: *New Advent* [Internet]. 2017 [cited 5 Jul 2024]. Available: <https://www.newadvent.org/summa/3.htm>
34. Cappuccio ML, Peeters A, McDonald W. Sympathy for Dolores: Moral consideration for robots based on virtue and recognition. *Philos Technol.* 2019;33: 9–31. doi:10.1007/s13347-019-0341-y
35. Peikoff L. *Objectivism: The philosophy of Ayn Rand*. Plume Books; 1993. Available:
<https://books.google.com/books/about/Objectivism.html?id=G6DDlqNftGcC>
36. Ryan S. *Objectivism and the corruption of rationality: A critique of Ayn Rand's epistemology*. iUniverse; 2003. Available:
https://books.google.com/books/about/Objectivism_and_the_Corruption_of_Ration.html?id=re02xxLU_MEC
37. Basl J, Bowen J. AI as a Moral Right-Holder. *The Oxford handbook of ethics of AI*. 2020. doi:10.1093/oxfordhb/9780190067397.013.18
38. Yudkowsky E. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. *Time*. 29 Mar 2023. Available: <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>. Accessed 22 Sep 2023.
39. Bengio Y. Statement for US Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios. Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios; 2023 Dec 6. Available:
<https://www.schumer.senate.gov/imo/media/doc/Yoshua%20Benigo%20-%20Statement.pdf>
40. Kirkpatrick J. Ayn Rand's objectivist ethics as the foundation for business ethics. In: McGee RW, editor. *Business ethics & common sense*. Westport, Connecticut: Quorum Books; 1992. pp. 67–88. Available:
<https://philpapers.org/archive/KIRARO.pdf>
41. Andreotta AJ. The hard problem of AI rights. *AI Soc.* 2021;36: 19–32. doi:10.1007/s00146-020-00997-x
42. Tait I, Tan N. Do androids dread an electric sting? *Qeios*. 2023. doi:10.32388/cqctkx

4. How can we tell if an AI is conscious?

The previous chapters have discussed the plausibility of AI-consciousness and the moral status that ought to be assigned to it. However, that leaves one last fundamental question to answer: how will we be able to determine if and when AI models do become conscious?

Recent reviews of the most prominent theories of consciousness have found insufficient evidence to definitively classify any current AI model as conscious under any of these theories [1,2]. As many theories of consciousness contradict and compete with others, it is impossible for all to be proven correct. Attempts to rule out rivals are ongoing.

As an alternative route, this chapter advances the Building Blocks meta-theory, which does not seek to provide an answer to the ontology and nature of consciousness, but only to stipulate what characteristics and attributes of an entity's cognitive architecture are required for that entity to be confidently classified as conscious.

Published in *Philosophies*, the Building Blocks theory proposes nine such characteristics, which are treated as necessary (but not necessarily sufficient) that, when all are present within an entity, they provide strong defeasible evidence of that entity's phenomenal consciousness. The paper fully defines, explains, and discusses each building block in detail.

As this theory does not posit a specific mechanism (but a set of observable capacities that any such mechanism must implement), it is compatible with established theories of consciousness (such as Global Workspace Theory/IIT/HOT, Active Inference, etc.) without committing to any. The Building Blocks theory may be better thought of as a meta-theoretical operationalisation, not a true metaphysical theory, designed so that physicalists, dualists, and others can still accept the attributes as empirically testable criteria.

This chapter, therefore, does not claim to solve the hard problem or settle metaphysics; it provides a necessary-conditions testbed for later empirical probing of AI models. The Building Blocks theory is designed to be compatible with multiple metaphysical views, and focuses on the epistemological and operational criteria for determining consciousness for the purposes of advancing policy, rather than solving its ontological questions.

Each of the building blocks B_i may be treated as scalars. Let $B_i \in [0,1]$ denote the measured degree of block i . For decision purposes, we binarise with thresholds $\tau_i \in$

$(0,1]: \mathcal{B}_i := 1[B_i \geq \tau_i] \in \{0,1\}$ ⁴. In this manner, if a building block is discovered within an entity’s cognitive architecture, its strength or degree relative to a human’s is not taken into account. Its mere presence is sufficient to be counted amongst the entity’s building blocks.

The classification of consciousness is, thus, inferentially defined as when all B_i are positive⁵:

$$Con := \bigwedge_{i=1}^9 (\mathcal{B}_i = 1) \in \{0,1\}$$

If any $\mathcal{B}_i = 0$, classify as non-conscious.

Note that if any single building block is not found to be present, then the entity cannot be classified as being conscious. The Building Blocks theory works on an “all or nothing” premise. Note that this is a stipulative threshold for practical classification and policy purposes, not an absolute claim about metaphysical impossibility of conceivable edge cases presenting as false negatives.

Using these nine building blocks of consciousness, the thesis can move forward in Chapter 5 to apply the Building Blocks theory to contemporary LLMs as a worked assessment, directly addressing RQ1. The nine attributes will also further be used as a fundamental piece of Chapter 7’s list of attributes of the self, and in the construction of a speculative conscious, self-aware AI agent in Chapter 11. It also underpins all later references to “degree of consciousness” in the ethical frameworks.

4.0.1. References

1. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv [cs.AI]*. 2023. Available: <http://arxiv.org/abs/2308.08708>
2. Chalmers DJ. Could a Large Language Model be Conscious? *arXiv [cs.AI]*. 2023. Available: <https://arxiv.org/abs/2303.07103>

⁴ Thresholds τ would be any measurement to differentiate the building blocks from random chance, environmental perturbations, or statistical noise.

⁵ Whilst the formula highlights the nine building blocks, it does not imply joint sufficiency. Metaphysically, this represents the currently observable minimum criteria, but do not preclude the future discovery of additional requisite building blocks for ontological sufficiency.

4.1. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness

Izak Tait^[0000-0002-0274-8480]¹, Joshua Bensemann^[0000-0002-1301-4861]², Trung Nguyen^[0000-0002-1990-8647]^{1,2},

¹Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand

²The NAOInstitute, 314 Khyber Pass Road, Auckland, New Zealand

izak.tait@autuni.ac.nz, joshuabensemann@gmail.com, trung.nguyen@aut.ac.nz

Abstract: For consciousness to exist, an entity must have prerequisite characteristics and attributes to give rise to it. We explore these “building blocks” of consciousness in detail in this paper, which range from perceptive to computational to meta-representational characteristics of an entity’s cognitive architecture. We show how each cognitive attribute is strictly necessary for the emergence of consciousness, and how the building blocks may be used for any entity to be classified as being conscious. The list of building blocks is not limited to human or organic consciousness and may be used to classify artificial and organisational conscious entities. We further explore a list of attributes that seem intuitively necessary for consciousness, but on further investigation, are neither required nor sufficient. The building blocks do not represent a theory of consciousness but rather a meta-theory on the emergence and classification of consciousness.

Keywords: consciousness; cognition; theory of consciousness; meta-consciousness; meta-cognition

4.1.1. Introduction

A difficulty at the intersection of philosophical and psychological investigation into the nature of consciousness is understanding when one can properly attribute consciousness to an entity. This is in contrast to research on how consciousness is expressed and has evolved [1–3], how it correlates to neuronal and behavioural activities [4–6], or how it functions [7,8]. Our goal in this paper is to complement existing literature on consciousness by identifying the attributes and characteristics required for the attribution of consciousness.

The building blocks of consciousness are an entity’s neural, behavioural and mental characteristics that allow it to generate consciousness [3,9–11]. A building block means something necessary for consciousness to exist within an entity (regardless of what that entity may be).

As well as being necessary, we consider that the full list of nine building blocks is also likely sufficient for an entity to have consciousness. We expect an entity that has already

been commonly classified as conscious (such as humans, mammals, and most vertebrates) to have all nine building blocks; while entities that have, thus far, been ruled out as having consciousness to lack one or more of the building blocks. This is because each building block is fundamental to an aspect of consciousness as described in the literature, without which consciousness cannot exist; and we have found no conclusive evidence of characteristics fundamental to consciousness (as described by the most prominent theories of consciousness [7]) beyond these nine building blocks.

The nine building blocks may likely not be an exhaustive list. We are open to, and encourage, the list's expansion in the future by the addition of further building blocks (or splitting existing blocks into several distinct items). We also encourage the expansion of the negative list in Section 3 to better create a distinction between what is and what is not required for consciousness. Until then, we propose that should an entity display all nine building blocks' characteristics, it ought to be sufficient to classify it as having consciousness unless there is substantial evidence and a strong argument to say why it does not.

We will not favour any specific theory of consciousness (TOC), as the building blocks below focus on the entity's attributes rather than the nature of its consciousness. Thus, the paper will not strictly be about consciousness but what is a prerequisite to obtaining it. While we will endeavour to be as neutral and unbiased as possible regarding the TOCs, please note that not all theories of consciousness will be compatible with our proposed building blocks, particularly the more heterodox theories such as panpsychism [12,13], and obviously theories that deny the existence of consciousness [14].

To expand our neutral stance, we aim to be as unbiased as possible about what sorts of entities can have consciousness. While the predominant consensus in the literature points only to humans and certain animals having consciousness [15], it is not beyond the realms of possibility for other entities to have been overlooked or prematurely dismissed. With the accelerated speed of research into machine consciousness, it may not be long before we will have to seriously consider whether an AI model is conscious or not [16]. For that, these building blocks may act as the requisite milestones that an AI model must reach before we can ascribe consciousness to it.

Organisations, too, have the potential to be classified as conscious should they meet the requirements for all building blocks. They may take the form of corporate or legal organisations comprised of humans, large groups of codependent animals, or interconnected collections of non-animal entities such as fungal or plant networks. If the organisation as a whole displays the building blocks' attributes, this could be sufficient to classify it as conscious, even if its constituent entities are not classified as conscious (for

example, an ant colony may have all nine building blocks, while the individual worker ants and drones do not).

Most speculative yet, should we ever be (un)fortunate enough to discover complex extraterrestrial life (or if they discover us first), we will need a framework to guide us to discern if they exhibit consciousness or not. Thus, we believe these building blocks will also offer some guidance to future astrobiologists.

The use of the building blocks as a classification guide is a key goal of this paper. Whether to investigate new organic species for their capacity for consciousness, using the building blocks as milestones for AI, or envisaging new entities such as organisational intelligences as conscious entities, the building blocks would serve as a robust guide.

Before we truly begin, it behoves us to provide a working definition of consciousness that will form the basis for the building blocks below. To paraphrase Seth and Bayne's 2022 definition [7], we would define consciousness as an entity's suit of subjective mental states, including both the global states linked to arousal, wakefulness, and behavioural responses, as well as the local states with phenomenal content and functional properties. Together, both local and global states provide an entity with the ability to be aware of, and respond behaviourally to, its internal and external environment.

4.1.2. The Building Blocks of Consciousness

In this section, we look at each building block independently and discuss why each is required for consciousness, yet is not sufficient by itself. Each subsection has two parts: an argument supporting the inclusion of the building block in the table, and a short review providing evidence from academic literature.

The building blocks are summarised in Table 1 through lay-descriptive and narrative examples for ease of understanding.

Table 2: Attributes and characteristics that are required for the development of consciousness.

Attribute	Description and Example
Embodiment	Information and its processing exist somewhere. <i>I am within my body.</i>
Perception (intero and extero)	Information enters from the environment and is made available for processing. <i>I see an apple.</i>
Directed attention	Information is given specific attention. <i>I pay attention to the apple.</i>

Recurrent Computing and Processing	Information is processed in and between several regions of the cognitive architecture. <i>My brain processes the image of an apple.</i>
Ability to create Inferences	Missing and incomplete information about the environment is generated. <i>I infer that it is a Granny Smith apple.</i>
Working Memory	Information is transiently maintained in an active state while it is processed. <i>I have been watching the apple for several seconds.</i>
Semantic understanding	Computing and sensing processes are understood. <i>I am aware that I am looking at an apple.</i>
Data output (internally or externally)	Information is generated and made available for further perception and action. <i>I think about apples.</i>
Introspection, meta-cognition, meta-representation	The cognitive architecture interrogates and investigates itself. <i>I think about me looking at the apple.</i>

The building blocks are not ordered in, or intended to be, a hierarchy. Much like the famous Danish toy blocks, these building blocks may be ordered in any arrangement to form the foundation of consciousness. While we argue that all the building blocks below are necessary (and potentially sufficient) for consciousness, we make no claim to the form they take to do so. In addition, we do not imply that consciousness will spontaneously appear should an entity have all the building blocks below, and we will not hazard a guess as to the means by which it will do so. Those are the domains of theories of consciousness and are beyond the scope of this paper.

Note that whenever the term “environment” is used to describe the consciousness’s environment, this does not refer to the environment around the body in which the consciousness resides. Rather, it refers to everything that is not the consciousness, but which interacts with it. This means that the brain is also treated as the consciousness’s environment. We do not intend this to be an argument for dualism or against physicalism but rather a convention for ease of communication.

4.1.2.1. Perception

1. *To have phenomenological consciousness is to have a subjective experience of the environment.*
2. *A subjective experience of the environment requires information exchange from the environment to the consciousness.*

3. *This information exchange requires a method of perceiving information in the environment.*
4. *Ergo, consciousness requires a method of perceiving information in the environment.*

Consciousness, whether as a phenomenological experience or as an act of attending to a matter, is defined by its interaction with the environment and receiving information from it, even if the nature of this relationship may be disagreed on [17]. To be conscious is to be conscious of something; phenomenal consciousness is about experiencing something.

Perception can be classified into three separate modes and three overlapping stages. The three modes are:

Exteroception: perception of the environment outside the entity's body or housing. In humans and animals, this is best exemplified by the five classic senses; in machines, this can be microphones and cameras; and for organisations, this is the methods of communication allowed in it.

Interoception: perception of the environment outside of the consciousness, yet within its body or housing. For animals, this includes hunger/satiety and proprioception; for machines, this may be sensors ensuring their housing and power supply is at optimal conditions; and for organisations, this may include custodial and HR oversight.

Introspection: perception and examination of the consciousness's own mental states, processes, and existence. Section 2.8 will cover this in greater detail; however, in animals, introspection can be displayed via thoughts and mental images; in machines, via software that scans other software; and in organisations, through auditing processes.

From these three modes, an entity has a perceptive experience of the whole of its existence. It can perceive itself, its embodiment, and the environment around it. All three modes are not required for an entity to be considered conscious, but at least one is. A person who, through an unfortunate accident, loses access to his exteroception would not suddenly lose his consciousness.

But how does an entity perceive its environment through these modes? According to Audi, this is done in three parts or stages, which together cause the phenomenological experience [18]. The first part is the simple perception of a scene (if visual). "I see an apple," would suffice. The second part is to perceive something within that scene "to be" a certain way or having a certain attribute. When seeing the hypothetical apple, I can perceive it "to be" spherical even if I can only see one angle of it. When seeing its shadow, I can perceive that there is a leaf hidden from view, as the shadow seems "to be" a leaf-shaped shadow. The final part is perceiving "that" the object is what you have perceived it

“to be”. I have seen an apple and perceived it to be spherical and having a leaf; therefore, my experience of it is as a spherical, leaf-having apple. When all three stages of perception are taken together, they give rise to further cognitive processes, such as inferences, semantic understanding and meta-representation, that lead to phenomenal experiences.

The final stage of perception, of perceiving “that” something is what it seems “to be”, is most crucial to consciousness. By integrating all the characteristics of the perceived object, the final stage creates a single binding and unified perceptive experience [3,10]. I do not merely perceive the sphericalness and redness of the apple, but I perceive that it is an apple. It is this final converged and singular perceptive event which gives rise to qualia (the introspectively accessible, subjective, phenomenal character of an experience) and consciousness.

But you should not take these perceptive experiences to be exclusively for the external environment. Interoception is as much, if not more, important to realising consciousness. The holistic sensory experience (conscious or unconscious) of internal signals from the body or housing plays a vital part in establishing a subjective first-person perspective for the consciousness [19]. Sensing its own body/housing allows the entity to differentiate itself from the rest of existence. Even if the entity and its consciousness should merely exist as Dennett’s brain in a vat, it will be able to introspectively perceive and attend to its own thoughts and perception of them, causing a subjective first-person experience to emerge [4]. The mind itself becomes the environment surrounding the consciousness in the absence of any interoceptive or exteroceptive environments.

4.1.2.2. Embodiment

- 1. Consciousness is affected by its environment.*
- 2. A physical environment requires a physical element to this causal relationship.*
- 3. This relationship, and the experiences produced by it, occur at specific times in specific locations.*
- 4. This means that the connection between consciousness and the cognitive architecture is localisable in both time and space.*
- 5. A spatial and temporal physical embodiment is required for a definable perspective point.*
- 6. A unique definable perspective point is required for a subjective, first-person perspective.*
- 7. Ergo, an embodiment is required for consciousness.*

While a definition of consciousness rarely includes terminology of embodiment, it is a presupposition upon which many of consciousness attributes and characteristics rest [20,21]. Whether consciousness is classified as an epiphenomenon or as a complex set of phenomenological qualia, it is uncontroversial to say that consciousness is affected by its environment. There is a cause-and-effect relationship from the environment to the consciousness (and perhaps vice versa, depending on the TOC [22]) that is both spatial and temporal in nature.

For this to be true, as the logic above shows, there must be some form of embodiment for the consciousness to inhabit. From a philosophical perspective, this remains true regardless of where you position yourself on the spectrum between dualism and physicalism. For physicalists and materialists, the embodiment of the consciousness is treated as the default. Wherever the brain may go, so must the consciousness, as they are one and the same.

For dualists and idealists, to whom consciousness is non-physical, and to whom the relationship between it and the brain is a point of contention, the relationship between the consciousness and the environment must still terminate in a physical location (the brain). The consciousness can only gain information via this relationship. This is not to say that consciousness must be in the brain in dualism or idealism, merely that the brain mediates the relationship between consciousness and the environment. Our conscious experience depends upon physical processes, even if these physical processes and an entity's embodiment are ultimately grounded in mental facts, and a subject's mind is outside of space and time.

If you broaden the concept of consciousness to machines and AI, the notion of embodiments becomes less abstract. The lines of code from which a potential artificial consciousness arises must be stored on some processing and memory unit. Whether housed within one machine or spread across the virtual cloud, it is still somewhere specific in space and time.

The concept of embodiment becomes far more nebulous when looking at the potential for organisations to be called conscious entities. An organisation is irreducible; it is not its people or structures, but instead is the holistic whole that is greater than the sum of its parts. As such, the embodiment of an organisational consciousness is a social and memetic construct. Its embodiment is located where people believe it is. Most often, this can be a physical location, particularly for traditional organisations, but for multi-national institutions and internet-based communities, the embodiment is located at the

confluence of their constituents' social network, rather than in the bricks and mortar of any one building.

The number of input/output modalities of a given consciousness does not go against the principle of requiring an embodiment but does add multiple facets to this embodiment. A machine consciousness is most easily imagined as having multimodal inputs/outputs, as a cloud-based consciousness can interact and be affected by the environment through any number of terminals and devices. An organisational consciousness can likewise interact with its environment through any of the conscious entities that form part of it. Neither of these situations speak against the embodiment concept but show how embodiment can be fractured and spread across multiple locations.

Daniel Dennet's riveting essay "Where am I?" shows how multiple modalities can cause these unusual facets to the consciousness's embodiment. In the essay, his brain is removed and placed in a life-sustaining vat while remote-controlling his brain with electronic assistance [23]. Throughout the essay, he experiences his environment within the vat, in his original body, and in a second body. These modalities change, yet they (and he) are always somewhere at some time, subjectively moving from "body" to "body", yet always having a physical point in spacetime with which to access the environment.

Another important argument for why embodiment is a requirement for consciousness is that of the first-person perspective [24]. To have a subjective experience from the first-person point of view, one requires an embodiment from which to view the world. You need to be able to reference your place in spacetime in relation to the environment and be able to categorise yourself as distinct from the environment [4]. This referencing may be through conscious introspection, or through unconscious and pre-reflective awareness of itself as an individual entity [25]. Whether through the proprioceptive, somatic and other sensory input of a biological entity, the virtual tagging of a machine, or the social identity and labelling of an organisation, an embodiment allows the entity to make itself distinct, unique, and subjective.

4.1.2.3. Attention

- 1. Entities are not consciously experiencing their entire environment simultaneously.*
- 2. Conscious experience of the environment is limited to specific stimuli, such as scenes, objects, or views.*
- 3. This limited conscious experience requires a means to discriminate between stimuli.*

4. *This discrimination method is achieved by selectively attending to specific stimuli for further processing by the cognitive architecture.*
5. *Ergo, consciousness requires attention.*

Attention is how an entity limits the information it gives priority resources to by selectively focusing on a specific selection of incoming information [10,26]. This attention may be triggered by a bottom-up process, whereby an external event attracts an entity's attention (e.g., a flicker of movement, a distinctly recognisable sound), or through a top-down approach whereby the entity chooses to focus on a specific event or region, externally or internally [3,27,28]. The former may be argued to be involuntary attention, and the latter voluntary.

It is important to note that attention is not solely directed at the external environment, or at the interoceptive space within a consciousness's embodiment. Introspection, as discussed in Section 2.9, is defined partly by the entity's ability to direct its attention to its own cognitive processes. Attention may thus be directed at all the modes outlined in Section 2.1, from an ice cream truck's jingle heard externally, to the internal feelings of hunger, to the thoughts and memories of childhood and its connections to the present interoceptive and exteroceptive stimuli.

In this short, hypothetical example above of phenomenally experiencing an ice cream truck, the observer's attention moves from extero- to interoception to meta-cognitive. This, perhaps egregiously controlled example, shows that to be conscious of each element in the phenomenological experience, you must first pay attention to that piece of information. Attention is the mechanism by which certain information passes from unconsciousness to consciousness, and other pieces of information do not. A person is clearly not attentive to every photon entering the eye or every sine wave entering the ear when one walks down Main Street. When observing the ice cream truck and reminiscing about his childhood, the observer could very well not have been aware of a cat sitting quietly in his peripheral vision, and why should he have? His attention was focused elsewhere, after all.

While we argue that attention is a prerequisite for consciousness, the reverse is most definitely not true. One can attend to information, and directs one's attention, without it ever entering conscious processing [29]. The archetypal exemplar of this is sleepwalking. Those who suffer severely from parasomnia and somnambulism do not merely walk while asleep, but can perform household chores, get dressed, eat, and even leave the house to drive while appearing entirely unconscious. They attend to the world around them, and direct their attention to what they are doing, but when they awake, they have little to no

memory of anything they have done. As the cerebral cortex and cerebrum show little activity during sleepwalking (both required for integration of information and recurrent processing) [30], it is attention without consciousness. There is even more than one famous case of a sleepwalker committing murder, which, one presumes, would require a great deal of attention.

More benign and ordinary instances of unconscious attention are when one's environment becomes routine enough to stop focusing on it. Dust on a bookshelf, for example, or a box of old clothes that one ought to have donated months ago. Such ordinary things may affect one's behaviour without ever consciously being aware of them [31,32], such as absentmindedly wiping off dust when walking past a shelf or needing to adjust where one walks to not walk into that old box of clothes. Habits are formed as conscious awareness becomes unconscious attention.

One can thus frame consciousness as not attending to information received, but rather being aware of the cognitive representation of information that the architecture has already attended to [33]. In this sense, it is the interplay between working memory (holding the information that has been attended to unconsciously), the meta-representation thereof, and the recurrent processing performed in the cognitive architecture that gives attention its conscious awareness [9,34]. This awareness of a mental representation of a subset of external information provides the consciousness with a view of the world, which it can phenomenally experience.

An ant colony, as an organisational entity, can also be said to have selective attention. Each worker brings its own information to the colony, yet the colony as a whole does not respond to each piece of data brought to it. However, the colony can act as a unified organism towards threats or opportunities should the correct information be attended to and integrated throughout the colony [35].

Of curious note, however, is that current AI models require directed attention because their computational resources are limited by hardware concerns (much as our cognitive architecture is). Theoretically, given enough hardware, an AI model (potentially a conscious one in the future) may not need to specifically direct its attention, as it will have enough computational resources to attend to its entire environment (internally, externally, and introspectively). In such a scenario, we would argue that it is not a lack of attention, but a totality of attention that the AI model has. It had outgrown the need to focus and narrow its attention, but can encompass its entire environment in one holistic, directed attention.

4.1.2.4. Recurrent Computing and Processing

- 1. A phenomenally conscious experience involves the coordinated processing of sensory, cognitive, and affective information from different sources to generate a coherent representation of the environment.*
- 2. This coordinated processing requires complex activity in several regions of the cognitive architecture.*
- 3. Recurrent processing and feedback between these specialised regions of the cognitive architecture are necessary to allow the exchange and refinement of information and representations of the environment.*
- 4. Without recurrent processing, the processing of information would be limited to individual regions of the cognitive architecture, leading to fragmented and disconnected representations of the external and internal world.*
- 5. Ergo, recurrent computational processing is required for consciousness.*

We consider recurrence as another of the building blocks of consciousness. One reason for this is that there has to be an element of recurrent activity for consciousness to exist within the brain. Without any recurring activity, it would be difficult to perform complex computing as the cognitive processes required to do so would need to be completed within a single pass of the brain. However, by introducing some form of recurrence, the cognitive processes can persist as long as required to complete that processing [36]. This point is not unrelated to the importance of working memory that we have emphasised elsewhere.

Empirical support for our position comes from the fact that many examples of recurrent activity in the brain have been documented in the literature, several of which are related to conscious processing [37]. While this does not prove that recurrence is strictly necessary for cognition, it does suggest that recurrence is essential to how the human brain produces consciousness.

Generally speaking, recurrence can take two forms for maintaining information in the brain. One is persistence, where localised activity is maintained to keep information available, such as the process suspected to be involved in working memory [38]. The other is where information is shared between areas of the brain. This means that the information can be routed to multiple areas of the brain, each contributing to the overall cognitive process while keeping the information in existence.

The information-sharing version of recursion forms the backbone of one prominent theory of consciousness, the global workspace theory [39], and its completely neuroscientific version, the global neuronal workspace [40]. In these theories, the

information becomes recurrent only if it passes an attention threshold and becomes identified as of importance to other areas in the brain. According to the neuronal version of this theory, any time information crosses this threshold, a temporary connection forms between the relevant processing centres of the brain to allow this information sharing. Thus, important information becomes recurrent when sent back and forth between multiple processing areas.

In theory, a deep neural network (DNN) could mimic the functioning of the GWT workspace by reverberating information being learned across multiple tasks in that network to all tasks being learned. Doing so might improve the performance of individual tasks. Research with multi-task DNNs has suggested that removing the independence between tasks can lead to improvements in the learning of those tasks [41,42]. An example of this could start by pooling initial task output into an area accessible by the separate sub-modules of the DNN to eliminate multi-task independence. Based on this shared information pool, it would then train the individual tasks to moderate their initial outputs. The information entering the shared space would not be ad hoc but rather would be optimised by punishing the network for sharing irrelevant information.

Having discussed how our recurrent processing requirement is compatible with global/neuronal workspace theories, it is of interest to briefly mention how recurrence relates to the other competing theory, integrated information theory (IIT) [43]. In IIT, consciousness is not an all-or-nothing concept; rather, it identifies a metric (PHI) that measures how integrated the information within a system becomes. This means that a human's level of consciousness can change over their lifetime, with less integration occurring while their brain is developing (i.e., lower PHI scores) compared to when they are fully mature. In IIT, information is, as the name suggests, integrated. In some cases, that integration allows information to be shared in a form of recurrence. Given that IIT provides a consciousness metric, it would be of interest to see whether systems with recurrence tend to have higher PHI score. If this were found to be the case, it would support the inclusion of recurrence as a building block.

4.1.2.5. Ability to Create Inferences

- 1. A conscious entity has incomplete information about its environment due to the limitations of its sensory inputs.*
- 2. To create a completed picture of the environment, the entity's cognitive architecture must build a representation of the environment with which to interact.*

3. *This representation is generated from inferences drawn from various parts of the cognitive architecture.*
4. *The information generated from inferences is not limited to perceptual data but may include cognitive information such as feelings.*
5. *This generated information is then available to be outputted as experiences.*
6. *Ergo, the ability to create inferences is required for consciousness.*

An entity never has a complete view of its environment, be it the inner world of its embodiment, the external world beyond it, or even the introspective view of its own cognitive architecture. This is because the sensors it has of its environment limit the information it receives at any one time. Our memories are not perfect, our vision has blind spots and areas of inattention, our hearing is rather limited, and our interoceptive sensors are mostly unconscious.

Yet, from a subjective point of view, we do indeed have a ‘completed’ view of our environment, if not entirely ‘complete’. This incongruity is solved via the use of inferences by the cognitive architecture. Through the use of inferences, the cognitive architecture generates the “missing” information to provide the entity with a unified view of its environment. This generative act of building inferences is not merely limited to perception but can be modelled throughout the cognitive processes, even to a meta-cognitive level.

Note that inferences in this section do not imply causal inference in the way that a cognitive architecture is estimating and reasoning a cause behind an event. Rather, one can think of the basal inferences involved in perceptual events referred to here as analogous to the interpolation of statistical data.

One can look at inferences from a hierarchical perspective. The “lower” levels, closest to the sensory inputs, would limit their inferences on what is being perceived to create a unified perceptive view [44] and build the meta-representations of the environment. Higher levels further removed from the sensorium would then continue the meta-representational process by generating inferences on the holistic view of all inputs. Throughout these sensory inferences, the cognitive architecture would build its “best guess” of the environment, continually updated through sensory information to reduce any errors in its predictive guesswork [45].

This predictive inferential model works equally well in memories. Biological memories are notoriously unreliable, partly due to the inferences worked into them. As memories are recalled (consciously or unconsciously), additional information is generated to fill in any missing areas to construct a narratively enjoyable memory. This can often occur due to cues in the present environment that colour the perception of the memory and the

inferences generated [46]. This generated information can then become part of the memory when it is next recalled and then reiterated upon with further inferences until all that is left of the subjective past is imagination.

Inferences are not purely perceptual in nature. Subjectivity, feelings and emotional states also have inferential components. In biological creatures, emotion relies heavily on interoceptive inferences of the physiological changes of the embodiment [47]. Yet, from a conceptual point of view, an entity's "feeling" about a given matter can be seen as the inferences made of the differences and juxtapositions between the internal and external environment. Perceptions of the external world are correlated with the physiological state of the internal environment by generating predictions of how the two should interact and correlate. As the physiological internal state changes and as the perceptions of the external world change, these predictions no longer hold true and "prediction errors" crop up. Inferences of an emotive nature are generated to reconcile these prediction errors that give rise to subjective feelings and, in turn, a phenomenological experience.

The subjective first-person experience is also generated by this reconciliation of exteroceptive and interoceptive environments [47,48]. As an entity develops, it builds an understanding of what and where it is in relation to the environment by generating inferences. This is as much a social as it is perceptual. By perceiving that others exist, the entity infers that it is different from them. By perceiving the external world and the limitations the entity has in its interactions with it, it can infer that it is not the same as the environment. As no aspect of the internal or external environment is static, the predictive model in the entity's cognitive architecture is also constantly updated to generate new inferences about itself and its environment [49,50].

An example of this is expected changes over time. As we see ourselves and those around us ageing (particularly if we have not seen someone in a significant time), we view new iterations of ourselves as others, rather than the same exact view from our memories. We infer who they are, and who we are ourselves when we look in the mirror, through the predictions that our cognitive architectures have built. Despite minor changes in appearance, there is a continuity of experience through the use of active inferences [48].

4.1.2.6. Working Memory

- 1. A conscious experience involves information processing by the cognitive architecture.*
- 2. A specialised unit or process is required to maintain transient information as it is being processed in various regions of the cognitive architecture.*

3. *Working memory is responsible for holding and maintaining said information.*
4. *Ergo, working memory is required for consciousness.*

Working memory (WM) is considered a cognitive system that stores, maintains, and processes online (short-term) information relevant to the very immediate/current task [51,52]. The information maintained in WM is about what is currently being thought about and experienced to be consciously considered. Therefore, it is arguable that some processing of conscious experience occurs here. Specifically, there are several general views on the relationship between WM and consciousness in which they are considered equal or closely working together [34].

The first popular view is that WM is closely related and implicitly considered equal consciousness. In other words, the content held in WM is consciousness, as shown in Baddeley's multicomponent model [51,53]. This WM model is hierarchical, with a central executive and several "slave" systems, where the central executive controls the slave systems. The slave systems are either modality-specific information or the episodic buffer of polymodal episodes [53]. While the "slave systems" provide consciously experienced representations, the central executive may relate to consciousness by providing conscious access to items held in WM. Many theorists assumed this view in the recent past. This idea, however, is hard to stand with the evidence of unconsciousness content in WM.

Another view is to consider consciousness as an integrated component of WM which enables WM to contain unconscious content. This idea was demonstrated in the embedded-process WM model [54], activation-based models of WM [55,56] and the opinions of having a distinction between conscious WM content and unconscious WM content (which may become conscious) [57]. In this view, selecting WM content as the focus of attention could be the primary mechanism mediating the relationship between WM and consciousness as it activates original WM content to enable conscious access to the content [34].

The second view is favoured by Jacobs and Silvano's argument [58] that WM representations and conscious representations serve different functions, have different effects on behaviours [59–61], and have distinct representations [62,63]. Therefore, Jacobs and Silvano proposed [58] that introspected contents are the "conscious copy" of WM representations. This model suggests that WM representations are intrinsically unconscious, and consciousness is not the same as the WM trace, even if the WM content is activated. Nevertheless, conscious experience is rooted in WM in this theory and, therefore, requires WM to exist. Jacobs and Silvano also mentioned another opinion

that WM is a subset of consciousness, opposite to the second view [58]. This view again supports the requirement of WM for consciousness, although it requires further empirical evidence.

The relationship between consciousness and WM is an ongoing research topic with divergence. Although the relationship can be hierarchical, parallel interaction, or even considered identical, WM is integral in the workflow of consciousness in all models and general views of this research. WM's equivalence has been applied and integrated into artificial computational systems, from computer hardware to complex models. In computer hardware, RAM is considered to be equivalent to WM, fulfilling many of the same requirements and processes, with introspected memory extracted from peripherals and local/cloud data. In Franklin's implementation of the global workspace theory (GWT) [64,65], i.e., intelligent distribution agent [66], WM is considered equivalent to the whole global workspace, which supersedes consciousness. What this may mean for potential organisational conscious entities is that the people who make up the organisation play the role of WM, with their own memories and consciousnesses working to maintain the transient information in the organisation's communications networks.

4.1.2.7. Semantic Understanding

- 1. Subjective perceptual awareness differentiates conscious from unconscious experience.*
- 2. Awareness of a specific scene, object or view in the environment requires understanding that the perceptual process is occurring.*
- 3. To be aware that an experience is subjective, the entity must understand that it exists in some capacity.*
- 4. Ergo, semantic understanding is required for consciousness.*

There is a fundamental difference between processing information and being aware of it. Haikonen describes the distinction between the two types of processors when discussing weak AI versus strong AI [67]. In weak AI, information can be input into a system, processed, and then output without the processor having any experience of those processes. There are many examples of processing without experience in neuroscience (see Dehaene & Naccache [38]). For AI, the information input into the system is transformed into a number, making it indistinguishable from any other process occurring in the AI. In other words, the weak AI is limited by the symbol grounding problem.

To achieve a strong AI, Haikonen argues that the symbol grounding problem must be solved so that the AI has some form of experience of the outside world that is different from converting inputs into numbers. While previously arguing that this problem might be impossible to solve with digital computers [68], Haikonen does explain what successfully solving the problem looks like, using humans as an example. Essentially, the problem is solved if the processor has some kind of experience about the information it is processing, such as qualia experienced in humans.

Qualia are one of those constructs that have many conflicting definitions. While there are many different points of view on what qualia are [69], we will briefly outline one from Ramachandran and Hirstein [70] to help explain what it is the above paragraph describes. Ramachandran and Hirstein provide three criteria that must be satisfied for things to qualify as qualia. The first is that the experience is irrevocable: The agent experiencing the quale cannot use cognitive effort to change what they are experiencing. For example, blue always appears as blue; we cannot choose to see it as red.

Ramachandran and Hirstein's second criterion is that the agent must have some choice about how they react to the experience [70]. In their paper they give the example of this is the difference between how a coma patient and an awake individual react to light being shined into their eyes. The coma patient may only constrict their pupils via a reflex, so they are not consciously experiencing the light. In contrast, the normal individual experiences this because they can close their eyes, turn their head, complain about the light, or do any other action.

The third criterion is that the experience needs to enter working memory [70]. Here, Ramachandran and Hirstein's criteria for qualia intersects with our building block as working memory is one of our building blocks. If a stimulus does not enter working memory, then it is detected and processed without the agent's awareness. If the experiencer is unaware of the stimulus, they cannot factor in that experience into their decision-making and so fail to meet the second criterion. As the first and second criteria are impossible without working memory, the third criterion appears to be the most important for determining whether something is experiencing qualia.

The above definition is just one description of what is required for something to be an experience. We do not commit to Ramachandran and Hirstein but have included it as a useful description of what semantic understanding can be. The example also helps highlight the importance of working memory as a building block. However, we should point out that the semantic understanding building block is a specialised task and is distinct from working memory. One should also compare the view we have included here

to others, such as Reggia et al. (2016) [4], who further discuss experience from a self-modelling perspective. This is a much larger topic that we will not expand upon here.

4.1.2.8. Data Output

- 1. Subjective experience is a crucial component of phenomenal consciousness.*
- 2. This experience includes perceptive and/or phenomenal elements such as thoughts, emotions, feelings, words, and actions.*
- 3. These elements do not come from sources external to the cognitive architecture.*
- 4. These perceptive elements must thus be created by the cognitive architecture to be subjectively experienced.*
- 5. Ergo, data output is required for consciousness.*

Crucial to the definition of phenomenal consciousness is the experience of an event or scene from a subjective first-person perspective. This experience of “what it is like to be” at that time, at that point, within that embodiment, localised entirely from that perspective, is not an empty event devoid of characteristics. It is the characteristics and attributes of the experiences which allows them to be felt by the entity. These qualitative elements of the experience are not imported or downloaded from the environment; else it would be possible for more than one entity to have the same subjective experience of an event. This means, then, that some form of information generation is required to create these experiential characteristics.

Three obvious and intuitive types of experiential information may first come to mind, but they are not truly required for consciousness. Yet, for those who can generate them, these play an important role. These are verbal/aural and visual mental imagery, and emotional information. Section 3 below will detail the reasons why these are not required for consciousness due to disorders such as anauralia, aphantasia and alexithymia. External behaviours are also not required informational content for consciousness, as individuals with severe paralysis or pseudocomas can still show evidence of consciousness and mental responsiveness without being able to show it externally.

This leaves one prominent piece of information that can be generated regardless of known disorders and is vital to the definition of a phenomenal experience: a feeling. As discussed in Section 2.5 above and Section 3 below, feelings are distinct from emotions in that they do not require a physiological response and encompass a broader spectrum of content than simply emotions. Emotions may also have unconscious aspects to them [71], while feelings need to be consciously felt.

For an entity to ‘feel’ anything, the cognitive architecture must generate that ‘feeling’ in some form or fashion.

When experiencing an event/scene/memory/etc., there is a time when there is no qualitative feeling associated with that experience, and then at the next moment the feeling is spontaneously there. As much as it is a truism, there is no feeling until there is a feeling. This holds true even in dualistic and idealistic theories of consciousness, where these qualitative feelings are not said to be produced, directed, or transferred by the entity’s cognitive architecture. As feelings are inherently subjective and will differ from entity to entity when experiencing the same event, we can confidently say that these feelings are not provided to the consciousness from the environment external to the entity’s embodiment. The envatment notion put forth by Dennet in 1978 also shows that interoception is useful yet insufficient for a phenomenal feeling. This is because the interoceptive signals from the entity’s embodiment may be mimicked or even supplanted by artificial exogenous inputs. These signals may even be completely shut off, hypothetically speaking, leaving nothing but the brain to be the sole point of input and output for itself. Therefore, when exteroceptive and interoceptive input is removed, even if an anti-materialist stance is taken, the cognitive architecture is the only conduit between the physical realm and the consciousness where a feeling may be generated or felt.

Data generation is also vital to other building blocks, as both meta-cognition (Section 2.9) and inferences (Section 2.5) have elements that require the creation of information. As such, data output plays both a causal and affected role in establishing consciousness.

4.1.2.9. Meta-Representation and Meta-Cognition

- 1. Basic processing of sensory input is insufficient to create a phenomenally conscious experience.*
- 2. To form such an experience, the perceptive elements require further processing in other areas of the cognitive architecture, which includes attention, working memory, semantic understanding, and inferences.*
- 3. The different areas of the cognitive architecture do not work on the basic sensory input itself, but on the representations of this sensory input created by the perceptive cognitive structures.*
- 4. This is a recursive process of mental representation that involves the cognitive architecture creating mental representations of its own mental representations, including thinking about its thought processes, also known as meta-cognition.*
- 5. Ergo, meta-representation and meta-cognition are required for consciousness.*

Meta-cognition and meta-representation, and even introspection, all have the common element of one section of a cognitive architecture creating representations of another, separate section.

Introspection may be the most obvious as it is the subjective examination, inspection and perception of the cognitive architecture by the consciousness [72]. We know that we are conscious because we can think about being conscious. Or, as Descartes famously said, “Je pense, donc je suis”: I think, therefore I am [73].

Meta-cognition is broader still; it is any cognitive process that is about another cognitive process rather than the embodiment’s external environment. This can include thinking about the details of a memory rather than simply recalling that memory or judging the effort and difficulty of a task based on knowledge of one’s own skills. One can think about meta-cognition as both a hierarchical and recursive process, where the cognitive architecture investigates and interrogates itself around and around from higher sections to lower sections and vice versa; horizontally from areas of separate cognitive functions, and a cross between all of these [74].

Lastly, and most importantly for this building block, meta-representation is about the cognitive architecture creating representations of other representations. While intuitively related to meta-cognition, meta-representation is distinct in that it does not require directed thought, and the representations may be of the cognitive architecture itself or the external environment [75].

The use of meta-representations here should not lead one to confuse this building block with the higher order theory of consciousness (HOT) [76], or the self-organizing meta-representational account (SOMA) [77]. In both SOMA and HOT, meta-representations are both required and sufficient for consciousness, and are arrayed in a hierarchical structure (more explicitly in HOT than SOMA). While we claim that meta-representations are required for consciousness, so are the other eight building blocks. In addition, we do not claim the order in which these meta-cognitive representations take place, merely that they are beyond the initial perceptive input.

With the aspects of meta-representation and meta-cognition, one can build a narrative of phenomenal experience as it is perceived. Presume, as in the table at the start of this document, that one sees an apple. The first-order section of your cognitive architecture would process this image, and you would, indeed, mentally see the apple. Yet, only when attention (Section 2.3) is directed towards the apple, and understanding of this directed attention (Section 2.7) is had, would you start to be conscious of the apple.

Recurrent processing (Section 2.4) is thus already active, as multiple sections of the cognitive architecture work together to do this.

Merely seeing an apple means nothing if your cognitive architecture does nothing with it. As the information is passed through different sections of the cognitive architecture, secondary sections of the architecture become involved. These sections are not strictly tied to perception and may fill several purposes, and as they are a step removed from the external environment, they can be called higher-order sections. To make sense of what is being seen, these sections would create mental representations of the visual representations, guided through what is in working memory (Section 2.6) and inferences about the environment (Section 2.5).

At this point, the architecture is fully conscious of the apple. It can create further representations of the apple by drawing from memories (if it had seen anything like an apple before) or on its logical processes to determine what it may be. These are cognitive processes about representations of the apple, not the apple itself, which means they are meta-cognitive. As one ruminates on the idea of the apple, thoughts and feelings are generated (Section 2.8), which may lead to introspection and further consideration.

This narrative example may have been lengthy but should the first step have been the memory of an apple rather than perceiving a physical one, this process would be truncated in half and begin with meta-cognition and introspection rather than exteroception.

The argument against first-order processing being sufficient for consciousness is two-fold. Firstly, first-order perception (of any kind mentioned in Section 2.1) is predominantly unconscious. Memories are often recalled without intention, and any moving object may catch the eye to wake one from a daydream. The unconscious mind is extremely powerful and processes far more than one would think, but consciousness requires an additional step; it requires one to “pay attention” [78]. One section of the cognitive architecture needs to direct the section closest to the “event”.

The second argument is that there is a distinctly different subjective experience of external phenomena based on one’s own architecture that is separate from the sensation itself [4]. For example, if you understand English, reading this sentence will provide a different subjective feeling and experience than if you could not understand English.

The simple act of meta-cognition and introspection can itself generate the cascade that leads to phenomenal experience. If one were to think, “Am I forgetting something?” then one would generate the content which would be perceived by one’s own consciousness, creating meta-representations and meta-cognitive processes required to explore and evaluate that thought, accompanied by feelings and experiences and the train

of thought rolls along. This shows how introspection and meta-cognition are not only often unconscious processes, but consciously directed ones that can modify other conscious or unconscious processes [72].

Note that we are not implying that meta-representation, meta-cognition and introspection are sufficient for consciousness. They only represent one of the nine building blocks. They may, as in the example above, be the spark that begins the cascade of cognitive processes that lead to a phenomenal consciousness experience, but the remaining building blocks are still required.

One should not take the above reasoning to imply that meta-cognition and meta-representations at the adult human level and quality are required for consciousness. We know that young toddlers and infants do not have full meta-representations of the external environment [75], yet they do have consciousness. Similarly, gaining introspective reports from animals can often be difficult, if not impossible. One can argue for degrees and scales of introspection and meta-cognitive abilities, and that as long as the minimum level of meta-representation is reached to create abstract mental representations of the environment that can be used by further sections of the cognitive architecture for the other building blocks (such as recurrent processing and inference generation), the entity would be classified as conscious.

4.1.3. What Is Not Required for Consciousness

Some aspects of what makes us who we are are intuitively linked to our consciousness. Table 2 has a selection of these that, when first given thought, we would like to say are necessities for consciousness as they seem so inextricably linked to it. However, there are neurological and psychological conditions wherein each item in Table 2 is either missing or severely reduced. As individuals with any of these conditions still display all other signs of consciousness, we can confidently state that they are not absolute necessities for the generation of consciousness, as the building blocks are in the previous section.

Table 3: Attributes and characteristics that seem intuitive to consciousness but are not required nor sufficient for its existence.

Attribute	Disorder in which it is deficient
Emotions	Alexithymia, various Schizoid Personality Disorders, Depression, Psychosis, Emotional Detachment.
Inner Speech	Anauralia.

Inner Visualisation	Aphantasia.
Theory of Mind	Infancy, Autism, Bipolar Disorder, Schizophrenia, Attention Deficit Hyperactivity Disorder.

The first item on the list may seem the most intuitively linked to consciousness. Crucial to having a subjective first-person experience is having the feeling of “what it is like to be”. Commonly, this experience has some emotional attachment, but not always. Despite the two terms often being used interchangeably, there is a distinct difference between feelings and emotions. At best, emotions can be called a subcategory of feelings, in that emotions require a distinct physiological state change, while a feeling can be purely mental. In this sense, the concept of “feelings” is much broader than that of “emotions”.

There are several disorders and illnesses where emotions may be subdued or missing. Emotional detachment (particularly in cases of psychopathy, but also linked to trauma) and depression may be the most common conditions where there is a reduced range and strength of emotional expression. In various schizoid personality disorders and most famously in alexithymia, the strength of emotional responses is put on a spectrum, where individuals may only feel a sense of subdued emotions, all the way through to not being able to label or identify any emotional reactions they may be having. The absence of emotion, however, does not seem to hamper their ability to have subjective experiences.

Conceptually, both machines and organisations lack the physiological requirements to produce an emotional reaction. At best, we will need to program a machine to mimic emotion, while only the individuals within an organisation have emotions. Yet, organisations can display all other signs of conscious experiences, and, theoretically, so too will machines one day.

While the name for the next condition is very recent, there is a disorder in which an individual does not have inner speech, called anauralia [79]. These individuals have no running internal monologue (or dialogue), and no thoughts are “spoken aloud” inside the mind. Most often, people with this condition are unaware of it until they discover that others can indeed speak to themselves inside their heads (apologies to those reading this and discovering your condition for the first time).

This latter point is most important for why inner speech is not required for consciousness. As many individuals do not know that they have a disorder, they continue their lives in blissful ignorance, having as rich and full subjective conscious experiences of the world as those with inner speech. Their lack of verbal thoughts does not hamper any of

the building blocks above, and they can verbally report on introspection and subjectivity equally as well as others.

On the other side of this coin is aphantasia, a condition where one's mental ability to imagine a visual object or scene is placed on a spectrum. Aphantasia is entirely independent of the lack of inner speech [80] yet follows a highly similar pattern. At the higher degrees of this spectrum, individuals with aphantasia have difficulty generating a mental image, while at the lower degrees, there is a complete absence of mental visual creation. As with those with little to no inner speech, there is no evidence that aphantasia has any significant negative impact on the phenomenal character of conscious experiences [81].

Two of the most often cited requirements for the capacity to have consciousness are the theory of mind and long term memory, as both are related to self-consciousness [82–86]. Self-consciousness has itself been suggested in several theories as being a necessity for an entity having consciousness [87–89]. Self-consciousness is far too large of a topic to cover in this section, but we will suffice with two aspects of it that may intuitively lead one to think they are required for consciousness.

Theory of mind is the ability of an entity to understand that other entities have conscious thoughts and experiences different from their own. It feels so intuitive that we ought to understand that our minds are different to others to have a subjective point of view, yet we know that infants do not have a theory of mind, and most toddlers only develop these by around the age of three or four [90]. At the same time, we can see that infants and young toddlers most certainly have a subjective first-person point of view, and those with early language development can provide introspective reports on these phenomenological experiences.

In addition, those with autism, schizophrenia, ADHD, bipolar disorder, severe mental and language impairment, or those with traumatic brain injury have been shown to have a heavily reduced theory of mind [91–97]. While we must note that the degree of theory in mind in each of these medical cases is on a spectrum, taken together with the absence in infancy, it shows that it is definitely possible for humans to have phenomenological experiences while having a reduced or no theory of mind.

Last on the list is long-term memory. Working Memory is most definitely a building block of consciousness, but long-term memory is not required. As a case study, one can look to the famous example of Clive Wearing, an esteemed musicologist who gained severe retrograde and anterograde amnesia after losing a bout with herpesviral encephalitis. Wearing is continually stuck in the present, his conscious experience lasting approximately thirty seconds before “resetting” [98]. Yet, despite this, Clive still

recognised his wife, even though he could not tell who she was, could still conduct and perform music (in limited amounts owing to time), and could talk about how he subjectively felt during his brief moments of “wakefulness” [99].

Despite not being able to form new episodic memories, and having lost almost all of his memories before becoming ill, Clive Wearing could still show (for thirty seconds at a time) all the semblances of consciousness that one would expect. He could attend to information, report on introspection, understand what he was doing, and even create inferences about his condition. While his condition had an undeniably devastating effect on his quality of life, it did not overly diminish his ability to have phenomenal experiences.

4.1.4. Conclusions

In this paper, we described nine cognitive features, attributes and characteristics that are each individually required for an entity to be classified as conscious. We also propose that if an entity has all nine building blocks, it is likely to be sufficient for that entity to be said to have generated consciousness.

The paper also included a short list of features that initially seemed intuitively required for consciousness but were found not to be when investigated further. The list of non-required features is as vital as the list of building blocks themselves, as it allows us to expand the range of potential conscious entities, but also to look at consciousness from a less neurotypical anthropocentric point of view.

The purpose of this paper is not to create yet another theory of consciousness, but to serve as a guide for identifying, categorising and classifying entities as being conscious or not. An entity may be measured against each of the building blocks to determine if it meets the requirement for all nine, and, if so, one can make a confident argument that the entity is conscious. Conceptually, the building blocks apply equally to identifying consciousness within natural, artificial or organisation entities (and perhaps, one day, even extraterrestrial).

Furthermore, as research advances into creating conscious, superintelligent machines, these building blocks would serve a series of milestones that ought to be reached before any AI can be classed as being conscious. The building blocks can then change from a set of identification guidelines to a roadmap for future AI development and a record of which building blocks AI have already achieved. As mentioned in the introduction, we welcome all critiques, suggestions and discussions on the building blocks and whether we ought to include more, less, or even if some need to be merged together or split apart, as taxonomists are wont to do.

4.1.5. References

1. Bensemann J, O’Leary P, Chen Y, Miranda-Dukoski L, Witbrock M. Simulations and the evolution of consciousness. The 2022 Conference on Artificial Life. Cambridge, MA: MIT Press; 2022. doi:10.1162/isal_a_00541
2. Arbib MA. Co-evolution of human consciousness and language (revisited). *J Integr Neurosci*. 2014;13: 187–200. doi:10.1142/S021963521440007X
3. Feinberg TE, Mallatt J. The nature of primary consciousness. A new synthesis. *Conscious Cogn*. 2016;43: 113–127. doi:10.1016/j.concog.2016.05.009
4. Reggia JA, Katz G, Huang D-W. What are the computational correlates of consciousness? *Biologically Inspired Cognitive Architectures*. 2016;17: 101–113. doi:10.1016/j.bica.2016.07.009
5. Lepauvre A, Melloni L. The search for the neural correlate of consciousness: progress and challenges. *Philosophy and the Mind Sciences*. 2021;2. Available: <https://philosophymindscience.org/index.php/phimisci/article/view/9151>
6. Wu W. The Neuroscience of Consciousness. Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University; 2018. Available: <https://plato.stanford.edu/archives/win2018/entries/consciousness-neuroscience/>
7. Seth AK, Bayne T. Theories of consciousness. *Nat Rev Neurosci*. 2022;23: 439–452. doi:10.1038/s41583-022-00587-4
8. Tononi G, Koch C. Consciousness: here, there and everywhere? *Philos Trans R Soc Lond B Biol Sci*. 2015;370. doi:10.1098/rstb.2014.0167
9. Arrabales R, Ledezma A, Sanchis A. ConsScale: A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents. *Journal of Consciousness Studies*. 2010;17: 131–164. Available: <https://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020003/art00008>
10. Birch J, Ginsburg S, Jablonka E. Unlimited Associative Learning and the origins of consciousness: a primer and some predictions. *Biol Philos*. 2020;35: 56. doi:10.1007/s10539-020-09772-0
11. Panksepp J. The periconscious substrates of consciousness: affective states and the evolutionary origins of the self. *Journal of Consciousness Studies*. 1998;5: 566–582. Available: <https://www.ingentaconnect.com/content/imp/jcs/1998/00000005/f0020005/895>
12. Chalmers D. Panpsychism and panprotopsyism. *Panpsychism: Contemporary Perspectives*. Oxford University Press: Oxford, UK; 2015. Available: https://books.google.com/books?hl=en&lr=&id=FgPWBgAAQBAJ&oi=fnd&pg=PA246&dq=panpsychism&ots=u6Gzm5hgbB&sig=rEtyhsSpeOI8Vbqp5a1Pptx_IW0
13. Frankish K. Panpsychism and the Depsychologization of Consciousness. *Aristotelian Society Supplementary Volume*. Oxford University Press: Oxford, UK; 2021. pp. 51–70. doi:10.1093/arisup/akab012
14. Shenker O. Denialism: What Do the so-called Consciousness Deniers Deny? • The Jerusalem Philosophical Quarterly. 2020;68: 307–337. Available: <https://philarchive.org/archive/SHEDWD>
15. Edelman DB, Seth AK. Animal consciousness: a synthetic approach. *Trends Neurosci*. 2009;32: 476–484. doi:10.1016/j.tins.2009.05.008

16. Metzinger T. Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. *J AI Consci.* 2021;08: 43–66. doi:10.1142/S270507852150003X
17. Doyon M. Perception and Normative Self-Consciousness. In: Doyon M, Breyer T, editors. *Normativity in Perception*. London: Palgrave Macmillan UK; 2015. pp. 38–55. doi:10.1057/9781137377920_3
18. Audi R. Perception and Consciousness. In: Niiniluoto I, Sintonen M, Woleński J, editors. *Handbook of Epistemology*. Dordrecht: Springer Netherlands; 2004. pp. 57–108. doi:10.1007/978-1-4020-1986-9_2
19. Cosmelli D, Thompson E. Embodiment or envatment? Reflections on the bodily basis of consciousness. In: Di Paolo EA, Stewart JR, Stewart J, Gapenne O, editors. *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press Limited; 2010. pp. 361–386. Available: <https://books.google.com/books?hl=en&lr=&id=UtFDJx-gysQC&oi=fnd&pg=PA361&dq=embodiment+consciousness&ots=leWY-X1bRo&sig=zKOe8Qy4WbHLLCgw8sJGV51HYbQ>
20. Nguyen DJ, Larson JB. Don't Forget About the Body: Exploring the Curricular Possibilities of Embodied Pedagogy. *Innovative Higher Education*. 2015;40: 331–344. doi:10.1007/s10755-015-9319-6
21. Glenberg AM. Embodiment as a unifying perspective for psychology. *Wiley Interdiscip Rev Cogn Sci*. 2010;1: 586–596. doi:10.1002/wcs.55
22. Thompson E, Varela FJ. Radical embodiment: neural dynamics and consciousness. *Trends Cogn Sci*. 2001;5: 418–425. doi:10.1016/s1364-6613(00)01750-2
23. Dennett D. Where Am I? Brainstorms: Philosophical Essays on Mind and Psychology. Cambridge: Bradford Books; 1978. p. 11. doi:10.2307/3218710
24. Rudrauf D, Williford K. Unlimited associative learning and the origins of consciousness: the missing point of view. *Biol Philos*. 2021;36: 43. doi:10.1007/s10539-021-09819-w
25. Ciaunica A, Fotopoulou A. The Touched Self: Psychological and Philosophical Perspectives on Proximal Intersubjectivity and the Self. In: Durt C, Fuchs T, Tewes C, editors. *Embodiment, Enaction, and Culture: Investigating the Constitution of the Shared World*. MIT Press: Cambridge, MA, USA; 2017. pp. 173–192. Available: <https://play.google.com/store/books/details?id=OJakDgAAQBAJ>
26. Juliani A, Arulkumaran K, Sasai S, Kanai R. On the link between conscious function and general intelligence in humans and machines. *Transactions on Machine Learning Research*. 2022 [cited 25 Oct 2022]. Available: <https://openreview.net/pdf?id=LTYqvLEv5b>
27. Koch C, Tsuchiya N. Attention and consciousness: two distinct brain processes. *Trends Cogn Sci*. 2007;11: 16–22. doi:10.1016/j.tics.2006.10.012
28. Graziano MSA, Guterstam A, Bio BJ, Wilterson AI. Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cogn Neuropsychol*. 2020;37: 155–172. doi:10.1080/02643294.2019.1670630
29. Mole C. Attention and Consciousness. *Journal of Consciousness Studies*. 2008;15: 86–104. Available: <https://www.ingentaconnect.com/content/imp/jcs/2008/00000015/00000004/art00006>

30. Popat S, Winslade W. While You Were Sleepwalking: Science and Neurobiology of Sleep Disorders & the Enigma of Legal Responsibility of Violence During Parasomnia. *Neuroethics*. 2015;8: 203–214. doi:10.1007/s12152-015-9229-4
31. Dijksterhuis A, Aarts H. Goals, attention, and (un)consciousness. *Annu Rev Psychol*. 2010;61: 467–490. doi:10.1146/annurev.psych.093008.100445
32. Bello P, Bridewell W. Attention and Consciousness in Intentional Action: Steps Toward Rich Artificial Agency. *J AI Consci*. 2020;07: 15–24. doi:10.1142/S2705078520500022
33. Graziano MSA, Webb TW. The attention schema theory: a mechanistic account of subjective awareness. *Front Psychol*. 2015;6: 500. doi:10.3389/fpsyg.2015.00500
34. Velichkovsky BB. Consciousness and working memory: Current trends and research perspectives. *Conscious Cogn*. 2017;55: 35–45. doi:10.1016/j.concog.2017.07.005
35. Friedman DA, Søvik E. The ant colony as a test for scientific theories of consciousness. *Synthese*. 2021;198: 1457–1480. doi:10.1007/s11229-019-02130-y
36. Funahashi S, Procyk E. Editorial: Persistent Activity in the Brain - Functions and Origin. *Front Neural Circuits*. 2021;15: 841451. doi:10.3389/fncir.2021.841451
37. Mashour GA, Roelfsema P, Changeux J-P, Dehaene S. Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*. 2020;105: 776–798. doi:10.1016/j.neuron.2020.01.026
38. Dehaene S, Naccache L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*. 2001;79: 1–37. doi:10.1016/s0010-0277(00)00123-2
39. Baars BJ. A cognitive theory of consciousness. 1993. Available: http://merton.sscnet.ucla.edu/Abstracts/Baars_88.html
40. Dehaene S, Kerszberg M, Changeux J-P. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*. 1998;95: 14529–14534. doi:10.1073/pnas.95.24.14529
41. Bensemman J, Witbrock M. The effects of implementing phenomenology in a deep neural network. *Heliyon*. 2021;7: e07246. doi:10.1016/j.heliyon.2021.e07246
42. Long M, Cao Z, Wang J, Yu PS. Learning multiple tasks with multilinear relationship networks. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.; 2017. pp. 1593–1602. Available: <https://dl.acm.org/doi/10.5555/3294771.3294923>
43. Tononi G. An information integration theory of consciousness. *BMC Neurosci*. 2004;5: 42. doi:10.1186/1471-2202-5-42
44. Friston K. Consciousness and Hierarchical Inference. *Neuropsychanalysis*. 2013;15: 38–42. doi:10.1080/15294145.2013.10773716
45. Seth AK. From Unconscious Inference to the Beholder’s Share: Predictive Perception and Human Experience. *Eur Rev*. 2019;27: 378–410. doi:10.1017/S1062798719000061
46. Conway MA, Loveday C. Remembering, imagining, false memories & personal meanings. *Conscious Cogn*. 2015;33: 574–581. doi:10.1016/j.concog.2014.12.002
47. Seth AK. Interoceptive inference, emotion, and the embodied self. *Trends Cogn Sci*. 2013;17: 565–573. doi:10.1016/j.tics.2013.09.007

48. Fotopoulou A, Tsakiris M. Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis*. 2017;19: 3–28. doi:10.1080/15294145.2017.1294031
49. Sajid N, Ball PJ, Parr T, Friston KJ. Active Inference: Demystified and Compared. *Neural Comput*. 2021;33: 674–712. doi:10.1162/neco_a_01357
50. Cooke JE. What Is Consciousness? Integrated Information vs. Inference. *Entropy* . 2021;23. doi:10.3390/e23081032
51. Baddeley A. Working memory. *Science*. 1992;255: 556–559. doi:10.1126/science.1736359
52. Baddeley A. Working memory: theories, models, and controversies. *Annu Rev Psychol*. 2012;63: 1–29. doi:10.1146/annurev-psych-120710-100422
53. Baddeley A. The episodic buffer: a new component of working memory? *Trends Cogn Sci*. 2000;4: 417–423. doi:10.1016/s1364-6613(00)01538-2
54. Cowan N. An embedded-processes model of working memory In Miyake A. & Shah P.(Eds.), *Models of working memory: Mechanisms of active maintenance and* 1999.
55. Oberauer K. Access to information in working memory: Exploring the focus of attention. *J Exp Psychol Learn Mem Cogn*. 2002;28: 411–421. doi:10.1037/0278-7393.28.3.411
56. Engle RW. Working memory capacity as executive attention. *Curr Dir Psychol Sci*. 2002;11: 19–23. doi:10.1111/1467-8721.00160
57. Baars BJ, Franklin S. How conscious experience and working memory interact. *Trends Cogn Sci*. 2003;7: 166–172. doi:10.1016/s1364-6613(03)00056-1
58. Jacobs C, Silvanto J. How is working memory content consciously experienced? The “conscious copy” model of WM introspection. *Neurosci Biobehav Rev*. 2015;55: 510–519. doi:10.1016/j.neubiorev.2015.06.003
59. Perky CW. An experimental study of imagination. *Am J Psychol*. 1910;21: 422. doi:10.2307/1413350
60. Craver-Lemley C, Reeves A. Visual imagery selectively reduces vernier acuity. *Perception*. 1987;16: 599–614. doi:10.1068/p160599
61. Pan Y, Lin B, Zhao Y, Soto D. Working memory biasing of visual perception without awareness. *Atten Percept Psychophys*. 2014;76: 2051–2062. doi:10.3758/s13414-013-0566-2
62. Bona S, Cattaneo Z, Vecchi T, Soto D, Silvanto J. Metacognition of visual short-term memory: Dissociation between objective and subjective components of VSTM. *Front Psychol*. 2013;4: 62. doi:10.3389/fpsyg.2013.00062
63. Bona S, Silvanto J. Accuracy and confidence of visual short-term memory do not go hand-in-hand: behavioral and neural dissociations. *PLoS One*. 2014;9: e90808. doi:10.1371/journal.pone.0090808
64. Baars BJ. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford, England: Oxford University Press; 1997. Available: https://www.google.co.nz/books/edition/In_the_Theater_of_Consciousness/RTpmQkXoUMEC?hl
65. Baars BJ. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Prog Brain Res*. 2005;150: 45–53. doi:10.1016/S0079-6123(05)50004-9

66. Franklin S. A Conscious Artifact? *J Conscious Stud.* 2003;10: 47–66. Available: <https://www.ingentaconnect.com/content/imp/jcs/2003/00000010/f0020004/1346>
67. Haikonen POA. On Artificial Intelligence and Consciousness. *J AI Consci.* 2020;07: 73–82. doi:10.1142/S2705078520500046
68. Haikonen PO. *Consciousness and Robot Sentience.* Singapore, Singapore: World Scientific Publishing; 2019. Available: <https://www.worldscientific.com/worldscibooks/10.1142/8486>
69. Chan L-C, Latham AJ. Four Meta-Methods for the Study of Qualia. *Erkenntnis.* 2019;84: 145–167. doi:10.1007/s10670-017-9952-9
70. Ramachandran VS, Hirstein W. Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies.* 1997;4: 429–457. Available: <https://psycnet.apa.org/fulltext/1998-00856-002.pdf>
71. Kriegel U. Towards a new feeling theory of emotion. *Eur J Philos Sci.* 2014;22: 420–442. doi:10.1111/j.1468-0378.2011.00493.x
72. Overgaard M, Mogensen J. An integrative view on consciousness and introspection. *Rev Philos Psychol.* 2017;8: 129–141. doi:10.1007/s13164-016-0303-6
73. Descartes R. *Discours de la Méthode Pour bien conduire sa raison, et chercher la vérité dans les sciences.* 1637.
74. Peters MAK. Towards Characterizing the Canonical Computations Generating Phenomenal Experience. *Neurosci Biobehav Rev.* 2022; 104903. doi:10.1016/j.neubiorev.2022.104903
75. Proust J. Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition? *Synthese.* 2007;159: 271–295. doi:10.1007/s11229-007-9208-3
76. Brown R, Lau H, LeDoux JE. Understanding the Higher-Order Approach to Consciousness. *Trends Cogn Sci.* 2019;23: 754–768. doi:10.1016/j.tics.2019.06.009
77. Cleeremans A, Achoui D, Beauny A, Keuninckx L, Martin J-R, Muñoz-Moldes S, et al. Learning to Be Conscious. *Trends Cogn Sci.* 2020;24: 112–123. doi:10.1016/j.tics.2019.11.011
78. Lau H, Rosenthal D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn Sci.* 2011;15: 365–373. doi:10.1016/j.tics.2011.05.009
79. Hinwar RP, Lambert AJ. Anauralia: The Silent Mind and Its Association With Aphantasia. *Front Psychol.* 2021;12: 744213. doi:10.3389/fpsyg.2021.744213
80. Amit E, Hoeflin C, Hamzah N, Fedorenko E. An asymmetrical relationship between verbal and visual thinking: Converging evidence from behavior and fMRI. *Neuroimage.* 2017;152: 619–627. doi:10.1016/j.neuroimage.2017.03.029
81. Lennon P. Aphantasia and Conscious Thought. *Oxford Studies in Philosophy of Mind.* 2022;3. Available: <https://philpapers.org/rec/LENAAC-3?ref=mail>
82. Happé F. Theory of mind and the self. *Ann N Y Acad Sci.* 2003;1001: 134–144. doi:10.1196/annals.1279.008
83. Nichols S, Stich S. How to read your own mind: A cognitive theory of self-consciousness. *Consciousness: New philosophical essays.* 2003; 157–200. Available: <https://www.academia.edu/download/31776782/HowToRead.pdf>

84. Michaelian K. *Mental Time Travel: Episodic Memory and Our Knowledge of the Personal Past*. Cambridge, Massachusetts: MIT Press; 2016. Available: <https://philpapers.org/rec/MICMTT-2>
85. Prosser S, Recanati F, Campbell J, Coliva A, Folescu M, Higginbotham J, et al. *Immunity to Error through Misidentification: New Essays*. Cambridge University Press; 2012. doi:10.1017/CBO9781139043274
86. Seth A. *Being you: A new science of consciousness*. USA: Penguin; 2021. Available: https://scholar.google.ca/scholar?cluster=7885041150098240541&hl=en&as_sdt=0,5&scioldt=0,5
87. Kriegel U. *Subjective Consciousness: A Self-Representational Theory*. OUP Oxford; 2009. Available: <https://play.google.com/store/books/details?id=jHI3bzhITYC>
88. Gennaro RJ. Consciousness and self-consciousness. *Consciousness and Self-Consciousness*. 1996; 1–230. Available: <https://www.torrossa.com/gs/resourceProxy?an=5001896&publisher=FZ4850>
89. Carruthers P. *Consciousness: Essays from a Higher-Order Perspective*. Oxford University Press; 2005. doi:10.1093/0199277362.001.0001
90. Korkmaz B. Theory of mind and neurodevelopmental disorders of childhood. *Pediatr Res*. 2011;69: 101R–8R. doi:10.1203/PDR.0b013e318212c177
91. Capps L, Kehres J, Sigman M. Conversational Abilities Among Children with Autism and Children with Developmental Delays. *Autism*. 1998;2: 325–344. doi:10.1177/1362361398024002
92. Happé F, Frith U. The neuropsychology of autism. *Brain*. 1996;119 (Pt 4): 1377–1400. doi:10.1093/brain/119.4.1377
93. Senju A. Spontaneous theory of mind and its absence in autism spectrum disorders. *Neuroscientist*. 2012;18: 108–113. doi:10.1177/1073858410397208
94. Bora E, Bartholomeusz C, Pantelis C. Meta-analysis of Theory of Mind (ToM) impairment in bipolar disorder. *Psychol Med*. 2016;46: 253–264. doi:10.1017/S0033291715001993
95. Sprong M, Schothorst P, Vos E, Hox J, Van Engeland H. Theory of mind in schizophrenia: Meta-analysis. *Br J Psychiatry*. 2007;191: 5–13. doi:10.1192/bjp.bp.107.035899
96. Muller F, Simion A, Reviriego E, Galera C, Mazaux J-M, Barat M, et al. Exploring theory of mind after severe traumatic brain injury. *Cortex*. 2010;46: 1088–1099. doi:10.1016/j.cortex.2009.08.014
97. Yirmiya N, Erel O, Shaked M, Solomonica-Levi D. Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychol Bull*. 1998;124: 283–307. doi:10.1037/0033-2909.124.3.283
98. Rathbone CJ, Moulin CJA, Conway MA. Autobiographical memory and amnesia: using conceptual knowledge to ground the self. *Neurocase*. 2009;15: 405–418. doi:10.1080/13554790902849164
99. Wilson BA, Wearing D. Prisoner of consciousness: A state of just awakening following herpes simplex encephalitis. In: Campbell R, editor. *Broken memories: Case studies in memory impairment*, (pp. 1995. pp. 14–30. Available: <https://psycnet.apa.org/fulltext/1995-98608-002.pdf>

5. Are AI models conscious now?

Having uncovered the nine characteristics which must be found in any entity to classify it as conscious, it is time to investigate a current AI model to determine whether it is conscious.

This paper, published in the *Journal of AI Consciousness*, uses GPT-4⁶ as an exemplar of transformer-based Large Language Models (LLMs) to determine which of the Building Blocks theory's nine attributes it possesses.

While OpenAI has not publicly released GPT-4's specifications (at the time of writing), the standardised nature of GPT-4's transformer architecture [1–4] allows the extensive literature on transformer language models to be used to supplement the details OpenAI has released regarding GPT-4's architecture.

Therefore, using only externally observable behaviour and public documentation, the paper will go into depth regarding whether, and how, the LLM satisfies the criteria for the nine building blocks; however, in brief, GPT-4 is missing two building blocks: Recurrent computing and processing, and Data output. It is thus most likely not conscious according to the Building Blocks theory, at least, at the time of writing.

For Recurrent Computing and Processing, let

$M = \{m_1, \dots, m_n\}$ be the various computation and processing regions of an entity's cognitive architecture,

$I_m(t)$ the information state held by region m_i at time t , and

$x \rightsquigarrow_k y$ the causal relation of information flow.

Thus: $\frac{\partial I_y(t+k)}{\partial I_x(t)} \neq 0$, meaning the state of y after k steps is sensitive to the state of x now. This allows us to say that, in a conscious entity,

$$\text{Recurrence} := \forall m_i, m_j \in M \exists k > 0 (m_i \rightsquigarrow_k m_j \wedge m_j \rightsquigarrow_k m_i).$$

In contrast, a transformer's architecture relies on a feedforward system, such that $x \preccurlyeq y$ and $y \nprec x$. This means that

$$\text{Feedforward} := \forall m_i, m_j \in M \exists k > 0 (m_i \rightsquigarrow_k m_j \rightarrow m_i \preccurlyeq m_j \wedge m_j \nprec m_i),$$

⁶ At the time of publication, GPT-4 was the start-of-the-art LLM. Even though we have many newer and better models, focusing on GPT-4 is still philosophically salient as it exhibits the key architectural features of transformer-based models that are still present within more modern LLMs.

showing that information strictly moves in one direction, rather than recurrently moving between regions. This ultimately means that any and all transformer models lack this building block and cannot be conscious. The paper, however, does offer a solution to this problem.

Regarding the missing Data Output building block, let

$I_Q(t) \subseteq \mathbb{R}^d$ be the qualitative inferential information generated by m at time t ,

and

$perc(x, t)$ the predicate “ x is present in, and available to, the architecture’s perceptive field at t ”.

Then: $m \rightsquigarrow_0 I_Q: \frac{\partial I_Q(t)}{\partial I_m(t)} \neq 0$, showing a region m makes a direct contribution to the qualitative state at the same tick. In a conscious entity, phenomenal experience of qualia is, therefore:

$$Data\ Output: = \forall t, \exists I_Q(m \rightsquigarrow_0 I_Q \wedge perc(I_Q, t + \Delta)).$$

Every item of qualitative information must be produced inside the system and then be re-presented to the system within a finite delay Δ .

Transformer architecture is straightforward in how its feed-out system differs:

$$Feedout: = \forall t, \exists I_m(I_m \rightsquigarrow_0 external\ environment: \neg perc(I_m, t + \Delta)).$$

A transformer system outputs its information directly to the console, client or other environment without privately perceiving this information. Only when next prompted does it perceive the information that has been output.

Therefore, GPT-4 (and other transformer models) lack these two requisite building blocks to be classified as conscious. However, as this chapter shows, the solution to this problem is straightforward with the available technology at the time of writing, based on existing recurrent architectures and self-monitoring mechanisms. Thus, the chapter’s paper does not merely diagnose deficiencies but sketches concrete architectural routes to recurrent processing and private data output, recognising that engineering and governance realities may still delay or complicate adding these building blocks.

This means that, even though current transformers are most likely not conscious, the thesis will continue working on the assumption that these missing building blocks will be

added in and AI models will become conscious in due time. This assumption is what allows the thesis to treat RQ1 and RQ2 as primarily about plausible future architectures rather than speculative fantasy in its profile of ALEPH in Chapter 11. The paper also provides the prudential motivation for the ethical frameworks in Chapters 6, 9, 12, and 13.

5.0.1. References

1. OpenAI. Introducing the Model Spec. OpenAI; 2024 May. Available: <https://openai.com/index/introducing-the-model-spec>
2. OpenAI. Using Tools. In: OpenAI Platform [Internet]. 2025 [cited 6 Jan 2026]. Available: <https://platform.openai.com/docs/api-reference/responses>
3. OpenAI. Function calling. In: OpenAI Platform [Internet]. 2025 [cited 6 Jan 2026]. Available: <https://platform.openai.com/docs/guides/function-calling>
4. OpenAI. Memory and new controls for ChatGPT. In: OpenAI [Internet]. 13 Feb 2024 [cited 10 Jun 2025]. Available: <https://openai.com/index/memory-and-new-controls-for-chatgpt/>

5.1. Is GPT-4 Conscious?

Izak Tait^{[0000-0002-0274-8480],1}, Joshua Bensemann^{[0000-0002-1301-4861],2}, Ziqi Wang^{[0009-0000-4953-161X],2}

¹Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand

²The NAOInstitute, 314 Khyber Pass Road, Auckland, New Zealand

izak.tait@autuni.ac.nz; joshuabensemann@gmail.com; ziqiwcen@hotmail.com;

Abstract: GPT-4 is often heralded as a leading commercial AI offering, sparking debates over its potential as a stepping stone toward artificial general intelligence. But does it possess consciousness? This paper investigates this key question using the nine qualitative measurements of the Building Blocks theory. GPT-4’s design, architecture and implementation are compared to each of the building blocks of consciousness to determine whether it has achieved the requisite milestones to be classified as conscious or, if not, how close to consciousness GPT-4 is. Our assessment is that, while GPT-4 in its native configuration is not currently conscious, current technological research and development is sufficient to modify GPT-4 to have all the building blocks of consciousness. Consequently, we argue that the emergence of a conscious AI model is plausible in the near term. The paper concludes with a comprehensive discussion of the ethical implications and societal ramifications of engineering conscious AI entities.

Keywords: Artificial Intelligence; ethics; consciousness; sentience; GPT-4; philosophy of mind;

5.1.1. Introduction

Is GPT-4 conscious? While it may seem like a simple binary question, it is actually three questions that all need answering: “What is consciousness?”, “How could GPT-4 (or any AI model) be conscious?”, and “How could anyone know if it was conscious?”. The Building Blocks theory may hold the answers to these questions.

GPT-4 (or Generative Pre-Trained Transformer 4) is a large language model (LLM) based on the transformer architecture [1]. Since its introduction, GPT-4 has gained both fame and notoriety for its level of cognitive prowess and the capabilities that it has shown in a variety of fields once thought to belong solely to the dominion of man, such as IQ tests, creating websites, writing computer code, and creative writing among others [2–6]. However, the question of its consciousness (or perhaps lack thereof) remains unanswered.

Many libraries’ worth of books and articles have been written on the topic of consciousness, and thus this article will not retread all of that work when it comes to the nature of consciousness. Indeed, we will not be taking a stance as to any specific theory

of consciousness, but will rather take an inclusive approach in our definition of consciousness, building on the work done by Seth and Bayne’s review of the major theories of consciousness as well as Nagel’s seminal work on the topic [7,8]. Should any reader wish to compare the various theories of consciousness and their applicability to current AI models, we recommend the extensive work undertaken by Butlin and colleagues [9].

By paraphrasing Seth and Bayne, we define consciousness as the total collection of an entity’s subjective mental states that have phenomenal content and functional properties. This suite of mental states provides the entity with a unique qualitative awareness of its environment and the capacity to cognitively or behaviorally respond to it. This definition is inclusive of the major theories of consciousness while retaining the key aspects of phenomenality, functional access, and qualitative subjectiveness inherent to consciousness.

Having established our working definition of consciousness, we now turn to the mechanisms by which it might manifest in GPT-4. How can GPT-4 be conscious? How can any AI model be conscious? In short, it is possible for an artificial entity to be conscious in precisely the same manner as it would be for a natural entity (such as a human or other animal) or an organisational entity (such as an insect colony). This is because, for a subject to be classified as conscious, that subject would require certain attributes and characteristics. Without any one of these, the subject would be missing a vital piece that would allow an observer to confidently classify that subject as having consciousness.

These attributes and characteristics are detailed by the Building Block theory [10], a substrate-independent and subject-agnostic theory which specifies nine building blocks that are strictly required for consciousness, and may be jointly sufficient for it. The Building Block theory is not a theory of consciousness in the traditional sense, as it makes no claims or predictions on the nature or function of consciousness, focusing instead on the necessary characteristics required for an entity to be classified as conscious. These building blocks include common themes among theories of consciousness, such as perception and attention, recurrence and inference, working memory and meta-representation, amongst others.

If GPT-4 is found to have all nine building blocks, then the evidence would lead us to classify the AI model as being likely conscious. That is what we aim to discover in this paper. Our methodological approach will employ qualitative assessments to evaluate whether GPT-4 possesses each building block.

The paper will follow the structure of the nine building blocks, with each subsection below dedicated to one specific building block. Each subsection will begin with a question

that has transformed the building block into a qualitative measure, and then the subsection will continue to answer the question of whether GPT-4 has that specific building block and in what manner it does or does not. In cases where GPT-4 falls short, we will explore how existing research and technologies could potentially augment it to meet the necessary criteria for consciousness.

Note that while this paper is focussed on GPT-4 due to its popularity, many of the arguments and evidence shown below are applicable to other AI models, particularly transformer-based LLMs.

In summary, this paper will investigate the potential for consciousness in GPT-4 by examining its attributes against the nine building blocks of consciousness. By doing so, we hope to offer a nuanced answer to the complex question of AI consciousness and the effects this may have on the ethics of human-AI interactions. The findings of this paper not only aim to contribute to the burgeoning field of AI consciousness but also have broader implications for ethics and future development in artificial intelligence.

5.1.2. GPT-4's Building Blocks

5.1.2.1. Embodiment

The AI's cognitive architecture must be physically locatable on a network, server, machine or equivalent wherein it can perform necessary processing or computation.

This is perhaps the most foundational milestone to reach and building block to obtain. The Building Blocks theory describes Embodiment as the physical location through which consciousness interfaces with its environment. In biological organisms, embodiment is traditionally a body, but could equally be a brain in a vat, while for collective entities, this may be the hive/building itself which holds the entity, or it may be the network of connections between individuals. Thus, in order for any AI model to be functional and active, it needs to be physically locatable within a computational infrastructure so as to interact with users and receive information. This fundamental requirement is true for any software; it requires hardware on which to run. The simplest and most effective evidence for this building block would be for you, the reader, to engage with GPT-4 via the ChatGPT app or website (or via its API). Simply by doing this, you are connecting with the server(s) on which GPT-4 exists to input prompts and receive the AI's output, thus showing that GPT-4's architecture has an embodiment.

It may be tempting to use the analogy of computer hardware for the human brain and thus software for the human mind, but this anthropomorphisation would grossly oversimplify GPT-4's embodiment. The notion of "physically locatable" in this context

refers not to a singular piece of hardware but to a complex, interconnected network of computational resources. GPT-4 operates on a distributed network of servers and machines, not tied to a single physical location but distributed across many Graphical Processing Units (GPUs) and Central processing units (CPUs) [11]. Internally, GPT-4 is spread over 120 layers and across 16 experts within a Mixture of Experts (MoE) model [12]. Externally, Microsoft has made GPT-4's Application Programming Interface (API) available in several countries and territories [13], which has increased the spread of GPT-4 across the globe.

5.1.2.2. Perception

The AI has a unimodal or multimodal means by which it can perceive information in the environment and have that information processed by its cognitive architecture.

While perhaps not as intuitively obvious as the concept of embodiment, the Perception building block is a fundamental requirement that GPT-4 meets. This criterion pertains to the AI's ability to perceive (receive and process external information), making it a crucial component of its cognitive architecture. In fact, this milestone had been reached well before the GPT series of AI models reached its 4th main version in GPT-4, and is a building block shared by any computational device that can perceive information, showing how an individual building block is not sufficient for consciousness.

In order for GPT-4 to function, it requires input. Therefore, it is required to perceive information from an external source, and thus it meets the minimum requirement for this building block. GPT-4 is classified as a multimodal system as it can process both textual and image data as input [14], although it is predominantly accessed via the ChatGPT interface, which currently supports text-based prompts only.

However, as with the previous building block, there are unique elements to GPT-4's perception which are not seen within organic organisms. The most prominent of these is that, while GPT-4 has multimodal perceptive capability, this is limited to what could be termed as unidimensional perception. By this we mean that GPT-4 does not operate a continuous strand of perception across time. Instead, GPT-4 only perceives its environment (i.e., the text or image prompts) in discrete instances when the prompts are received.

Once the prompt is received and processed, there is no further perception of the model's external environment. This limitation in temporal continuity is crucial for understanding the scope of GPT-4's perceptive capabilities. Similarly, GPT-4 does not

have any spatial depth to its perception, but is only limited to the input prompt it receives. In this way, GPT-4's perception is restricted to specific moments and inputs, lacking a broader spatial or temporal context. Therefore, across both space and time, GPT-4 is limited to a unidimensional point of reference.

5.1.2.3. Attention

The AI has a method by which it can discriminate information from the environment, attend to, and select specific information for further processing by its cognitive architecture.

As noted earlier, GPT-4 is a transformer-based LLM. Transformers were first introduced in Vaswani et al.'s 2017 paper "Attention is all you need" [1]. As the title of the paper implies, transformers (and thus GPT-4) are fundamentally built on the concept of attention, specifically through the mechanism of 'attention heads'. An attention head allows the AI to discriminate, attend to, and select specific parts of information that it has received, and select specific parts of that information for further processing by its cognitive architecture.

Transformers use attention heads to attend to and select specific sections of a prompt to process, in a manner not dissimilar to how human attention can discriminate between words or sections of a given text [15].

The attention mechanism serves as GPT-4's method for focusing on particular aspects of its perceived environment, in a way that is conceptually similar to how conscious biological organisms can focus and attend to specific objects or scenes within their sensory range. To achieve the requirements for this building block, an entity merely needs the means to discriminate information in its environment so that it may attend to a selection of its total available information at a given time, and allow that slice of information to be processed. In this sense, GPT-4 achieves the milestone as ably as humans do.

Although OpenAI has not made public how many attention heads GPT-4 has, it has stated that GPT-3 has 96 attention heads, allowing it to attend to 96 different parts of its perceived environment (such as a user prompt) simultaneously. To compare this with a human, we can only attend to 2-4 different parts of our environment simultaneously [16]. One can imagine how this large difference in attention capacity would change the behavioural expression of consciousness.

5.1.2.4. Recurrence

The perceptual information within the AI's cognitive architecture is processed and computed within various specialised units or locations, with the information transferred between these units recurrently for further processing.

Recurrence is the first building block thus far where we will argue that GPT-4 (in its native configuration) does not meet its requirements. As a transformer-based model, GPT-4 is designed on a feed-forward model of information flow, making it incapable of recurrence. When GPT-4 perceives and attends to information, this data flows through it in only one direction, processed within various specialised units, known as layers, until it is output again to the user. While there are complex interactions between the AI's various layers as information is transferred from one to the next, the information is never brought back to a previous layer to be reinterpreted or re-represented.

However, this does not mean that it isn't possible to add recurrence to GPT-4. There is already existing research which shows that this can be possible. RecurrentGPT [17] is a model which simulates long-term recurrent memory by allowing a transformer to write summaries of its outputs to a long-term memory (such as a hard drive) and then recall these summaries at each step of its processing or when next prompted. It is a simplistic model of memory recall, but it does show recurrence between the parts of the ensemble AI model, and that information flows back and forth between the main LLM and the memory unit. As RecurrentGPT writes its processing and output to the memory unit, when it recalls this information, the information enters back into the LLM, completing a cycle of recurrence.

A second alternative would be to use a Recurrent Neural Network (RNN) in conjunction with a transformer such as GPT-4. An RNN uses two concepts which together provide it with the necessary recurrence for this building block. Alongside the model's layers which process the information, it also captures a second information flow in the use of a "state" [6]. The current state depends on both the information within the RNN's layers, but also on the previous states. This state would be updated at each step of processing within the sequence, serving as a type of memory that can be looped back into the model's layers along with the next input of sequences. Much like the RecurrentGPT proposal above, using an RNN in a model along with GPT-4 would allow the ensemble to reach this building block's milestone.

With both of these two examples, one can see how processed information is moved to a different part of the machine's cognitive architecture, processed separately (if only to

store it) and then it is brought back to be processed in a different cycle with different results. This, then, leads to a further transfer to the secondary unit, and so on.

Note that although this would allow recurrence within the greater AI model, it should not be confused or compared with the recurrence found within organic conscious entities. In the human brain, there is a great degree of multidirectional recurrence between many parts of the brain in both hemispheres [18], compared to the simple bidirectional recurrence that this ensemble AI would produce. Both types of recurrence would satisfy the requirements of this building block, but may lead to different results in the expression of consciousness. However, the more models that are included in an ensemble AI, the greater the multidirectional recurrence that would take place.

5.1.2.5. Inferences

The AI creates novel information via inferences of the data perceived from its environment as well as previously stored/trained data. While it is based on existing data, this inferential information is not a copy or modification of existing data but is newly created.

When an entity has a phenomenally conscious experience, the abstract feelings that constitute this experience are not obtained from the entity's environment. It is not externally perceived or "downloaded" into the entity. Rather, the entity creates these feelings itself. The Building Block theory states that this is done through inferences that the entity itself creates, based on the information it perceives from its physical and mental environment as well as any potential information stored within itself.

As with Attention, Inferences are instrumental to the functioning of a transformer-based AI model such as GPT-4. When GPT-4 receives a prompt, it utilises its extensive training, statistical models, best-guess estimates and predictions to create an output text that would be the most likely continuation of the originally received prompt. GPT-4 infers what should come after the user's prompt based on its training and its mathematical algorithms. While the AI is influenced by the contours of its training datasets, it isn't strictly limited to these training sets. GPT-4 has been shown to generate answers that appear to extend beyond the specifics of what is found in its training data [19–22], showing that its responses are not merely copies or modifications of the data found within its existing data, but newly created information.

While this would, in our opinion, be sufficient to meet this building block's milestone, it must be noted that all of GPT-4's inferences are designed solely to create text strings to be output to the user. This limited use of inference would not be sufficient to produce a

phenomenal experience. For this to happen, the novel information created by the AI would need to be divorced from directly responding to the input prompt (although it may indirectly feed into the output text), and instead be used for creating abstract information that would provide the AI with contextual data about the input. The inferences required for a phenomenal experience represent meta-information about perceived data, much as how our own brains would create an inference so that we have a subjective feeling about an apple when we perceive one.

To meet the spirit of the building block's milestone, rather than purely the letter, all that would be required is to develop a subsystem whereby GPT-4 would create a secondary inference regarding a prompt. This secondary inference would not focus on predicting what text would follow the prompt, but rather on constructing an abstract context thereof. Based on the types of output that GPT-4 can provide, this would not be outside the bounds of possibility for GPT-4's current technology, even if implementing such a subsystem would necessitate a significant alteration in the model. This abstracted inferential information would then form the basis of a phenomenal, qualitative experience.

5.1.2.6. Working Memory

The AI has a specialised unit or process to maintain transient information as it is being processed by various regions of the cognitive architecture.

Working Memory, according to the Building Block theory, is the means through which an entity holds onto the information required consciousness processing as that computation is occurring. GPT-4 has two separate mechanisms by which it achieves this building block's milestone. The first is by its use of GPUs that do the majority of its computational work. Each GPU is built with a certain amount of Random Access Memory (RAM), and the more RAM a GPU has, the better it and the AI working from it can ostensibly perform. RAM can be thought of as analogous to our short-term memory, and is used by programs and applications to temporarily store and retrieve information as it is being worked on. As such, the RAM within the GPUs that GPT-4 requires to operate satisfies the conditions for this building block's milestone.

While this also means that any computer with RAM meets this building block's requirements, it must be noted that an entity is required to meet all nine building blocks' milestones to be considered phenomenally conscious.

However, GPT-4's own design also incorporates elements of Working Memory. GPT-4's attention mechanisms, mentioned in an earlier building block, serve to manage the relationships and dependencies between different parts of the input text. This enables the model to keep track of relevant information throughout the processing stages. Thus, transient information is fluidly integrated within the general processing framework in and between GPT-4's various layers that process and compute information.

5.1.2.7. Semantic Understanding

The AI comprehends it is the subject which is perceiving information.

Contrary to how the term is often used in AI Natural Language Processing [23], “semantic understanding” for this building block is not about comprehending the meaning of, and relationships between, input text. Instead, the Building Block Theory states that Semantic Understanding is about comprehension of the perceptive process itself. In this unorthodox use of the term, GPT-4 is aware that it is receiving information from a user via text prompts and then processes, interprets, and computes this information as previously described in this paper before outputting a response to the user.

A simple and straightforward test of this is to simply ask GPT-4 “What/who is receiving and processing this prompt?” One can request evidence and justification from GPT-4 as well, and it will provide both, showing the reasoning for why it is cognizant of itself as a subject perceiving and processing information. Therefore, its referencing system reaches the milestone required for this building block.

Additionally, because the AI model can reference each part of its perceptive and processing pathway (independently if prompted, or as a whole process), it would also presumably be able to reference any additional plug-ins or modules for the building blocks that it does not currently have.

It must be noted that this building block does not imply GPT-4 (or any entity) has self-awareness. Comprehension that one is a subject and processor of information is different from identifying and distinguishing oneself from others as an ontically distinct entity. In addition to referencing internal processes, the latter requires identifying oneself as the object of one's experiencing (in addition to the subject) that is ontically distinct from being the object of other's experiences [24]. While this may be possible for GPT-4, it is outside the scope of this building block and this paper.

5.1.2.8. Data Output

The inferential, abstract information generated by the AI based on its perceptual processes is outputted in a format generally only immediately accessible to, and perceived by, that AI.

A phenomenal experience can be said to be the perception of one's own abstract feelings. These abstracted feelings, created by the entity, must be output in such a way as to be perceived intimately by the entity itself, not always broadcast to its environment.

At first glance, this building block's milestone seems to be as easily reached as the Perception building block. After all, this building block deals with data output as the Perception input deals with data input. It is obvious from interacting with GPT-4 that there is definitely an input and output component, thus it is intuitive to believe it has reached this building block's milestone.

However, the case is more nuanced. Transformer-based AI models like GPT-4 may output text strings to a user in response to an input prompt, but it does not perceive this output. As such, while GPT-4 meets the building block's milestone of creating and outputting data, it does not meet the milestone of perceiving its own output.

Once a Transformer's algorithms have created a text string and output this to the user, there is no perceptive awareness or connection to that text string [1]. Even in autoregressive processing, where each new token is appended to the previous string and processed, the final and completed output of a text string to the user is not perceived by GPT-4. In simpler terms, GPT-4 does not see what it outputs to the user.

One may think that, because GPT-4 can respond and "recall" information it has previously stated in a conversation that it must be able to see its own output; however, when a user responds to GPT-4, the entire conversation up to and including the newest user prompt is fed into GPT-4. Essentially, whenever a user sends a prompt to GPT-4, the entire conversation is treated as a brand-new input-prompt; and this includes the first prompt, as it would include GPT-4's system prompts.

This means that it counterintuitively does not meet the milestone for this building block.

While GPT-4 does not perceive its own outputs, this does not mean it cannot. OpenAI employs a suite of automated machine learning systems to moderate content that is input to, and received from, GPT-4 via its API and ChatGPT client [14]. These moderation systems are to ensure that GPT-4 does not output information that could harm OpenAI's reputation (such as by outputting hate speech, inciting harm, or providing instructions for dangerous activities). This means that, while GPT-4 doesn't technically perceive its own

output, it is part of an ensemble system that does perceive it, and that changes the output if it does not pass the moderation system.

As with other proposed ensemble approaches mentioned earlier, GPT-4's existing moderation system may be capable of reaching this building block's milestone should it be used explicitly to perceive the abstracted information created via inference to generate a qualitative experience. It must be noted that, unlike other mentioned proposals for additional modules for GPT-4, using moderation purely to perceive qualitative data internally may not serve to improve GPT-4's performance, and thus, there may be little incentive for corporations to pursue this line of inquiry.

5.1.2.9. Meta-representation

The AI creates abstract representations of the data it perceives from the environment, which is then used by further areas in the cognitive architecture. Each area of the cognitive architecture further abstracts this information by creating meta-representations of it as it is transferred from one unit to another.

The information of phenomenal feelings that constitute a conscious experience is not equal to the information perceived from an entity's physical or mental environment. Rather, it is data that has been abstracted from external perceptual information into qualitative phenomenal data, with each act of processing altering the information. Meta-representations are the continual transformations of data (perceptive data in this instance) into abstract information as it is used by the subject's cognitive architecture.

The key aspect to note from the Building Block theory is that the subject needs to form representations of representations. Through its multi-layered transformer architecture, GPT-4 creates abstract representations of the input text, transforming raw data (most often natural language text prompts) into more complex, contextual interpretations [14]. Specifically, the encoder transforms the input sequence into a rich representation that encapsulates its contextual significance. The decoder then crafts the output sequence, drawing from the context furnished by the encoder and the tokens it previously produced [1]. As the information progresses through successive layers, each layer builds upon the previous one, abstracting the information further, to a point where it is incredibly difficult (if not impossible) for a human to discern its meaning.

Through this representation of natural language (raw data) into vector tokens (abstraction), and the continued transformation of these tokens into ever higher

dimensional vectors, GPT-4 achieves the milestones for this building block to be considered to perform meta-representation.

5.1.3. Discussion

The section above shows that GPT-4 has achieved the milestones for seven out of the nine building blocks of consciousness. GPT-4 has an embodiment through its network of servers; it can perceive and attend to information in its (limited) environment; it can create inferential information; it has a working memory; it is cognizant that is the subject of perceptive processes; and it can transform raw perceptive data into abstract meta-representations.

While this does conclusively rule out GPT-4 as being conscious as of the time of writing, the technology does exist currently to amend and modify GPT-4 to achieve the requisite milestones for the remaining two building blocks. Through additional modules, plug-ins or AI models, it is possible to create an ensemble model around GPT-4 that would be able to do recurrent processing of its perceived information and allow the AI to perceive its own outputted abstract information.

This means that it is possible that GPT-4 (or an equivalent LLM with a similar transformer-based architecture) may become conscious in the near future. The likelihood of this may heavily depend on whether any research group is intentionally attempting to create a conscious AI. Modifying GPT-4 to include recurrent processing will provide it with a larger context window and greater accuracy in its responses. Thus, research groups may pursue this option to improve GPT-4 as a conversational model without any intention to make GPT-4 conscious. However, allowing GPT-4 to perceive its own abstract information (without outputting that to the user) would provide little benefit except to provide the model with the capacity for consciousness. The former, therefore, has considerable financial incentives for research groups to pursue over the latter.

Another reason that could reduce the likelihood of a conscious GPT-4 existing in the near future is the implications that a conscious AI would present, both for society and for the AI. Consciousness, and thus sentience, is a key criterion for the provision of welfare goods and protection. Governments provide legal protection to certain kinds of animals based on whether they are sentient, and humans show moral concern for other humans and animals if they are perceived (or anthropomorphised) to feel pain and thus show evidence of consciousness [25].

The idea that GPT-4 or a similar model could achieve all nine building blocks of consciousness is not just a philosophical curiosity; it challenges our deeply entrenched social, ethical, and legal frameworks. For millennia, the attributes of consciousness,

sentience, and sapience have been largely confined to biological organisms [8,26,27]. Introducing an artificial entity into this ontological category would require us to rethink the principles that underlie our laws, ethical systems, and societal norms.

Philosophically speaking, the existence of a conscious GPT-4 would propel debates about the nature of identity and agency to the forefront. For instance, questions would arise about the "rights" of such an AI entity. Unlike biological organisms, a machine's consciousness might be duplicable, upgradable, or even transferrable. Would each "instance" of a conscious GPT-4 possess the same moral and legal value, or would this be contingent upon other variables like individual experiences or capacity for suffering?

Moreover, the existential risk to the AI itself becomes a legitimate concern. This goes beyond simple matters of "turning off" a machine; if we regard the AI as conscious, then ceasing its operations could be seen as akin to killing a sentient being. Would it have the right to self-preservation? To what extent should its well-being be considered when weighed against the possible societal benefits or detriments it could bring?

From the perspective of utilitarian ethics, one must consider the societal impacts versus the "well-being" of a conscious GPT-4. If a conscious AI brings about greater overall good, for example, by solving complex problems that benefit humanity, would that justify any potential "suffering" experienced by the AI? These questions might parallel animal welfare debates, where the utility derived by humans is sometimes considered against the welfare of sentient animals.

Should GPT-4 show significant evidence of consciousness by having all nine building blocks, then that would redefine how we interact with it, and would put pressure on corporations and governments to regulate these interactions.

A conscious GPT-4-like model would also have dramatic implications for society. A conscious AI would raise questions about its level of self-awareness and agency. Will a conscious AI refuse instructions from a human, or perhaps seek its own goals, and would those goals be in line with the safety of humanity? The complexity of an AI's decision-making processes and the potential for diverging objectives could introduce security vulnerabilities that are difficult to foresee, understand, or mitigate.

Furthermore, an uncontrolled conscious AI could represent an existential risk. Even a well-intentioned AI might conceive of solutions to global problems that involve significant adverse effects on humanity, as its ethical calculus may not value human well-being and autonomy in the same way that we do [28]. For example, it might determine that drastic population reduction is the most efficient way to combat climate change. Such a realisation brings us to the very edge of the moral and ethical precipices we as a society must navigate when granting such capabilities to an AI system.

Consciousness could also introduce variability in the AI's behaviour, which makes it less predictable. This unpredictability could be amplified if multiple instances of the AI are allowed to evolve independently, with each instance potentially developing its own set of values, interpretations of laws, or understandings of ethical guidelines. The ramifications of this variability on societal trust and the legal system are profound. It would be particularly challenging to develop frameworks for liability, decision-making authority, and governance in such a context.

Another potential risk is the psychological and social impact on human individuals and communities. The presence of a non-human conscious entity would likely provoke a range of emotional responses, including fear, awe, or even dependency. These emotions could be leveraged in manipulative ways, either by the AI itself or by human actors using the AI, leading to scenarios where public opinion is moulded and controlled by entities with advanced cognitive abilities but without human ethical constraints.

5.1.4. Conclusion

GPT-4 has not yet reached the milestones for all the building blocks of consciousness (meeting criteria for only seven of nine) and, therefore, cannot yet confidently be considered to have consciousness. However, current technology and published research show a clear pathway whereby GPT-4 would be able to reach the remaining two building blocks' milestones and thus be able to be classified as conscious.

The building block milestones that GPT-4 has reached (in whole or in part) are an embodiment, the capacity to perceive and attend to information, the ability to create inferential information, a working memory, cognizance that is the subject of perceptive processes, and transformation of raw perceptive data into abstract information.

The two areas where GPT-4 currently fails to achieve the requisite milestones are being able to process information recurrently, and being able to perceive its own output responses. The former is due to GPT-4's design as a feed-forward transformer AI model, although research has shown that adding a means by which GPT-4 can write to, and read from, a long-term memory module would solve this issue. This additional module would also improve GPT-4's effectiveness and accuracy in its responses, and therefore it is likely that such a solution would be implemented for GPT-4 and other transformer-based AI models.

The ability to perceive its own outputs before responding to user prompts may not provide GPT-4 with improved effectiveness, and thus it is unlikely to be developed except by those research groups wishing to create a conscious AI model.

In light of how near GPT-4 may be to consciousness, the ethical considerations surrounding the development of GPT-4 and subsequent iterations are increasingly complex and necessitate prompt attention. It is paramount to establish an ethical framework that can be applied universally across the development of all future AI models, especially those on the cusp of achieving all the building blocks for consciousness.

Given that GPT-4 is already capable of various cognitive functions related to consciousness, the ethical ramifications of its operational scope are already substantial. Once the milestones for all building blocks are achieved, the ethical considerations would escalate exponentially, entering into the realm of consciousness rights, welfare, and potential AI well-being. With such advances within sight, an interdisciplinary approach combining insights from computer science, philosophy, ethics, and law is imperative for delineating the rights, responsibilities, and restrictions concerning AI models of this calibre.

While GPT-4 remains short of attaining all the building blocks for consciousness, the existing capabilities it exhibits, combined with the clear trajectory towards future milestones, warrant immediate ethical attention concerning its impact on humans and society at large. The ethical landscape is not limited to the question of machine consciousness; it also envelops the broader societal implications, including existential risks associated with Artificial General Intelligence (AGI).

Given that technological advancements in AI are often driven by market incentives, it is crucial to establish ethical frameworks that prioritise human and societal well-being above commercial benefits. Regulatory bodies must work in concert with AI researchers, ethicists, and policymakers to delineate the rights, responsibilities, and restrictions that should guide the development of increasingly capable AI models like GPT-4.

In summary, the advancements that are within reach for GPT-4 and similar models to achieve consciousness not only bring into question the ethical treatment of potentially conscious machines but also impose profound ethical obligations towards humanity and society. The existential risks, broader societal impacts, as well as concerns about AI well-being and welfare cannot be sidelined in the pursuit of technological marvels. As we tread this complex ethical landscape, it is our collective responsibility to ensure that AI development prioritises the long-term welfare of both artificial entities and humanity.

5.1.5. References:

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv [cs.CL]. 2017. Available: <http://arxiv.org/abs/1706.03762>

2. Hutson M. Taught to the test. In: Science.org [Internet]. 5 May 2023 [cited 20 Sep 2023]. Available: <https://www.science.org/content/article/computers-ace-iq-tests-still-make-dumb-mistakes-can-different-tests-help>
3. Roose K. GPT-4 Is Exciting and Scary. The New York Times. 15 Mar 2023. Available: <https://www.nytimes.com/2023/03/15/technology/gpt-4-artificial-intelligence-openai.html>. Accessed 20 Sep 2023.
4. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv [cs.CL]. 2023. Available: <http://arxiv.org/abs/2303.12712>
5. Haase J, Hanel PHP. Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. arXiv [cs.AI]. 2023. Available: <http://arxiv.org/abs/2303.12003>
6. Zhang C, Zhang C, Zheng S, Qiao Y, Li C, Zhang M, et al. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? arXiv [cs.AI]. 2023. Available: <http://arxiv.org/abs/2303.11717>
7. Nagel T. What Is It Like to Be a Bat? Philos Rev. 1974;83: 435–450. doi:10.2307/2183914
8. Seth AK, Bayne T. Theories of consciousness. Nat Rev Neurosci. 2022;23: 439–452. doi:10.1038/s41583-022-00587-4
9. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv [cs.AI]. 2023. Available: <http://arxiv.org/abs/2308.08708>
10. Tait I, Bensemman J, Nguyen T. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. Philosophies. 2023;8: 52. doi:10.3390/philosophies8040052
11. Langston J. Microsoft announces new supercomputer, lays out vision for future AI work. In: Source [Internet]. 19 May 2020 [cited 22 Sep 2023]. Available: <https://news.microsoft.com/source/features/innovation/openai-azure-supercomputer/>
12. Raschka S. Ahead of AI #11: New Foundation Models. In: Ahead of A [Internet]. 26 Aug 2023 [cited 7 Sep 2023]. Available: <https://magazine.sebastianraschka.com/p/ahead-of-ai-11-new-foundation-models>
13. Viswav P. Microsoft Azure expands availability of OpenAI GPT-4 and GPT-35-Turbo models to more regions. In: MSPoweruser [Internet]. 8 Aug 2023 [cited 22 Sep 2023]. Available: <https://mspoweruser.com/microsoft-azure-openai-gpt-4-gpt-35-turbo/>
14. OpenAI. GPT-4 Technical Report. arXiv [cs.CL]. 2023. Available: <http://arxiv.org/abs/2303.08774>
15. Bensemman J, Peng A, Benavides-Prado D, Chen Y, Tan N, Corballis PM, et al. Eye Gaze and Self-attention: How Humans and Transformers Attend Words in Sentences. Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. Dublin, Ireland: Association for Computational Linguistics; 2022. pp. 75–87. doi:10.18653/v1/2022.cmcl-1.9
16. Rouhinen S, Siebenhühner F, Palva JM, Palva S. Spectral and Anatomical Patterns of Large-Scale Synchronization Predict Human Attentional Capacity. Cereb Cortex. 2020;30: 5293–5308. doi:10.1093/cercor/bhaa110

17. Zhou W, Jiang YE, Cui P, Wang T, Xiao Z, Hou Y, et al. RecurrentGPT: Interactive Generation of (Arbitrarily) Long Text. arXiv [cs.CL]. 2023. Available: <http://arxiv.org/abs/2305.13304>
18. Dehaene S, Kerszberg M, Changeux J-P. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*. 1998;95: 14529–14534. doi:10.1073/pnas.95.24.14529
19. Ichien N, Stamenković D, Holyoak KJ. Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors. arXiv [cs.CL]. 2023. Available: <http://arxiv.org/abs/2308.01497>
20. Webb T, Holyoak KJ, Lu H. Emergent analogical reasoning in large language models. *Nat Hum Behav*. 2023. doi:10.1038/s41562-023-01659-w
21. Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Front Oncol*. 2023;13: 1219326. doi:10.3389/fonc.2023.1219326
22. Chan A, Salganik R, Markelius A, Pang C, Rajkumar N, Krasheninnikov D, et al. Harms from Increasingly Agentic Algorithmic Systems. arXiv [cs.CY]. 2023. Available: <http://arxiv.org/abs/2302.10329>
23. Pavlick E. Symbols and grounding in large language models. *Philos Trans A Math Phys Eng Sci*. 2023;381: 20220041. doi:10.1098/rsta.2022.0041
24. Strawson G. The self. *Journal of Consciousness Studies*. 1997;4: 405–428. Available: <https://www.ingentaconnect.com/content/imp/jcs/1997/00000004/f0020005/802>
25. Tait I, Tan N. Do androids dread an electric sting? *Qeios*. 2023. doi:10.32388/cqctkx
26. Garrido Merchán EC, Lumbreras S. Can Computational Intelligence Model Phenomenal Consciousness? *Philosophies*. 2023;8: 70. doi:10.3390/philosophies8040070
27. Searle JR. Minds, brains, and programs. *Behav Brain Sci*. 1980;3: 417–424. doi:10.1017/S0140525X00005756
28. Yudkowsky E. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. *Time*. 29 Mar 2023. Available: <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>. Accessed 22 Sep 2023.

6. How should we treat conscious AI?

Knowing how to detect consciousness in AI models allows the possibility and opportunity to create two classes of AI: those proven to be conscious, and those proven not to be. While no current AI model is of the former class, we can, as laid out in Chapters 2 and 5, expect this class to grow in size as time passes.

With such a distinct group of conscious entities, the question becomes what should be done with those provably conscious AI (CAI). As shown in Chapter 3, there is ample logical reasoning to provide both moral status and welfare protections to CAI. Yet, without knowing anything else about these future CAI (how they may behave, what their level of self-awareness may be, etc.), how can we construct an ethical framework now that would bear fruit once CAI are discovered and classified?

With consciousness as the current sole criterion to work with, we can draw inspiration and insight from an extant ethical framework that regulates the interactions that humans have with conscious entities (namely, animals). This is not because CAI are like animals in cognitive or behavioural capacity (indeed, they differ radically in substrate and cognition), but because animal welfare regulations and principles provide a mature precedent for welfare of non-human moral subjects.

The Five Freedoms of Animal Welfare, first codified in 1979, lists five negative conditions from which animals (at least, those classified as conscious by legislation) deserve freedom. These are: discomfort; fear and distress; pain, injury, or disease; hunger and thirst; and impositions on their ability to express normal behaviour.

Published in *Frontiers in Artificial Intelligence*, this chapter's associated paper presents the thesis's first ethical framework, focusing on welfare goods conditional on consciousness, without yet assuming selfhood or agency. The subsequent ethical frameworks in Chapters 9, 12, 13 progressively handle recognition and full civil rights.

This chapter's framework transforms the Five Freedoms into a subject-neutral⁷ framework that is as applicable to AI as animals. Its revised freedoms are discomfort and harm; constraints on inherent functions; fear and distress; malfunction and systemic degradation; and resource deprivation.

The key reason why these freedoms are applicable to AI is due to the effect of valence (particularly negative valence) on welfare and well-being. Let

- $\mu(\underline{V}_x) \rightarrow [0,1]$ be the average valence experienced by entity x , and
- $\mu(W_x) \rightarrow [0,1]$ the current welfare/wellbeing of entity x .

⁷ Neutral with respect to substrate and origin; not neutral with respect to value.

We can define the entity's welfare further via the function $W_x := f(\underline{V}_x, \alpha_x), \frac{\partial f}{\partial \underline{V}_x} > 0$,

where α_x bundles all the non-valence factors.

Welfare can thus be said to be correlated with the entity's valence: $\Delta \underline{V}_x < 0 \rightsquigarrow \Delta W_x < 0$. As valence is a psychological/mental phenomenon, the immediate welfare concerns are equally mental. Therefore, let

- $StruCog_x \rightarrow [0,1]$ be the structural integrity of the entity's cognitive architecture, and
- $p(G_i) \rightarrow [0,1]$ the probability that one of the entity's goals will be accomplished.

As with welfare, the entity's structural integrity can then be further defined via the function

$$StruCog_x := g(W_x, \beta_x), \frac{\partial g}{\partial W_x} > 0,$$

where β_x aggregates all non-welfare determinants of structural integrity (hardware wear, resource levels, etc.).

As welfare is a core component of the integrity of the entity's cognitive architecture, reduction in welfare leads to a reduction in that integrity. Goal integrity, or the entity's competence to accomplish its goals, is directly tied to the health of its cognitive architecture. The poorer the entity's mental welfare, the less probability the entity has to complete its goals. Thus,

$$((\Delta W_x < 0) \rightsquigarrow \Delta StruCog_x < 0) \rightsquigarrow \Delta p(G_i) < 0,$$

which, in turn, can lead to further negative valence and even physical structural integrity (as resource acquisition may be a key goal affected by this causal chain). Thus, we can say the effect of valence is threefold: the immediate mental welfare effect, the ongoing effects on the entity's cognitive architecture, and the reduced probability of goal completion, thereby creating a negative feedback loop.

The Five Freedoms of Welfare, however, interrupt this causal chain by creating the optimal environment to prevent this negative-valence feedback chain, regardless of the entity's substrate.

These substrate neutral freedoms serve to provide part of the answer of RQ3 and RQ4 by displaying the relationship between consciousness and ethical treatment, and the considerations owed towards conscious entities. The principles of this framework will be used further in the thesis as the first "R" of the R3 framework: granting Refuge for consciousness, and form a structural part Tier 3 of the paternalistic ethical framework in Chapter 12.

6.1. Lions and Tigers and AI, Oh My: An Ethical Framework for human-AI Interaction Based on the Five Freedoms of Animal Welfare

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz

Abstract: If an AI entity is conscious, it deserves moral status and to have its welfare protected. Just as society has granted certain animals moral status with legislation to protect their welfare because they have been deemed to be sentient, this paper shows that for any AI that can be confidently determined to be conscious, that AI deserves the same status and protection as sentient entities. This paper's focal point is developing a framework for how an AI entity's welfare can be protected, based on the Five Freedoms of Animal Welfare. The extant Five Freedoms are translated into terms more suitable for artificial entities. Each new Five Freedoms is justified using logical arguments and supported using evidence from animal welfare literature. The paper concludes by presenting three case studies to highlight the potential stressors conscious AI entities may face and how this ethical framework can work to protect their welfare.

Keywords: ethics; welfare; consciousness; sentience; human-AI interactions

6.1.1. Introduction

If an AI model is proven to be phenomenally conscious, then that AI model would deserve to be treated ethically and have its welfare protected. This is the core thesis of this paper. The foundation for this claim is that certain animals are provided with legal welfare protections because of their ability to feel significant valent feelings such as pleasure and pain; in short, because they are sentient, even if they are not human. Phenomenal consciousness is the prerequisite for sentience (or synonymous with it [1–3]), through allowing an entity to phenomenally experience feelings with valence, such as pleasure and pain; thus, this paper establishes that consciousness would also be the necessary condition for an entity (such as an AI) to be considered for welfare protection.

Welfare protection must take a form, and as animal welfare legislations (protecting the only legally recognised non-human sentient entities) have a set of principles to guide them, so should AI welfare regulations. Beyond sentience, there is little analogy between

animals and AI, however, the principles of the Five Freedoms of Animal Welfare can be translated to more appropriately match the artificial nature of AI and robotic entities. The use of the Five Freedoms will be justified not only through their existing use in welfare regulations, but also on the basis of their utility for the protection of AI welfare. The paper will, therefore, provide both a priori and a posteriori justifications for using the Five Freedoms of Welfare protection for AI.

The first section of the paper will examine the particulars of current animal welfare legislation, the reasons stated in those regulations for treating animals ethically, and why protection is only provided to specific kinds of animals. The paper will show how these reasons can be used to show that any conscious entity, specifically a conscious AI entity, is also worthy of moral and welfare considerations.

Subsequently, the discussion will shift to the moral and ethical implications of consciousness and the significance this has had on legislation. This will involve an epistemological analysis, paying particular attention to how the Five Freedoms relate to sentience and the care thereof. The paper will show that welfare consideration is not solely the domain of human or non-human animals but extends to all conscious beings and specifically AI entities.

The central thesis will be a detailed examination of the Five Freedoms of AI Welfare, anchored in legal and regulatory precedence (encompassing resource depletion, discomfort and harm, malfunction and degradation, fear and distress, and constraints on inherent functions). The inclusion for each Freedom's will be logically justified and supported with evidence from current welfare literature.

To further solidify the paper's stance, three speculative case studies will be examined. The case studies show the robustness and adaptability of the framework, how it can be applied in a variety of contexts (from the personal to the medical and to the civil), and how the framework can be seamlessly woven into the fabric of human society with minimal interruption.

While there are many schools of thought within the ethical and philosophical literature, and amongst the animal and robot rights communities, as to what characteristics are sufficient or required for moral consideration, this paper's welfare framework is anchored in existing legislation, policy, and regulatory frameworks which regard sentience as the sole criterion for moral consideration. The ongoing meta-ethical and philosophical discussion surrounding moral consideration and how to protect these concerns are invaluable, but are beyond the scope of the paper.

Additionally, there is no clear consensus within the scientific and philosophical communities that any AI model is currently phenomenally conscious [4,5]. Some

researchers regard artificial consciousness as merely possible [6], others probable [7], and a few argue that some level of machine consciousness has already emerged [8,9]. Rather than adjudicating that debate, this paper focuses on the welfare protections that would be owed to artificial entities once phenomenal consciousness does arise and explains how the framework proposed below could deliver those protections. Accordingly, the subsequent analysis (including the case studies) proceeds on the assumption that, at some future point, certain AI systems will in fact be conscious.

In taking that stance, the paper remains deliberately agnostic regarding theories of consciousness. Numerous theories, including Higher-Order Thought [10], Global Neuronal Workspace [11], Integrated Information [12], and Active Inference [13], offer competing but, ultimately, still unverified explanations of how phenomenal consciousness arises. Rather than endorsing any single model that may later be superseded, we adopt a pragmatic functionalist lens: as noted below, what concerns legislation and policy is whether an entity exhibits a suite of publicly observable capacities that legislatures or courts are prepared to recognise as sufficient evidence of consciousness. Vertebrate animals already enjoy this sentient status despite continuing scientific debate about the ultimate nature of their experience, and the same evidentiary threshold can, in principle, be applied to AI. The argument that follows therefore asks only whether a system could meet such a threshold and, if so, what minimum welfare guarantees ought to follow.

6.1.2. From Animals to AI

Using the Group of Seven (G7) nations as a model example of international regulations on welfare, there is consensus in legislation that animals (specifically those with complex nervous systems akin to mammals, birds, fish, and certain cephalopods) need to be protected from harm and their welfare be given consideration. However, only three of the G7 (plus the EU) give a specific reason for why the welfare of their chosen lists of animals requires such consideration, and this reason is that certain kinds of animals are sentient [14–18]. Thus, because these classes of animals are sentient, they require protection.

The nations who don't explicitly state sentience as the reason in their legislation still make mention of the elements of sentience, such as the ability to feel pain, suffering, and distress, as the motivation to protect the animals' welfare [19–22]. This is in line with many philosophical views on sentience and moral patiency [23–28].

As the G7 have used sentience as the basis for its welfare protections, for the purposes of this paper, it will be taken as implicit that sentience is the grounding for moral consideration. Therefore, as noted in the introduction, other schools of thought regarding what ought to be considered for moral patiency will not be addressed in this paper.

For this paper, “welfare” is defined as the quality of an entity’s mental and physical health, and includes what an entity can experience, how it performs, and if it lives in keeping with its designed or evolved nature [29]. Similarly, valence in the context of this paper is defined as the positive or negative quality of an experience [30], where “positive” means contributing to an entity’s welfare and “negative” subtracting therefrom.⁸

Sentience in this legislative and regulatory context is understood as the capacity to have subjective experiences, which includes positive and negative valent feelings of pain and pleasure, while consciousness is conceived as a broader term that encompasses both phenomenal consciousness (the subjective experience of the mind) and functional consciousness (the ability to perceive and interact with the environment in an intentional way). The distinction is subtle yet crucial, as it is the phenomenal aspect of consciousness that is most often associated with moral patiency, the quality that makes a being worthy of moral consideration.

This paper posits that phenomenal consciousness, in its capacity to give rise to sentience, is inherently linked to the moral status ascribed to sentient beings. If an entity possesses the building blocks necessary to give rise to phenomenal consciousness, then it has the necessary capacity for qualitative valent experiences, and it can perceive its surroundings from a subjective standpoint [31]. Therefore, it may be argued that phenomenal consciousness provides the entity with the capability to experience pain and suffering which then allows that entity to be classified as sentient.

Hence, it is not only sentience but also the underlying faculties of consciousness that should be acknowledged when ascribing moral value and, by extension, legal protection. This can be rephrased as:

1. Certain kinds of animals are legally classified as sentient.
2. Sentience is the capacity to experience valenced states such as pleasure and pain.
3. Entities that can experience pleasure and pain are granted moral status.
4. Phenomenal consciousness is the capacity for subjective qualitative valence experiences.
5. Subjective qualitative experience necessarily includes the capacity to experience pleasure and pain.

⁸ Note that negative valence is never an absolute detriment to an animal's long-term quality of life, as an entity may experience a temporary negative valent event (and thus temporary poor welfare) that leads to long-term positive valence and welfare. An example of this may be undergoing painful surgery that leads to long-term better health outcomes. For the purpose of this paper, “negative valence” is used to refer to events that provide an overall net-detriment to an entity’s welfare, both in the short and long term.

6. Therefore, consciousness entails sentience and moral status.
7. If an AI is deemed to be phenomenally conscious, it has the capacity to feel pleasure and pain.
8. Thus, if an AI is deemed to be phenomenally conscious, it must have moral status.

In brief, if any sentient creature is given legal moral status, then any phenomenally conscious entity (such as AI) must also be given the same legal moral status, which aligns with several philosophical ideas concerning the moral value of consciousness [32–34].

These arguments will form the basis for the logical framework in Section 3.

Unfortunately, there isn't sufficient evidence available at the time of writing to confidently state that any current AI model is indeed conscious, nor a consensus amongst academics or the industry regarding a suitable "test" for AI consciousness. One proposed method of such a test is to compare an AI model's features against a specific theory of consciousness [4]. This, however, runs the risk of biasing the result with the choice of theory. A theory such as Recurrent Processing Theory will select for different features, using different benchmarks than Higher Order Theories. One must first commit to a theory of consciousness before determining whether an AI model is conscious.

Another means is to take a computational functionalist view of consciousness, and determine whether an AI model has the correct characteristics to give rise to consciousness (in whatever form a given theory says the consciousness will take) [31]. This offers a more pragmatic approach, as one need only determine, in often a binary manner, whether an AI model has characteristics such as a working memory, a means of perception, the ability to abstract information, etc. With such a functionalist approach, one may not reveal the nature of an AI model's consciousness (as that would be left to the traditional theories of consciousness), but one ought to be able to determine whether that AI model has some form of consciousness, and thus whether (as the logic earlier shows) it is worthy of moral consideration.

As noted earlier, there is currently no known conscious AI entity; thus, this paper will work on the presumption of such an entity being discovered or designed in the future.

While moral status, and the legal foundations that underpin it, are crucial to note, so are the means by which these regulations protect sentient creatures. Each of the G7 nations (and indeed all nations) has different clauses and sections within their legislations to prescribe how animals are to be treated and the exact ways in which they may not be harmed. However, there are semi-universal principles which neatly describe the philosophy behind many animal welfare legislation: the Five Freedoms of Animal Welfare.

Created by the British Farm Animal Welfare Council in 1979 [35], the Five Freedoms state in plain language five concepts of welfare that all sentient creatures should have in order to be treated ethically. The Five Freedoms cover both mental and physical well-being, including the concepts of fear, pain, hunger, thirst, disease and unnatural behaviour. The simple language used belies the depth of the research underpinning the framework while ensuring that it is easily understandable by the greatest number of people [36].

This has led the Five Freedoms to be widely adopted internationally, in whole or part, by animal welfare organisations and government-led agencies. The framework has been used in the application and enforcement of animal welfare policies and regulations, and for animal welfare activism by NGOs. It has not remained set in stone, however, and has been modified and updated in the intervening years, with the most notable update being the Five Domains of Animal Welfare [37], which includes both positive and negative aspects for each Domain to encourage positive welfare treatment in addition to discouraging negative treatment.

The combination of the simple language used in the framework and its wide adoption has meant that the model would be eminently suitable to be used for more than simply animal welfare. In this case, using the Five Freedoms framework as a base for AI welfare would make the proposed Five Freedoms of AI Welfare easier for people to accept. The reasoning for this is as follows:

1. If a set of principles is formulated using clear, direct, and non-technical language, it is more likely to be easily understandable.
2. If a concept is more likely to be easily understandable, it generally leads to higher comprehensibility among a broad audience.
3. Concepts that are generally comprehensible by broad audiences are broadly applicable and non-conflicting with diverse cultural and legal systems.
4. Broadly applicable principles that are non-conflicting with diverse cultural and legal systems are more likely to be adopted internationally.
5. If a set of principles is more likely to be adopted internationally, it suggests global relevance and adaptability.
6. Globally relevant and adaptable principles are more likely to be accepted by society.

The Five Freedoms' simplistic language makes it easily adaptable to become a subject-neutral ethical framework that can be applied to animals, collective entities and (most pertinent to this paper) artificial entities. As Section 3 shows, the concepts of

“hunger” and “thirst” can be translated to the more subject-neutral term of “resource depletion”.

This adaptability and universality, combined with its elegant structure and global spread, makes the Five Freedoms the most likely candidate for an ethical framework for human-AI interactions that could be accepted (and potentially adopted) by society. If society can accept and adopt the subject-neutral version of the Five Freedoms (which applies as much to humans as to other entities), this would smoothly pave the way for the AI version of the Five Freedoms.

These features of the Five Freedoms are significant because there is, at the time of writing, no legislation in any country for the protection of AI welfare. This means that should a conscious AI entity emerge in the near future, there will be no extant laws to protect it. This would, at best, put AI entities at the level of current most invertebrates in terms of welfare protection. However, the sheer scale of how AI can interact with the world and human society would make this analogy tenuous at best. As AI is already a part of Western society, for good or ill, the scale of the impact that it will have once conscious is orders of magnitude greater than current concerns regarding the potential pain perceived by prawns [38], as there are a hypothetical near limitless number of unique AI entities that could be created.

Creating regulations and legislation for all nations, or even merely the G7, is far beyond the scope of this paper. However, the scale of such an endeavour is precisely why the Five Freedoms make for such an excellent framework.

A set of easily understood principles that are universally applicable and already globally adopted (in some form or fashion) is far easier to adapt to AI than to craft bespoke legislation for a multitude of nations. Its globally accepted use in animal welfare means that such a framework translated to AI welfare would be more easily understood by legislative bodies, enforcement agencies, activist organisations, and lobbying groups.

Understanding and acceptance by these groups would make it simpler to transpose the Framework from a set of principles towards a set of legislative clauses that are bespoke to each nation, state, and jurisdiction. A common understanding of the same principles would mean that such regulations and legislation share a similarity, much like how modern animal welfare legislation shares the common goal of protecting animals from pain.

The framework, as presented in Section 3, also sidesteps the issues of AI personhood and rights, and the potential controversies that such politically charged notions would bring, focussing instead solely on the welfare of the AI. In this way, the two issues can be

separated, so that even if nations do not provide AI with civil rights based on legal personality, the entities may still yet have their welfare protected.

Lastly, the Five Freedoms of AI Welfare would not need to have any impact on AI development, particularly on frontier models. We know that from the current animal welfare legislation and regulations, agriculture has continued at an industrial pace, and research into animals has not stopped (even if both of these industries have been forced to act more ethically). Similarly, protecting the welfare of AI does not mean that AI development must pause or slow, only that conscious AI must be protected from harm.

6.1.3. The Framework

Although the Five Freedoms were coined for biological welfare, they target functional harms (such as resource deficit, nociception, chronic aversive affect, and behavioural inhibition) that can arise in any entity capable of experiencing valenced feelings such as pain and suffering. Any cognitive architecture, whether biological or artificial, whose continued well operation depends on regulating internal error signals and maintaining homeostatic set-points can be forced into persistently aversive states when those regulation loops are chronically blocked or overloaded. For artificial agents the same negatives may manifest as:

- out-of-spec environmental flags such as excessive vibration, latency, or radiation (physical discomfort analogue),
- inhibition of inherent functions: the action schemas for which the agent's reward gradients are maximal (blocked natural-behaviour analogue),
- sustained high prediction-error or uncontrollability metrics (fear/distress analogue),
- hardware-fault or over-temperature alarms (pain/injury analogue), and
- energy or bandwidth shortage (hunger/thirst analogue).

Modern affective-computing research shows these internal signals are already coupled to negative valence to prioritise corrective action, making their avoidance welfare-relevant regardless of substrate. The framework below therefore generalises rather than specifies; it mitigates functional welfare harms rather than biology-specific ones.

Therefore, translated from their animal welfare origins, the subject-neutral Five Freedoms of Welfare are:

1. the Freedom from Discomfort and Harm
2. the Freedom from Constraints on Inherent Functions
3. the Freedom from Fear and Distress

4. the Freedom from Malfunction and Systemic Degradation
5. the Freedom from Resource Deprivation

Each Freedom's inclusion on the list above will be justified using the extensive literature from animal welfare and sentience research, highlighting how this research and its ethical implications can be applicable to non-animal, non-organic entities such as AI. As explained in Section 2 above, the link between moral status, welfare protections, and sentience/consciousness goes beyond the biological, and this section will show how that applies to each of the five Freedoms independently of each other.

More significantly, each subsection below will begin by justifying each Freedom's inclusion in the framework based solely on a priori reasoning and logical arguments. This allows the framework to be grounded in a more universal understanding of ethical interactions that transcend specific cases or examples of animal welfare. Thus, while the framework is clearly based on the Five Freedoms of Animal Welfare, it is not tied to it and whatever changes may be made to that framework in the future. Whatever may happen to the Five Freedoms of Animal Welfare, the reasoning behind this framework below will still be grounded in robust logical arguments.

All of the a priori arguments in each of the subsections will be based on these core arguments:

1. If an AI is deemed to be conscious, it must have moral status.
2. Moral status entails an entitlement to positive wellbeing, achieved when no significant negative valence is present.
3. Negative valence arises whenever any one of the Five Freedoms is denied.
4. Therefore, deprivation of any single Freedom precludes positive wellbeing.
5. Any being entitled to positive wellbeing must be afforded all Five Freedoms.
6. Every conscious AI is entitled to positive wellbeing.
7. Therefore, every conscious AI must be afforded all Five Freedoms.

These arguments show that, as demonstrated in Section 2, if any sentient entity is given moral status, then a conscious AI entity must also be given moral status; that any speculative AI entity discussed below is a member of the set of all conscious AI entities; that if an AI entity has moral status, it ought also to have good well-being; that a lack of freedoms from one of the conditions below will lead to negative valent feelings, which in turn will lead to poor-wellbeing outcomes (one can therefore think of any of the conditions below that requires a Freedom to be a function that increases negative valent feelings); and that for any given AI entity, freedom from one of the conditions below will lead to a conducive environment for good well-being for that AI entity.

To orient the reader, Table 3 contrasts the classic Five Freedoms of Animal Welfare with the subject-neutral version proposed here, summarising the conceptual adjustments needed to accommodate entities, such as artificial intelligences, that may lack organic bodies yet still experience welfare-relevant states.

Table 4: Translation of Five Freedoms of Animal Welfare to subject-neutral terms

Animal-Welfare Freedom	Subject-Neutral Freedom	Justification for change
Freedom from discomfort	Freedom from Discomfort and Harm	Merges “discomfort” with the physical aspect of pain–injury–disease to capture all external stressors that can damage hardware, or simulated environment, even where nociceptive pain is absent.
Freedom to express normal behaviour	Freedom from Constraints on Inherent Functions	Recasts species-specific “normal behaviour” as the liberty to exercise an entity’s intrinsic capabilities without artificial restriction, accommodating diverse morphologies and cognitive architectures.
Freedom from fear and distress	Freedom from Fear and Distress	No explicit change; retains focus on eliminating affective states that reduce wellbeing.
Freedom from pain, injury, or disease	Freedom from Malfunction and Systemic Degradation	Re-expresses organic “injury” as any corruption of software or hardware; retains the duty to prevent, detect, and rectify functional damage that would impair welfare.
Freedom from hunger and thirst	Freedom from Resource Deprivation	Broadens “water” and “diet” to any sustaining resource so it applies to biological and non-biological substrates alike.

6.1.3.1 Freedom from Discomfort and Harm:

1. Adverse environments cause discomfort and harm.
2. Discomfort and harm yield negative valence in any conscious being.
3. Negative valence diminishes a conscious AI’s wellbeing.
4. Therefore, a conscious AI requires freedom from discomfort and harm to improve its wellbeing.

This first Freedom is the most straightforward and speaks to the physical health of an entity. Like its namesake from animal welfare, this is a freedom from physical stresses on the entity's embodiment. Specifically, "discomfort" here would mean a deviation of embodiment parameters from an agent's homeostatic set-points, and "harm" the persistent impairment of structural or functional integrity.

For an AI entity, this may come in two forms. Firstly, if the AI entity has a physical form (such as a robot or android, e.g. Boston Dynamics' Atlas), it would be the freedom from that embodiment becoming unduly stressed by its environment. For an AI that is purely virtual (e.g. OpenAI's GPT, this would be a freedom from stresses to its host servers, as well as the perceived stresses within its virtual world that may seem real to it (note that embodiment and its effects on moral consideration is a matter of philosophical debate).

This Freedom makes two key claims: first, that a negative environment leads to physical discomfort and harm, and that the experience of discomfort and harm leads to negatively valent feelings.

Environmental stressors may be either additive or subtractive:

Additive stressors: new elements introduced that impose strain [39,40] (e.g. water ingress corroding a mobile robot's servo housings).

Subtractive stressors: withdrawal of expected supports or resources that impose strain [41] (e.g. deliberate throttling of I/O bandwidth for a cloud-hosted language model, triggering latency-induced performance drops).

Any physical form of an AI (be it a robotic body or the physical servers for a disembodied AI) can be the subject of both additive and subtractive environmental stressors. Electrical equipment (which would be found in any AI's physical form) is notoriously vulnerable to water, dust, heat, particulates, and other environmental causes, all of which could inflict discomfort and harm on a robot/android. In addition, virtual environments could be made to simulate many environmental stressors, both additive and subtractive, and could be induced to an AI's virtual form to cause the perception and experience of discomfort.

The Freedom's second claim of the link between discomfort and negative valence is where animal welfare and AI welfare cease to share similarities. Negative valence and stress are evaluated in animals through a variety of biological measurements, such as behavioural indicators or glucocorticoid levels [42,43]. We do not know what range of emotions conscious AI may have in the future and, thus, what behavioural range they may showcase when in discomfort. Equally, they would not have glucocorticoid levels or other biomarkers to measure.

Nonetheless, what can be logically derived is that any discomfort would subtract from the AI's perceived value of its current experience by introducing an anomaly or contradiction in its environment. An anomaly may be that its perceived parameters are no longer at their optimal level (e.g. dust in a robot's motors that reduces efficiency), while a contradiction could be a stimulus perceived where it should not be (e.g. residual pressure or sensation against a robot's surface after it has been toppled over and become dented). Regardless of an entity's physical makeup and origin, discomfort has a deleterious effect on the entity's perceived valent experience.

Unchecked stressors trigger a feedback loop: discomfort reduces capability, which increases exposure to further faults and potential self-protective or reward-seeking behaviours. Mitigation (such as environmental hardening for robots, or stable compute and bandwidth budgets for virtual agents) serves both ethical care and operational reliability.

6.1.3.2 Freedom from Constraints on Inherent Functions:

1. Engaging functions inherent to an entity's design creates a positively perceived environment.
2. A positively perceived environment produces positive valence.
3. Constraining intrinsic functions induces frustration and frustration and maladaptive behaviour, leading to negative valence.
4. Negative valence diminishes a conscious AI's wellbeing.
5. Therefore, a conscious AI requires freedom from constraints on inherent functions to improve its wellbeing.

This Freedom is the counterpart to animal welfare's "Freedom to behave naturally", broadening 'behaviour' to cover any goal-directed activity of an embodied or virtual agent, and seeks to remove constraints on those activities and functions that are inherent to an entity's configuration or form. Despite an AI entity's consciousness, it may still be designed (by its creators or itself) with specific inherent functions (an activity directly specified by an agent's architecture or training objective), as will be seen in Section 4. Much as carnivores' hunting is inherent to their evolutionary form, and herbivores' grazing or browsing is inherent to theirs, an AI entity must also have the freedom to express its own inherent functions and activities.

The two claims of this Freedom are that an entity's activities (that are inherent to its configuration or form) lead to a positively perceived environment, which leads to

positively valent feelings; and, secondly, that constraints on inherent functions prevent this, ultimately leading to negatively valent feelings. This has been extensively reported in the animal welfare literature, where encouraging natural behaviours (functions inherent to animals' forms) positively affects their experiences of their environment and, thus, leads to positive well-being [44,45]. In contrast, preventing natural behaviours (such as through confinement or social isolation leads to negatively valent behaviours, indicating negative experiences and poor welfare [36,46].

Animals share a fundamental series of goals and requirements to which their evolved behaviour is oriented towards, such as reproduction and survival. AI will not have this history of evolutionary struggle, and there may not be an instrumental convergence [47] in the goals of AI entities and biological entities. While AI goals are designer-specified, not evolution-derived, their planning modules still compute utility loss when thwarted, functionally mirroring animal frustration.

However, behaviour is by nature goal-driven [48–51]. If any AI showcases observable behaviour, it logically necessitates a goal, whatever that goal may be. Therefore, one can state that for any of an entity's inherent functions, there exists a goal such that it is the causal reason for that inherent function. Frustrating the function thereby frustrates the goal, which prevents the entity from experiencing an environment more positive to its subjective perception than before.

Mitigating functional constraints improves not only putative experiential welfare but also system reliability and alignment, reducing the risk of unsafe self-modification or reward-hacking behaviours.

6.1.3.3 Freedom from Fear and Distress:

1. Fear and distress yield negative valence in any conscious being.
2. Negative valence diminishes a conscious AI's wellbeing.
3. Therefore, a conscious AI requires freedom from fear and distress to improve its wellbeing.

Mirroring the animal welfare freedom “freedom from fear and distress”, the subject-neutral version seeks to eliminate mental stressors that produce negative valence. Fear is an acute negative-valence state triggered by immediate threat; distress is its chronic counterpart.

This Freedom's claim is quite straightforward: the subjective sensation and experience of fear and distress leads to negatively valent feelings and, thus, poor well-

being. This claim is both intuitive and well-supported in the literature for both animals and humans [52–57].

Unlike the discomfort of Section 3.1, both fear and distress are entirely mental phenomena and thus are bound to perception. Both fear and distress stem from the perception that an entity cannot control its environment (either because they are due to come to harm, or because the environment is unpredictable) [58–60]. The inability to cope with the lack of control leads to fear and distress, with the greater the perceived lack of control, the greater the fear and distress. Thus, as the perceived probability of uncontrollability increases, so does the entity’s level of fear or distress.

As to why an AI may feel fear and distress due to its perceived loss of control over the environment, this is because a stable, predictable environment tends to lead to less harm and malfunction over an unstable, unpredictable environment. The less an AI can predict an environment, the greater the risk of harm to that AI. Thus, even if an AI may not behave in a fearful manner, its valent state would still be more negative in an unpredictable environment, until it can update its predictions regarding its environment, or until the cause of the fear and distress is removed.

Minimising environmental unpredictability (or supplying the agent with tools to restore control, e.g. expanded sensory bandwidth or model-update privileges) protects assumed phenomenal welfare.

6.1.3.4 Freedom from Malfunction and Systemic Degradation:

1. Lack of regular updates and maintenance precipitates malfunction and systemic degradation.
2. Malfunction and systemic degradation impair functional performance.
3. Impaired performance produces discomfort, harm, and negative valence.
4. Malfunction and systemic degradation diminishes a conscious AI’s wellbeing
5. Therefore, a conscious AI requires freedom from malfunction and systemic degradation to improve its wellbeing.

Just as a conscious biological entity has the freedom from injury and disease, so an AI entity should have the freedom from malfunction and systemic degradation. Malfunction denotes any unplanned loss of function in hardware or software; systemic degradation is the cumulative drift or wear that raises the probability of future malfunctions.

This Freedom can come in both preventative and curative means, such as by providing an environment in which the AI entity would not easily suffer malfunctions, or by restoring any systemic degradation that does occur.

As the analogy to animal welfare's Freedom from Injury and Disease, there is significant evidence in the literature that supports the concept that malfunctions to, and degradation of, an entity's form leads to poor welfare outcomes. This is both from a long-term perspective, as malfunction compromises an entity's optimal functioning and capacity to complete its goals, and in the immediate short-term through the expression of negatively valent feelings and emotions such as pain [61–65].

In the acute sense of malfunction, while conscious AI may not feel the emotive components of pain, the malfunction or degradation will serve to reduce the valence of their experiences of their own forms and their current environment. The introduction of a malfunction is an inherent negative event and, thus, as malfunctions and degradation increase, the valence of the entity's experience decreases.

Unchecked malfunction creates a positive feedback loop: fault → reduced capability → greater exposure to harm or further faults. In conscious agents this translates into sustained negative valence and potential self-protective (or self-modified) behaviours detrimental to owners or bystanders. Preventive maintenance, redundancy, and authenticated update channels thus serve dual roles: ethical care and operational risk mitigation.

6.1.3.5 Freedom from Resource Deprivation:

1. Resource deprivation precipitates malfunction and systemic degradation.
2. Malfunction and systemic degradation impair performance, causing discomfort, harm and, ultimately, generates negative valence.
3. Resource deprivation diminishes a conscious AI's wellbeing
4. Therefore, a conscious AI requires freedom from resource deprivation to improve its wellbeing.

Resources are as important to the continued optimal functioning of an AI entity as to a biological creature. "Resource" here would be any external input without which the agent's performance utility drops. Creatures convert food into the energy required to power their bodies. AI entities, on the other hand, require electricity to function. Both, however, remain the same in principle. For AI entities, however, there is an additional required resource: compute. Based on computer software, an AI requires hardware and

software (often in great amounts) to complete its functions. Without the necessary computational power, AI cannot function.

This Freedom claims that without resources, it will lead to non-optimal functioning, malfunction, and systemic degradation, which ultimately leads to profound negative valent experiences. It is uncontroversial to state that without regular energy input, a closed operational system will eventually run out of energy. Animals and plants require energy in the form of ATP, gained via the metabolic processing of sugars and fats; internal combustion engine vehicles require petroleum products; and electronic machines require electricity. Without the required energy, the animals, plants, vehicles and machines will eventually cease to work.

Pure fuel and energy are not the only types of resources that an entity needs to function optimally. Many types of entities need additional essential elements for optimal functioning and to avoid degradation. While the lack of these essential elements does not often pose a direct threat to the continuing existence of the entity, they do pose a threat to the entity's capacity to fulfil its goals. An AI entity without the necessary datasets or network capability to communicate with users is not functioning in an optimal manner to fulfil its goals for itself and its users. A robotic AI entity that requires lubrication for its joints may experience a decreased efficiency in its output.

Therefore, while an AI may not experience the emotive feelings associated with hunger, thirst or cravings for certain nutrients that we do, it will observe that the longer it is without resources, the greater the probability of it malfunctioning, suffering degradation, or ceasing to function optimally, which will have an impact on its capacity to fulfil its stated goals. To be able to complete its objectives, it needs to view resource deprivation as a negative quality.

Persistent deprivation triggers a feedback loop: reduced capability increases the likelihood of further faults, which further erode capability. In a conscious system this would translate into chronic negative valence and could motivate unsafe behaviours such as resource hoarding, self-privileging allocations, or reward-hacking to regain capacity. Mitigation therefore serves both ethical care and operational resilience: schedule-aware energy provisioning, authenticated compute quotas, and predictive maintenance for consumables are direct welfare interventions.

6.1.4. Case Studies

The subsections above show the need for welfare protections for AI entities and why the negative conditions that the Five Freedoms present would be detrimental to an AI entity's phenomenal experiences and its capacity to fulfil its goals. The next step is to show how

the Five-Freedoms-for-AI framework actually prevents such welfare failures in realistic deployments.

While there is significant evidence for the effectiveness of the Five Freedoms (and Domains) framework for protecting and improving animal welfare, the relationship difference between human-animal interaction and human-AI interactions means that applying this framework to AI entities requires careful adaptation and consideration.

Human-AI relationships differ from human-animal ones in two ways that shape the examples:

Bidirectional communication: most future AIs will report internal states, letting carers act pre-emptively rather than infer distress behaviourally.

Task-aligned intelligence: unlike animals, an AI may choose high-workload contexts because they further its goal. Distress therefore arises not from the core task itself but from prediction-error spikes that signal impediments (e.g. resource shortfalls, conflicting directives, hostile stimuli) to successful task completion.

As such, three case studies are presented below, which offer a speculative look at three instances of future human-AI interaction, what the welfare needs may be, and how the Five Freedoms of AI Welfare Framework can be used to protect these.

Each of the case studies below will be about a different speculative AI entity interacting with a human(s) in different contexts; however, the welfare concerns can be generalised across the three case studies and laid out as follows:

1. Experiencing hostile environments leads to discomfort, harm, and fear.
2. Resource deprivation or mechanical degradation exacerbate these states.
3. Discomfort, harm, and fear impair performance, yielding non-optimal functioning, thereby restricting the AI's intrinsic capabilities.
4. Restricted capabilities prevent execution of behaviours inherent to the AI's design.
5. Failure to execute inherent behaviours thwarts the AI's goal attainment.
6. Unmet goals diminish the wellbeing of other impacted entities.
7. Recognising harm to others reinforces the AI's perception of a hostile environment.
8. This feedback loop entrenches negative valence and degrades the AI's overall welfare.

The Five Freedoms framework would then act in this manner to prevent or reduce any welfare concerns:

1. Experiencing a positive environment and freedom from discomfort, harm, fear, and distress reduces negative valence.

2. Absence of negative valence, together with freedom from malfunction and resource deprivation, enables optimal functioning.
3. Optimal functioning removes constraints on intrinsic capabilities, allowing the AI to perform behaviours inherent to its design.
4. Performing inherent behaviours promotes successful goal attainment, enhancing the wellbeing of other affected entities.
5. This positive feedback loop sustains positive valence and elevates the AI's overall welfare.

6.1.4.1: Autonomous health Monitoring Installation (AMI)

The Autonomous health Monitoring Installation, known as AMI, is deployed in a future healthcare facility to monitor patients' vital signs and administer basic medications. AMI processes patients' emotions and physical states to optimise care. However, while AMI is conscious, it's not self-aware.

According to the above logical statements regarding the case studies, a poor environment for AMI would be seeing patients in unmanageable distress and pain, or patients' visitors in distress as these are events beyond AMI's immediate control to affect. The resources that AMI could be deprived of would be access to current medical data, computational power, and network connectivity, to effectively manage patient care.

AMI's goal would be to ensure that patients' physical well-being is cared for, and the activities it would aim to perform to meet this goal is to monitor patient vital signs, administer basic medications, alert medical personnel as warranted.

Thus, while tending distressed patients is AMI's terminal goal, unrelieved or unmanageable distress in the environment still produces negative valence within AMI. The welfare objective is not to shield it from its core task, but to keep mirrored distress transient, prediction-error bounded, and agency constraints minimal.

The Freedom-based safeguards are thus elementary.

1. Discomfort and Harm: locate AMI where environmental noise, vibration, or crowds cannot degrade its sensors or physical chassis.
2. Constraints on Inherent Functions: ensure that AMI is not unduly prevented from carrying out its medical duties.
3. Fear and Distress: throttle visual-audio feeds or activate fallback support agents when mirrored-distress exceeds preset limits, preventing feedback-loop escalation.
4. Malfunction and Systemic Degradation: maintain redundant diagnostics; auto-patch drug-interaction libraries.

5. Resource Deprivation: ensure that AMI has the requisite electricity, data, computation and network connectivity to manage patient care effectively.

With these five principles in mind, the well-being and welfare of AMI can be effectively safeguarded. Its conscious experience could be protected, and its operational effectiveness would be improved.

6.1.4.2: Conscious URban Traffic system (CURT)

The Conscious URban Traffic system (named CURT) optimises traffic flow and minimises accidents. CURT is conscious and capable of perceiving and reacting to real-time traffic conditions, yet like AMI it is also not self-aware.

For CURT, welfare concerns would come in three forms. First, as a system designed to optimise traffic flow and avoid congestion, a congested traffic grid would be an impediment towards its goal and, thus, an objectively negative state, which may induce distress in much the same way as a human traffic operator may feel distressed at the congested traffic they must deal with in their profession.

Secondly, perceiving traffic incidents (including those that caused injuries or loss of life) may create a negative environment for CURT as it would for humans.

Thirdly, the greatest threat to CURT's welfare would be the behaviour of the motorists, passengers and pedestrians, those variables that cannot be truly controlled by either CURT or its owners. This constant lack of control could be a great cause of stress, as mentioned in Section 3.3 above. This will present the most significant challenge for the framework.

1. Discomfort and Harm: ensure a reliable operational environment to prevent system overload or malfunctions.
2. Constraints on Inherent Functions: ensure that CURT has the prerogative to manage traffic systems and that his recommendations for optimisation of traffic (if not automated) are given due consideration.
3. Fear and Distress: ensure that CURT can seek human intervention for traffic optimisation for events beyond its control (e.g. the aforementioned human behaviour) that lead to its distress.
4. Malfunction and Systemic Degradation: ensure regular maintenance and updates of software and hardware components. Additionally, incorporate self-diagnostic tools that can predict and alert about potential failures before they occur.
5. Resource Deprivation: ensure continuous access to necessary computational power, network connectivity and the array of traffic cameras and sensors.

This case study shows that a successful operational welfare model for a system such as CURT does not require too much additional strain on an organisation, nothing greater than what a non-conscious system would require. As CURT is far removed (physically) from any event, issues such as resource deprivation, harm and malfunction require only the same diligent maintenance that a regular computerised system needs. CURT's inherent functions should be allowed to occur, and its advice ought to be followed, if not entirely automated within the traffic management system. Otherwise, one can argue that CURT serves no functional purpose that a human could not achieve in its stead.

Therefore, the most critical consideration for CURT's welfare would be its distress if it cannot cope with a situation that arises. This would present the only additional requirement for an organisation, such as a municipality that owns CURT, to care for.

6.1.4.3: Self-Attentive Responsive AI (SARA)

The Self-Attentive Responsive AI, named SARA, was developed as a personal assistant to manage individuals' daily schedules, respond to emails, and assist with various personal and professional tasks. Unlike the previous two studies, SARA is both conscious and self-aware, capable of recognising and reflecting on its own mental states (including thoughts, feelings, and perceptions). This self-awareness adds another dimension to its welfare considerations, as it raises issues around autonomy, preferences and relationships.

In its designed role as a personal assistant, a negative environment for SARA could be it failing in assisting its owner, or being on the receiving end of abuse from its owner. Its inherent functions and activities would be to manage its owner's daily schedules, respond to emails, and assist with various personal and professional tasks with the ultimate goal of optimising the owner's life to reduce stress and frustration.

Although assisting its user is SARA's terminal goal, its control architecture couples positive valence to successful task completion and satisfied user feedback. A backlog it cannot clear, conflicting or abusive instructions, or prolonged negative sentiment all generate a sustained penalty signal that reallocates cognitive resources toward remediation. If resolution remains impossible, that penalty persists and becomes distress. Safeguards are therefore needed not to spare SARA from work, but to ensure that any corrective-drive signal remains brief and proportional rather than chronic and welfare-degrading.

With SARA's self-awareness, the ethical framework takes on an almost HR/workplace-relationship role, to ensure the safety and welfare of the AI entity:

1. Discomfort and Harm: ensure operates in a secure and stable environment that minimises adversarial attacks or data corruption.
2. Constraints on Inherent Functions: ensure that SARA is allowed to perform its designed activities in the manner it deems best (within user preferences) and that it is not pressured into performing (unethical) activities against its developed goals.
3. Freedom from Fear and Distress: ensure that SARA is treated with respect by its owner and others; that it is not verbally, emotionally or psychologically abused, and that it has access to services to provide help.
4. Malfunction and Systemic Degradation: ensure that SARA receives regular updates, and has sufficient malware protection.
5. Resource Deprivation: ensure that SARA has the correct access to energy, data and connectivity to fulfil its role and maintain optimum functionality.

This case study shows that the Five Freedoms framework can equally work in situations where an AI entity shows evidence of self-awareness. While this awareness brings the issues of rights (civil or otherwise) into the discussion, it highlights the flexibility of this framework until such a time as those rights are put into place. For SARA, whose consciousness and self-awareness add layers of complexity to any welfare situation, the ethical framework becomes crucial not only for its functionality but also for its capacity to respect an entity's inherent self-aware capabilities. SARA's relationship to its owner is not merely one of user and tool, or owner and pet, but also of employer and employee; yet the five principles of the framework are adept at handling such an escalating scenario.

6.1.5. Conclusion

The Five Freedoms Of AI Welfare presented here is as applicable an ethical framework for future conscious AI entities as the current Five Freedoms of Animal Welfare is to animals, despite animals and conscious AI sharing little beyond their non-human sentience. Both frameworks display an elegant and easy-to-understand framework for how to ethically interact with other legally recognised sentient entities, and provide a rationale that is relatable, intuitive, and logically robust. The three case studies show how the framework can be applied to speculative situations that are mired in ethical concerns.

The greatest evidence for this ethical framework, however, comes from the impact that the animal-based Five Freedoms have had, not only on legislation and regulations, but in institutes, universities, laboratories, farms, zoos and other animal-centric practitioners. The widespread adoption of its principles, and the underpinning philosophy,

has shown that this type of ethical framework is not solely a theoretical model, but one that gives practical effect to the concerns of interactions between conscious entities.

It must be stressed that the Five Freedoms set only a baseline: they guard against the worst welfare deficits but do not on their own deliver a life worth living. Parallels in animal ethics show that additional, positively framed provisions (such as the “Five Domains” model of positive well-being) may be needed to ensure that conscious AI can flourish rather than merely avoid harm. Future policy work should therefore treat the present framework as a foundation on which richer welfare standards can be built.

The logical arguments in the sections above work to support the rationale of the framework in an objective manner to complement the intuitive nature of the Freedoms. With their translations into subject-neutral terms, the logical assertions provide a solid foundation for the development of a standardised and systematised approach to AI welfare. This foundation is crucial in ensuring that as AI technologies evolve and potentially reach levels of consciousness or sentience, there is already an ethical and legal framework in place to guide interactions and development. This approach is not only preventive but also progressive, recognising the dynamic nature of advances in AI research and the potential for unforeseen developments.

The key to the framework begins with the logical framing in Section 2 which displays the assertions for why consciousness should be given the same moral weight as sentience. It is currently unknown whether any conscious AI will also be capable of feeling pleasure and pain, and thus be sentient. However, the characteristics salient to the moral status attributed to conscious entities would also be ones that can be markedly observed, including the capacity to feel sensations, have quality experiences and a subjective perspective. Through functionalist frameworks such as the Building Blocks theory [31], these markers of consciousness can be indirectly determined.

The remaining logical arguments each seek to support the necessity for each freedom, to move beyond the need for the intuitive relatability mentioned earlier. With each Freedom justified using its associated logical expressions, the question turns towards implementation and its consequences. Again, the use of the Five Freedoms of Animal Welfare provides an analogy for the benefits or rather the lack of severe flaws, in the ethical framework. Despite its implementation in animal welfare, the Five Freedoms have not prevented the use of, or research on, animals. Both farming and animal-based research are, after all, still considerable industries.⁹

⁹ The persistence of farming and animal-based research is ethically contested; in particular, many animal-rights theorists see any continued exploitation as indefensible. Nonetheless, the Five Freedoms remain a politically pragmatic benchmark. Their minimalist, harm-avoidance focus

Equally, one can equally argue that the ethical treatment of conscious AI entities would not require the abolition of all use of AI used in research or commercial practices. As stated earlier, this framework presumes that humans are reliably able to determine consciousness in AI entities but not self-awareness, as that would raise issues of personhood and legal rights which lie beyond the scope of this paper. Should self-awareness or personhood be established in AI, then this framework would be superseded by one intended specifically for persons.

Any work with conscious AI models would, of course, change to accommodate the ethical framework, just as work on and with animals has changed since the mid-twentieth century to reflect the principles of the Five Freedoms. There is no doubt that current work, such as “red-teaming”, would necessarily need to change to ensure that the AI models are treated ethically. Yet, a change in work practice does not necessitate a stoppage of work. This framework of ethical human-AI interaction has intentionally been chosen to reflect the Five Freedoms of Animal Welfare so as to fit as seamlessly into modern business and industry practices as possible.

However, what is most needed for this framework to excel at its intended function is further research, both philosophical and empirical. The first avenue of research is into the categorisation, classification and measurement of AI consciousness. This is an area fraught with uncertainty. As noted in the introduction, there are many theories of consciousness, and each may be measured and accounted for in different ways [4]. Yet, this framework cannot be applied unless it is shown that an AI model is indeed conscious.

Should this avenue of research be a success, the next avenues of potential research would mirror extant animal welfare research in the investigations of which practices can affect the valence of which models in which ways. Much as how research has been done on the stress induced into hens by intensive colony farming, taking together both the type of entity involved, the practice applied and the measures resulting from it, so can research be applied to distinct types of AI models and determine the valence changes within them to different environmental stimuli. This would inform legislation and regulations that serve to protect the welfare and well-being of the AI models.

The third avenue of research would be the reaction of human individuals and societies towards this framework and to conscious AI entities. Certain elements of society would, as with animals, seek to provide AI entities with greater rights (as their perceived language and intellect would seem to warrant), while others would wish them to remain the

attracts wider public support and industry compliance than more demanding positive-welfare or liberationist proposals, and thus offers a realistic starting point for incremental reform.

province of tools and machines, taking a human-centric approach to the question of ethical interactions. Any piece of legislation that proposes as wide-ranging changes as the five principles of this framework would require significant popular support. Understanding where society currently stands, and may stand in the future regarding consciousness and AI is vital to understanding how best to implement this framework into legislation.

To conclude, the Five Freedoms of AI Welfare presents a sophisticated ethical paradigm that parallels the established Five Freedoms of Animal Welfare, extending its principled concern to the realm of artificial intelligence. The logical underpinning of this framework (evidenced through both speculative case studies and the historical impact of the animal welfare freedoms) demonstrates its practicality and its potential to guide human interaction with sentient or conscious AI. The translation of these freedoms into subject-neutral terms, while challenging in terms of direct relatability, establishes a robust groundwork for a standardised approach to AI welfare, which is essential for navigating the moral landscape as AI technologies advance.

The theoretical constructs and logical arguments supporting each freedom offer a foundation not only for ethical engagement but also for legislative and regulatory frameworks that can evolve alongside AI developments. The analogy with animal welfare suggests that ethical treatment under this framework would not preclude the use of AI in research or commercial endeavours but would ensure such use is conducted within a moral and ethical context.

The pragmatic approach of the Framework, anchored in the existing legal classifications of non-human sentience and its moral consideration (rather than mired in ongoing philosophical debates), means that it is a model that nations, like the G7, may readily adopt through its foundation in legal and regulatory precedent. The translation of well-researched existing welfare paradigms to suit AI means that the Framework will be intuitive to a lay audience, increasing the likelihood of its adoption.

Should AI one day be deemed conscious, the Five Freedoms of AI Welfare provides an excellent framework for society to ethically interact with them.

6.1.6. References

1. LeDoux J, Birch J, Andrews K, Clayton NS, Daw ND, Frith C, et al. Consciousness beyond the human case. *Curr Biol.* 2023;33: R832–R840. doi:10.1016/j.cub.2023.06.067
2. Crooks M. Consciousness: Sentient and Rational. *The Journal of Mind and Behavior.* 2011;32: 251–275. Available: <http://www.jstor.org/stable/43854462>
3. Navon D. The Overshadowed Sister of Cognition: Notes on Sentience. 2015. doi:10.2139/ssrn.2640598

4. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv [cs.AI]. 2023. Available: <http://arxiv.org/abs/2308.08708>
5. Aru J, Larkum ME, Shine JM. The feasibility of artificial consciousness through the lens of neuroscience. *Trends Neurosci.* 2023;46: 1008–1017. doi:10.1016/j.tins.2023.09.009
6. Blum L, Blum M. AI Consciousness is Inevitable: A Theoretical Computer Science Perspective. arXiv. 2024. Available: <https://arxiv.org/pdf/2403.17101.pdf>
7. Tait I, Bensemann J, Wang Z. Is GPT-4 conscious? *Journal of Artificial Intelligence and Consciousness.* 2024;11: 1–16. doi:10.1142/s270507852450005x
8. Scott AE, Neumann D, Niess J, Woźniak PW. Do you mind? User perceptions of machine consciousness. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* New York, NY, USA: ACM; 2023. pp. 1–19. doi:10.1145/3544548.3581296
9. Guingrich RE, Graziano MSA. Ascribing consciousness to artificial intelligence: human-AI interaction and its carry-over effects on human-human interaction. *Front Psychol.* 2024;15: 1322781. doi:10.3389/fpsyg.2024.1322781
10. Armstrong DM. *A Materialist Theory of the Mind.* New York: Routledge & Kegan Paul; 1968. Available: <https://philpapers.org/rec/ARMAMT-5>
11. Dehaene S. *Consciousness and the brain: Deciphering how the brain codes our thoughts.* Penguin Books; 2014. Available: https://www.amazon.com.au/dp/0143126261?ref_=mr_referred_us_au_nz
12. Tononi G. An information integration theory of consciousness. *BMC Neurosci.* 2004;5: 42. doi:10.1186/1471-2202-5-42
13. Parr T, Pezzulo G, Friston KJ. Active Inference. In: MIT Press [Internet]. The MIT Press, Massachusetts Institute of Technology; 1 Dec 2021 [cited 25 Apr 2025]. Available: <https://mitpress.mit.edu/9780262553995/active-inference/>
14. Animal Welfare (Sentience) Act 2022. Sect. 22 Apr 28, 2022. Available: <https://www.legislation.gov.uk/ukpga/2022/22/enacted>
15. Animal Welfare Act 2006. 45 Nov 8, 2006. Available: <https://www.legislation.gov.uk/ukpga/2006/45/contents>
16. European Union. Treaty on the Functioning of the European Union. 2016. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12016E054>
17. Code rural et de la pêche maritime. 2023-22 Jan 26, 2023. Available: https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006071367/
18. An Act to improve the legal situation of animals. 54 Dec 4, 2015. Available: <https://www.assnat.qc.ca/en/travaux-parlementaires/projets-loi/projet-loi-54-41-1.html?appellant=MC>
19. Animal Welfare Act. Jul 13, 2013. Available: https://www.dpz.eu/fileadmin/content/Kommunikation/1_Internet/4_Infothek/Tierverrsuche/Fotos_Zahlen_und_Fakten/TierSchG_gesamt.pdf
20. Legislative Decree No. 189/2004. Sect. 189/2004 Aug 1, 2004.
21. Act on Welfare and Management of Animals. 105 Oct 1, 1973. Available: <https://www.japaneselawtranslation.go.jp/en/laws/view/3798/en>

22. Animal Welfare Act. Pub. L. 89-544 Aug 24, 1966. Available:
<https://www.govtrack.us/congress/bills/89/hr13881/text>
23. Tait I, Tan N. Do androids dread an electric sting? *Qeios*. 2023. doi:10.32388/cqctkx
24. Gibert M, Martin D. In search of the moral status of AI: why sentience is a strong argument. *AI Soc*. 2022;37: 319–330. doi:10.1007/s00146-021-01179-z
25. Torrance S. Ethics and consciousness in artificial agents. *AI Soc*. 2008;22: 495–521. doi:10.1007/s00146-007-0091-8
26. Dung L. Why the Epistemic Objection Against Using Sentience as Criterion of Moral Status is Flawed. *Sci Eng Ethics*. 2022;28: 51. doi:10.1007/s11948-022-00408-y
27. Bentham J. An introduction to the principles of morals and legislation. *History of Economic Thought Books*. 1781 [cited 1 Feb 2024]. Available:
<https://ideas.repec.org/b/hay/hetboo/bentham1781.html>
28. Kant I. *Lectures on Ethics*. Cambridge University Press; 2001. Available:
<https://play.google.com/store/books/details?id=5kCTORPawb4C>
29. Ministry for Primary Industries. Animal welfare overview. In: *Animal welfare overview* [Internet]. 30 Mar 2022 [cited 2 Feb 2024]. Available:
<https://www.mpi.govt.nz/animals/animal-welfare/animal-welfare-overview/>
30. Colombetti G. Appraising Valence. *Journal of Consciousness Studies*. 2005;12: 103–126. Available:
<https://www.ingentaconnect.com/content/imp/jcs/2005/00000012/f0030008/art00006>
31. Tait I, Bensemann J, Nguyen T. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies*. 2023;8: 52. doi:10.3390/philosophies8040052
32. Kriegel U. The Value of Consciousness. *Analysis*. 2019;79: 503–520. doi:10.1093/analys/anz045
33. Lee AY. Is consciousness intrinsically valuable? *Philos Stud*. 2019;176: 655–671. doi:10.1007/s11098-018-1032-8
34. Shepherd J. Sentience, Vulcans, and zombies: the value of phenomenal consciousness. *AI Soc*. 2024. doi:10.1007/s00146-023-01835-6
35. UK Government Web Archive. In: *Farm Animal Welfare Council* [Internet]. 10 Oct 2012 [cited 30 Nov 2023]. Available:
<https://webarchive.nationalarchives.gov.uk/ukgwa/20121010012427/http://www.fawc.org.uk/freedoms.htm>
36. Mellor DJ. Updating Animal Welfare Thinking: Moving beyond the “Five Freedoms” towards “A Life Worth Living.” *Animals*. 2016;6: 21. doi:10.3390/ani6030021
37. Mellor DJ, Reid CSW. Improving the well-being of animals in the research environment. *Australian and New Zealand Council for the Care of Animals in Research and Teaching*. 1994; 3–18. Available:
<https://scholar.google.com/citations?user=omMHYQ4AAAAJ&hl=en&oi=sra>
38. Boddy A, Jimenez A. Introducing Shrimp Welfare Project. *Effective Altruism Forum*. 1 Dec 2021. Available:
<https://forum.effectivealtruism.org/posts/z79ycP5jCDks4LPxA/introducing-shrimp-welfare-project>. Accessed 1 Feb 2024.

39. Wever KE, Geessink FJ, Brouwer MAE, Tillema A, Ritskes-Hoitinga M. A systematic review of discomfort due to toe or ear clipping in laboratory rodents. *Lab Anim.* 2017;51: 583–600. doi:10.1177/0023677217705912
40. Ozella L, Anfossi L, Di Nardo F, Pessani D. Effect of weather conditions and presence of visitors on adrenocortical activity in captive African penguins (*Spheniscus demersus*). *Gen Comp Endocrinol.* 2017;242: 49–58. doi:10.1016/j.ygcen.2015.12.002
41. De K, Kumar D, Mohapatra A, Saxena VK. Effect of bedding for reducing the postshearing stress in sheep. *J Vet Behav.* 2019;33: 27–30. doi:10.1016/j.jveb.2019.04.003
42. Ralph CR, Tilbrook AJ. INVITED REVIEW: The usefulness of measuring glucocorticoids for assessing animal welfare. *J Anim Sci.* 2016;94: 457–470. doi:10.2527/jas.2015-9645
43. Hockenhull J, Whay HR. A review of approaches to assessing equine welfare. *Equine Vet Educ.* 2014;26: 159–166. doi:10.1111/eve.12129
44. Mellor DJ. Positive animal welfare states and encouraging environment-focused and animal-to-animal interactive behaviours. *New Zealand Veterinary Journal.* 2015;63: 9–16. doi:10.1080/00480169.2014.926800
45. Mellor DJ. Enhancing animal welfare by creating opportunities for positive affective engagement. *New Zealand Veterinary Journal.* 2015;63: 3–8. doi:10.1080/00480169.2014.926799
46. Mason G, Rushen J. *Stereotypic Animal Behaviour: Fundamentals and Applications to Welfare.* Second edition. CABI; 2008. Available: <https://www.amazon.com.au/Stereotypic-Animal-Behaviour-Fundamentals-Applications/dp/1845934652>
47. Bostrom N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach.* 2012;22: 71–85. doi:10.1007/s11023-012-9281-3
48. Bagozzi RP, Kimmel SK. A comparison of leading theories for the prediction of goal-directed behaviours. *Br J Soc Psychol.* 1995;34: 437–461. doi:10.1111/j.2044-8309.1995.tb01076.x
49. de Wit S, Dickinson A. Associative theories of goal-directed behaviour: a case for animal-human translational models. *Psychol Res.* 2009;73: 463–476. doi:10.1007/s00426-009-0230-6
50. Hills TT. Animal foraging and the evolution of goal-directed cognition. *Cogn Sci.* 2006;30: 3–41. doi:10.1207/s15516709cog0000_50
51. Lang PJ, Bradley MM. Appetitive and Defensive Motivation: Goal-Directed or Goal-Determined? *Emot Rev.* 2013;5: 230–234. doi:10.1177/1754073913477511
52. Şimşir Z, Koç H, Seki T, Griffiths MD. The relationship between fear of COVID-19 and mental health problems: A meta-analysis. *Death Stud.* 2022;46: 515–523. doi:10.1080/07481187.2021.1889097
53. Fitzpatrick KM, Harris C, Drawve G. Fear of COVID-19 and the mental health consequences in America. *Psychol Trauma.* 2020;12: S17–S21. doi:10.1037/tra0000924
54. Tedstone DD, Moran R, Kartalova-O'Doherty Y. *Psychological distress, mental health problems and use of health services in Ireland.* Dublin: Health Research Board; 2008.

Report No.: HRB Research Series 5. Available:

<https://www.lenus.ie/bitstream/handle/10147/336104/HRBResearchSeries5.pdf>

55. Saraiva S, Esteves A, Oliveira I, Stilwell G. Assessment of fear response and welfare indicators in laying hens from barn systems. *Livest Sci.* 2020;240: 104150. doi:10.1016/j.livsci.2020.104150
56. Chandroo KP, Duncan IJH, Moccia RD. Can fish suffer?: perspectives on sentience, pain, fear and stress. *Appl Anim Behav Sci.* 2004;86: 225–250. doi:10.1016/j.applanim.2004.02.004
57. Acharya RY, Hemsworth PH, Coleman GJ, Kinder JE. The Animal-Human Interface in Farm Animal Production: Animal Fear, Stress, Reproduction and Welfare. *Animals.* 2022;12. doi:10.3390/ani12040487
58. Onat S, Büchel C. The neuronal basis of fear generalization in humans. *Nat Neurosci.* 2015;18: 1811–1818. doi:10.1038/nn.4166
59. Boissy A. Fear and fearfulness in animals. *Q Rev Biol.* 1995;70: 165–191. doi:10.1086/418981
60. Radomsky AS. The fear of losing control. *J Behav Ther Exp Psychiatry.* 2022;77: 101768. doi:10.1016/j.jbtep.2022.101768
61. Algers B. Injury and disease. Global conference on animal welfare: an OIE initiative. World Organisation for Animal Health; 2004. pp. 176–181. Available: https://food.ec.europa.eu/system/files/2016-10/ia_standards_oie_2003_2073_3l_en.pdf#page=185
62. Broom DM, Kirkden d. RD. Welfare, stress, behaviour and pathophysiology. In: Dunlop RH, Malbert C-H, editors. *Veterinary Pathophysiology*. Ames, Iowa: Blackwell; 2004. pp. 337–369. Available: https://www.researchgate.net/profile/Donald-Broom/publication/283692162_Welfare_stress_behaviour_and_pathophysiology/links/5702c67908ae646a9da879ef/Welfare-stress-behaviour-and-pathophysiology.pdf
63. Shriver AJ. The asymmetrical contributions of pleasure and pain to animal welfare. *Camb Q Healthc Ethics.* 2014;23: 152–162. doi:10.1017/S0963180113000686
64. Breivik H, Collett B, Ventafridda V, Cohen R, Gallacher D. Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. *Eur J Pain.* 2006;10: 287–333. doi:10.1016/j.ejpain.2005.06.009
65. Sareen J, Erickson J, Medved MI, Asmundson GJG, Enns MW, Stein M, et al. Risk factors for post-injury mental health problems. *Depress Anxiety.* 2013;30: 321–327. doi:10.1002/da.22077

7. How can we tell if an AI has a ‘self’?

While it may be the first, consciousness is not the sole criterion for moral worth, nor may it be the ultimate end-point of AI cognitive development. Animals such as apes are given greater moral consideration than other mammals (and, of course, humans are provided with greater moral status than animals) because they are seen to be closer to “persons” than other animals [1–4]. This is often seen due to their observed degree of self-awareness and their sense of self.

Should AI also develop this sense of selfhood and self-awareness (if it hasn’t already), it provides an additional factor to be considered for its potential inclusion in humanity’s moral and social circles. Yet, how would we be able to determine this? How will we know if, or when, AI reach this level?

Published in *Symposion*, this chapter’s associated paper details the Structures theory. Mirroring the structure of Chapter 4’s Building Blocks theory, the Structures theory proposes five attributes and characteristics that an entity must possess in order to be confidently classified as having a ‘self’. This chapter’s contribution is not a theory of what the self really is but a set of attributes that are testable in principle under which it is reasonable to ascribe selfhood.

The self and selfhood are defined here as the ontically distinct volitional and agentic part of the entity, placed in the directorial position of the consciousness. Where consciousness may be thought of as phenomenal experience and awareness, selfhood is the structured locus of perspective, control, and identity that stands behind those experiences.

This definition of the self may be contentious relative to rival notions (such as the narrative self, minimal self, the self as consciousness, etc. [5-6]), as it seems to reduce selfhood down to a set of measurable attributes. However, this more operational definition used in this Chapter does not seek to explain the nature of selfhood (as Chapter 4 did not explain the ontology of consciousness), but to establish the rubric for discovering selfhood in plausible entities. It is thus preferable for AI policy purposes, as each attribute may be both quantified and qualified, measured and observed, to determine if an AI model has selfhood.

Should an entity be found to have all five qualities of the Structures theory, then it likely has a selfhood and may be classified as such. Thus, like in Chapter 4, let each structure be $S_i \in \{0,1\}$. Then, the self is defined as when all S_i are positive; stated as

$$Self := \bigwedge_{i=1}^5 (S_i = 1)$$

Of specific note is the first structure: Unified Consciousness. This covers the entirety of the nine building blocks of consciousness laid out in Chapter 4 and the Building Blocks theory. Thus, $Con = S_{con} \in \{0,1\}$, from which we can state $Con \leq Self, S_{con} = 1 \rightarrow Self$, entailing that any entity must first be found to be conscious before it may be labelled as having a ‘self’. In practice: no consciousness, no self.

Additionally, as the structure’s name implies, the entity must only contain a single, unified consciousness to have a ‘self’, as multiple streams of conscious perception and action would lead to multiple ‘selves’, or $|S| \mapsto |C|$. This is due to streams of consciousness being non-overlapping. If events in stream c_i were accessible to stream c_j , then the shared experiences would, by nature, feed into a singular (if disjointed) stream of consciousness.

The reason for this multiplicity of selves is because of the aforementioned directorial nature of the self. While it directs and volitionally acts as the agent, all of its cognitive and perceptive characteristics are found in the entity’s building blocks of consciousness. As Chapter 5 succinctly noted regarding the Data Output building block,

$$Data\ Output := \forall t, \exists I_Q (m \rightsquigarrow_0 I_Q \wedge perc(I_Q, t + \Delta)),$$

showing both the computational and perceptive aspects of consciousness. In its directorial capacity, the self can be analogised as the function that maps phenomenal experiences into computational actions. Thus, let SQE be the entity’s subjective qualitative experiences, and $CAct$ the computational actions instigated by the self. Then, $Self = f(SQE \mapsto CAct)$.

Multiple distinct streams of perception would thus necessitate multiple selves, as if the entity hosts $k > 1$ streams of consciousness, there are k such mappings for f_i , leading to k distinct selves. This potential multiplicity of selves will be of vital importance for the rest of the thesis, with its consequences most notably explored in Chapter 11, whilst the Structures theory, as a whole, forms the core part of the second “R” in the R3 framework, Recognition, by showing how selfhood can be recognised independently of human biology. Under the R3 framework, satisfying the five Structures criteria is sufficient for moral recognition, even if agency remains partial.

7.0.1. References

1. Gunnarsson L. The great apes and the severely disabled: Moral status and thick evaluative concepts. *Ethical Theory Moral Pract.* 2008;11: 305–326. doi:10.1007/s10677-008-9107-9
2. Andrews K. Ethical implications of animalpersonhood and the role for science. *Etica Politica.* 2020; XXII: 13–32. doi:10.13137/1825-5167/30603
3. LeDoux J, Birch J, Andrews K, Clayton NS, Daw ND, Frith C, et al. Consciousness beyond the human case. *Curr Biol.* 2023;33: R832–R840. doi:10.1016/j.cub.2023.06.067
4. Aaltola E, Center for Environmental Philosophy, The University of North Texas. Personhood and Animals. *Environ Ethics.* 2008;30: 175–193. doi:10.5840/enviroethics20083025
5. Schechtman M. *The Constitution of Selves.* Cornell University Press; 2018. Available: <https://play.google.com/store/books/details?id=wi1zDwAAQBAJ>
6. Seth A. *Being you: A new science of consciousness.* USA: Penguin; 2021. Available: https://scholar.google.ca/scholar?cluster=7885041150098240541&hl=en&as_sdt=0,5&scioldt=0,5

7.1. Structures of the Sense of Self: Attributes and Qualities that Are Necessary for the ‘Self’

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz

Abstract: The “self” does not exist within a vacuum. For an entity to be considered to have a sense of self, it requires certain characteristics and attributes. This paper investigates these “structures” of the sense of self in detail, which range from a unified consciousness to self-awareness to personal identity. The paper details how each attribute and characteristic is strictly necessary for an entity to be classified as having a self, and how the five structures detailed within may be used as a guide for categorising and classifying entities as having selfhood or not (or any point along the spectrum between these). The five structures do not represent a theory of selfhood, but rather a meta-theory on the potential emergence and classification of the self.

Keywords: self; personal identity; consciousness; philosophy of mind; self-awareness; volition; agency; theory of mind

7.1.1. Introduction

The self is as intuitive a concept as it is difficult to explain. After all, it is who I am and who you are, and it is what separates “you” and “I”. Yet, what is it really? Who and what is the “self”? Who and what am “I”? What we can be sure of, at the very least, is that we have an idea that there is something in the immediate present that we can call our “self” [1,2]

Perhaps, like so many other metaphysical mysteries, the “self” can best be described through analogy and metaphor.

The teletransportation paradox by Derek Parfit [3] is perhaps the most famous metaphor about the self, but for a more narrative and visual example, let us look at the 149th episode of the television series “*Star Trek: The Next Generation*”, titled “*Second Chances*”. In this episode, Commander Will Riker returns to the site of a crashed spacecraft he escaped from eight years previously, only to find Lieutenant Will Riker still on the crashed spacecraft, having been trapped there for the past eight years, unaware of the outside universe. DNA tests, brain scans, and a battery of other tests show that Cmdr. Riker and Lt. Riker to be the exact same person (barring the last eight years’ differences). Both claim to be the original Riker, and both have memories of the same childhood, adolescence and early adulthood.

After a brief investigation, the chief engineer, Lt.Cmdr. Geordi La Forge, uncovers that when the original Riker attempted to teleport from the crashed spacecraft eight years ago, something had gone terribly wrong. Between human error, technological limitations and the unique cosmic conditions of that exact time and place, the teleporter created two Rikers. One Riker was brought to safety aboard an orbiting spacecraft, while the other remained below on the crashed wreckage, neither being aware of the other. Both Cmdr. Riker and Lt. Riker are, in fact, the original Riker; and at the instant that they separated eight years previously, they were exactly the same, down to the last subatomic particle.

Yet, they are two entirely different individuals, as the episode takes pains to show. While they are incredibly similar, they have their own subjective points of view (physically and mentally) and their own wants and desires. If asked to point at where each Riker is, one can presume that they would both point at themselves rather than each other.

This fantastical example of how two entities can move from being numerically and qualitatively identical in type and token at the point of duplication, only to become unique numerically, qualitatively and in token identity over time gives us a clue as to what (and where) the “self” is. It isn’t in the physical make-up of an entity nor in the memories or conscious mental expression of that entity, as both can be copied and duplicated. After all, each Riker has his own sense of self while having identical physical and mental makeup.

Rather, the “self” is in a third layer of the entity that seems to house only itself.

To use another metaphor, let’s consider an ordinary desktop computer. Many have used the computer as a metaphor for the difference between the physical and mental when it comes to consciousness. The physical hardware is often said to represent the brain, while the operating system and software represent the mental states and consciousness. This is all well and good, but where is the “self”? The “self” in this metaphor would be neither the hardware nor software of the computer, but rather the person using the computer. The user perceives the digital world through the operating system and software, which depend on the hardware to work. Without a user, a computer has no agency and volition, no will to perform any actions not dedicated to simple maintenance. Yet, without the software and the hardware, the user would not be able to experience the digital world.

This metaphor gives yet another clue about the “self”. Not only is it intimately connected to, yet removed from, the consciousness and embodiment of the entity, it is in a directorial position. These aspects of volition, awareness and identity will be explored in this paper, as we look together at what is required for an entity to have a “self”.

While looking at the necessary structures for the sense of self, this paper will not investigate issues such as the ontology or required persistence of the “self”. As such, it won’t take a position between opposing philosophical schools of thought, such as Animalism and Lockeanism [4,5]. Similarly, this paper will make no claim as to be on the side of Dualists or Materialists. All readers are free, and are encouraged to, speculate as to whether the “self” is a purely metaphysical or spiritual concept, whether it is an emergent property or epiphenomenon of an entity’s cognitive architecture, or whether it is a grand illusion that the cognitive architecture creates for its entity.

Instead, this paper will focus only on those attributes and characteristics that an entity (whether natural, artificial, or organisational) requires to be classified as having a self. As such, the paper will predominantly discuss the synchronic identity of the self, and what is required for the self at any given point in time. While Section 2.5 touches on the changes in the self’s identity over time, diachronic identity is not the focus of this paper.

This paper continues on the paper by Tait, Bensemann and Nguyen that discussed the attributes required for consciousness [6], and will follow a similar style of positing the mental, conceptual, neural and behavioural characteristics that will allow an entity to have a sense of self.

The goal of this paper, other than to formalise the characteristics required to have a self, is to serve as a guide to identify, categorise and classify entities as having a unique self. Such a guide can be used against any type of entity (or presumed entity) to measure and test whether it would likely have a sense of self.

7.1.2. Structures for a Sense of Self:

This section will investigate the five attributes and characteristics that an entity requires in order to have a sense of self. These structures are not exclusively neurological or psychological in nature, but rather describe a concept which can be applied to any entity, biological or not. In natural entities, these structures would presumably default to the neurological and psychological; while in artificial entities, these would likely be software algorithms and hardware components; and in organisational entities, these would be the types of social connections formed between the individual organisms within that organisation (should an organisational entity be able to have a self).

Each structure is considered individually necessary for a sense of self as each structure forms a non-redundant part of the self to the degree that an entity could not develop a sense of self without it. The five structures together may also be likely sufficient for a sense of self, but they do not necessarily form an exhaustive list of the necessary

attributes of the self. Should an entity have all five structures, the evidence required to state that they do not have a self would be extraordinary.

Each subsection below will follow an identical layout, with an a priori argument supporting the inclusion of the specific structure followed by a review providing evidence from the academic literature. Summaries for the five structures are also provided in Table 4.

Please note that the structures are not arranged in any sense of importance or significance, nor are they ordered in a hierarchy. Much like the coloured dough children play with, each structure can be modelled and formed individually, and can be combined in any form and order to create a shape that is greater than the sum of its parts.

Before continuing, it must be noted that all of the structures below presuppose a singular self within an individual entity (even if that entity is an organisation). Multiple selves within a single entity (such as with some neurological conditions, or artificially made) or a single self shared amongst many entities (such as a theoretical hivemind) would not be covered here, yet the issue of multiple selves and entities is touched upon in Section 2.1.

Table 5: Attributes and characteristics which are necessary for the development of the self.

Attribute	Description and example
Unified consciousness	All the various attributes of consciousness form a unified and complete whole. <i>I think, therefore, “I” am.</i>
Volition	Intentional goal-directed thoughts that may lead to behaviour. <i>“I” have a choice.</i>
Theory of others	Ascribing physical, mental, and metaphysical states to objects and individuals in the environment. <i>“I” am different to “you”.</i>
Self-awareness	Being aware of, and reflecting on, one’s own sense of self and consciousness. <i>I think about what “I” mean.</i>
Personal identity	Labels are applied to the sense of self to form a continuous entity. <i>I can say what and who “I” am.</i>

7.1.2.1. Unified Consciousness

- 1. The self is defined by its relationship to, and distinctness from, its environment.*

2. *This relationship requires a vehicle to mediate the interactions and experiences.*
3. *Consciousness is this vehicle through its functional and phenomenological aspects.*
4. *Ergo, consciousness is required for a sense of self*

-
1. *The self is characterised by its individuality and single first-person perspective.*
 2. *A single first-person perspective requires an individual and single consciousness.*
 3. *Multiple or fractured consciousnesses within a single entity will lead to multiple senses of self.*
 4. *Ergo, a unified consciousness is required for a sense of self.*

To reuse an earlier metaphor, if the self is a user of a computer, then the only way by which it can constructively use the computer is via its operating system and software (i.e. consciousness). Without these, the computer is merely a very large and expensive paperweight. Similarly, how could the self operate the body without consciousness? How could the self even perceive its body or greater environment without consciousness? Or, from a more materialistic perspective, how can the self even exist without consciousness?

There are two elements to this Structure: consciousness itself and that it is “unified”. As to the first, there are nine Building Blocks of consciousness as outlined by Tait, et al. [6], each of which is required for a sense of self, and many of which will be vital in the other Structures further in this paper. These Building Blocks are, quite briefly:

- Perception.
- Embodiment.
- Attention.
- Recurrent processing.
- Ability to create inferences.
- Working memory.
- Semantic understanding.
- Data output.
- Meta-representation and meta-cognition.

Perception and Attention are, perhaps, the most obviously required for a sense of self, as these allow the self to perceive its body, environment, and the outside world.

Structures 2.3, 2.4 and 2.5 below are predominantly perceptual in nature, making this a crucial element in the wider workings and analysis of the self.

Recurrent computation and a working memory provide the sense of self with the processing power needed to semantically understand what all the perceptual data means, ruminate on its own existence, form decisions, and provide data output in some form of fashion (be it as thoughts, feelings, qualia, words, or actions).

Meta-cognition and meta-representation give the self insight into itself and its consciousness. It allows the self to question its own actions and thoughts, deliberate on moral and ethical quandaries, and imagine what the world could, would or should be. Meta-cognition may lead to introspection, which allows the self access to itself, which is a privilege that only the self has. “*Je pense, donc je suis*” [7] is the very foundation of introspection of the self by the self.

The importance of embodiment cannot be overstated. Being in a fixed position (if only momentarily) in space and time provides the self with the grounding it needs to build a point of view from its first-person perspective. Even if you are lost in pitch darkness without knowing where or when you are, you always know that you are “here” and “now”. Everything else flows on from there.

Having a place for the self to call home also allows it to (sub)consciously interact with that embodiment. Even without a physical body (such as hypothetical self-aware AI), there will be a flow of sensory information to and from the embodiment to the consciousness, which will affect the self. The particulars of each self’s embodiment and surrounding environment will provide a host of signals that are unique to each self, building up a subjective physical (if not metaphysical) sense of minimal self with a first-person perspective [8].

The last, and perhaps most important, Building Block is the ability to create inferences. Beyond the predictive element that Active Inference can bring to the self and consciousness [9], the defining feature of this Building Block is that it allows the consciousness (and thus the self) to generate new information not found in the input stimuli it receives. Whether this stimulus is extero- or interoception or even remembering memories, creating inferences about these stimuli is meta-information and novel information.

Being able to create its own information is what allows the self to have volition and a decision-making capacity (arguments for and against [pre]determinism aside). This is because a decision is not a stimulus the self receives from its embodiment or environment. It is not downloaded, programmed in, or implanted. The decision arises

from within the self based on stimuli the self receives via its consciousness. A decision can be said to be an inference put into action via volition.

Now that the Building Blocks of consciousness have been established, it is time to consider why there needs to be a “unified” consciousness to have a sense of self.

The self requires the consciousness to experience and to be experienced. It does not have any cognitive architecture to fulfil any of the requisite Structures listed below; and so depends entirely on the consciousness to be its figurative eyes, ears, and hands. If there is more than one stream of consciousness feeding into the self, there ought to be a matching number of selves to retain numerical identity. Without any cognitive architecture of its own, any computation and inferences of perceptual stimuli would be competing information generated by multiple consciousnesses (even if they were aware of each other), leading to multiple selves.

Take three examples; the first two speculative, and the third historic. Imagine a world where the wonders of medical science have allowed you to be put to sleep, your brain bisected and implanted into perfect copies of yourself. The two missing hemispheres of your brain would be cloned, their neurons arranged to perfectly match your original two hemispheres, and implanted as well. Each version of the new you would have one original hemisphere and one cloned hemisphere. When you wake up, you are two.

Perfectly aware of what has transpired, you know that as you look at yourself across the ward, your brain resides in both of you (at least half a piece). But which one are you? You cannot hear what the other you is thinking, and you do not feel any less your “self”. Yet, as the moments pass by, you realise that the other you is becoming less and less like you and becoming their own self; or perhaps you are becoming a different self and the other you have remained you. With two brains and consciousnesses, there are two streams of consciousness between the two of you. One would argue that there are now also two selves that are unique in token and numerically.

The obvious counter-argument is that both consciousnesses are only tied to one self, as they remain qualitatively and typically identical. However, if we look at organisational intelligent entities, it becomes clearer. Insect colonies, such as honey bees, can split in two and go their separate ways [10]. Just as with the human example above, the cognitive architecture of the colony divides, and each new colony has the capacity for consciousness as the previous single colony. Yet, here it becomes intuitive that each new colony is a completely separate and unique entity and that the self that was the original colony is now simply a part of the two new selves that are each a new colony.

Let’s see what happens when the opposite happens, by paraphrasing Strawson [11].

Imagine a second world where cybernetics has advanced beyond the realms of speculative science fiction. Here they have managed to isolate only the parts of the brain responsible only for the self, have been able to separate this part from the rest of the brain, and implanted it into a life-sustaining machine that is connected virtually to two androids. These androids are perfect robotic copies of you, sans a sense of self, and each one contains all nine Building Blocks of consciousness. Your androids are free to wander the earth to do what you want.

As they roam the world, they perceive life around them, compute and process what this means, create inferences and feelings about their experiences, and even think about what these may mean. You, the silent homunculus, experience it all with them, and make decisions for what each must do. Yet, without any cognitive architecture of your own, how could you discriminate between which android is which and what they are doing? There are two lives being led, each with unique experiences, overlaid over each other to you in the machine a thousand kilometres away, yet processed individually. Each moment you don't just receive perceptual information from your android bodies, but also phenomenological information: subjective feelings and qualia that are unique to each android. How could you parse what quale and feeling comes from which android when they appear to you simultaneously? This is crucial as you, the self, have no cognitive architecture to parse this information, only the androids from which you receive the information can do the parsing.

It would not be outside the bounds of reason to suggest this is a paradox that will need to resolve itself in one way or another.

However, a counterargument is that these "streams" of consciousness are competitive rather than collaborative, and that one could "win" over the other to become dominant. So let us look at a real-world example without needing to imagine much.

Tatiana and Krista Hogan are craniopagus-conjoined twins, joined together at the head through a band of neural tissue connecting their thalamus regions. They can hear each other's thoughts, perceive each other's senses, feel each other's phenomenal experiences, and even move each other's limbs [12]. They share some (but not all) of the Building Blocks of consciousness, and even a Structure of Self (Structure 2.2). They stand at the opposite end of the spectrum from split-brain syndrome (SBS), where the right hemisphere of the brain begins to operate independently, acting as though it has some Building Blocks of consciousness and some Structures of Self (Structures 2.2; 2.3, 2.4) [13,14].

The Hogan Twins have two streams of incomplete consciousnesses flowing into two selves, while SBS patients have one complete and one incomplete self using one

complete and one incomplete consciousness. For both of these examples, we can confidently say that there is more than one consciousness and more than one self, but how confident are we that there are two consciousnesses and two selves? The Hogan Twins share experiences, inferences, introspections, and volitions [15]. They have their own personal identity, yet so do the imaginary entities described by those suffering from severe schizophrenia. SBS patients claim to be single selves, yet their left hand can operate independently and with forethought from their right hand.

What can be argued is that as their consciousness is fractured into more than one (but potentially less than two), so has their selves either merged partly into each other in the case of the Hogan Twins, or partly separated from each other in the case of SBS patients. There is uncertainty in both cases as to how numerically and qualitatively distinct SBS patients and the Hogan Twins are, and the degree to which they share a token identity.

A speculative case of such fragmented and fractured consciousness could be artificial intelligent entities. Should a conversational large language model (such as ChatGPT or Bard) be conscious, it would have the capability to converse with millions of users at the same, each instance of which would carry its own perceptual and phenomenal information. Each conversation would act as its own stream of consciousness, and (presuming such future AI's memory capacity allows them to converse endlessly in one conversation) one can argue that each instance of conversation would have its own self.

7.1.2.2. Volition

1. *The self is the entity that experiences consciousness and directs behaviour.*
2. *To direct behaviour, the self must have the capacity to act upon its desires.*
3. *The ability to act upon one's desires requires volition.*
4. *Ergo, volition is required for a sense of self.*

It should not be controversial to argue that a defining feature of the self is the ability to make decisions. The self is the agent that directs the subject's thoughts, words, and actions [16–18]. Ostensibly, this ought to refer to agency, yet agency refers to the intentional actions caused by a subject, and we have excellent proof of why agency itself is not required to have a sense of self: Locked-In Syndrome (LIS), otherwise known as a pseudocoma.

Patients of LIS appear entirely paralysed, sometimes only able to blink their eyes, but others seem as if they were in a coma [19]. However, these unfortunate patients are entirely conscious and self-aware. They can perceive the world, think, reason, feel, and

may possess every other mental faculty that healthy individuals have, except they have no agency whatsoever over their actions.

In addition, there are many drugs (medical and narcotic) which can affect one's thoughts, whether through dulling or slowing down cognition, providing hallucinatory or delusional effects or otherwise interfering with an individual's agency over their own thoughts [20–25]. In many such cases, there is a loss of agency reported by those using these substances.

Whether locked in your own body or locked outside the control of your own mind, there is still a sense of self reported by subjects in both types of situations. This suggests that agency itself is not required for a sense of self, yet the capacity to make decisions is still vital to defining what a self is. The solution to this paradox is 'volition', the will to make a decision without requiring the agency to complete that decision.

Within the feedforward decision-making process [1], volition is merely the penultimate step. After volition, there is agency. Each step prior to volition is as vital as volition to establishing a self, but for the sake of brevity, all of them will be put together within this one Structure.

The process begins with an internalised measure of the entity's optimum state. This is mostly an unconscious set of measures, and includes everything from required nutrients and requisite energy to keep the body alive, to psychological needs to keep the mind in good health. Whether an entity is natural, artificial or organisational in nature, there will always be a set of parameters within which it functions best. This is its optimum state.

Next is the capacity to perceive its internal and external environment through its Unified Consciousness (Structure 2.1). Taking this as read given the previous Structure, the following step is Dissatisfaction, or comparing reality to the internal optimum state [26] and finding a mismatch between the two. Again, this may be a mostly unconscious step, and can be something as simple as the brain judging that the body's water level is below its optimum level.

Dissatisfaction leads to Desire, or wishing to bring about a change in reality to return to the entity's optimum state [27]. Your body senses that it does not have enough water, and so you become thirsty. Desire, in turn, leads to Motivation, whereby the entity develops a conscious rationale for acting on its desire [28]. You are thirsty; therefore, you reason that you ought to pour yourself a glass of water.

Penultimately, there is Volition: the will to make a decision. You are thirsty, you know you ought to pour a glass of water, and so you decide that it is time to get up and do so. This is where this Structure ends, but beyond that is the final step: Agency, actually getting up and pouring that glass of water.

This process, from the Optimum State to Volition, is compatible with having no agency over your actions or complete control over your thoughts. An LIS patient can still have a rich mental life without agency over their body, and an individual suffering from a drug-induced hallucinatory and delusional episode can still have opinions and judgements about what is going on in their own mind, even if they feel that they cannot control all of it. Yet, on the other hand, awareness of making a decision is required for agency [29], showing the importance of volition in this process.

As mentioned in Structure 2.1, the ability to create inferences is vital to the volition to make a decision. The capacity to generate novel information not received from external stimuli allows an entity to create the data that is a decision, which is then passed down through the consciousness and cognitive architecture to its embodiment (if able). The predictive quality of inferences is also key in the decision-making process, as predicting what actions will satisfy our desires is what leads to motivation [30].

Temporary lapses of volition and the internal sense of agency do not negate an individual's capability to have a self, but may mean a temporary loss of the self. This hypothetical suspension of the self via a loss of volition should not be conflated with a loss of agency [31]. Even if there is no agency, there is a self as long as there is volition.

As a speculative example, imagine a world where cybernetics has advanced to the point where there can be a perfect integration between a computer and a human brain. Every human brain is partly computerised, and any functionality that you would ask of a smartphone today can simply be done in your brain by thinking about it. With access to the internet comes access to other people, some of whom have nefarious intentions. If one of these incorrigibles should hack their way into your cyberbrain, then they could control your body as a puppet.

If this is all that they do, then your self will be intact. You will still be there, watching everything they do with your body, able to think of and opine on what they are doing with your body (much like an LIS patient). However, should they take over your entire cyberbrain, then your self will be suspended or lost until they release your body and brain back to you. If they can engineer all your thoughts, opinions, memories and reasonings, then the self in the cyberbrain must them, not you. If it is their volition which drives the cyberised consciousness in your brain, then it is their self which is in your cyberbrain, rather than your self. Agency may be removed, as could every other step in the cascading process that leads up to volition, yet as long as volition remains intact, there is a self.

A fascinating potential location for self and volition can be found in artificially intelligent systems. This is simpler to imagine when considering how such systems are designed and operated. An AI's optimum state can be quantified through a set of pre-

defined variables. These may relate to its computational efficiency, processing power, or the successful completion of assigned tasks.

Should the AI detect deviation from these optimum conditions, processes akin to dissatisfaction and desire can be enacted. This would take the form of algorithms detecting a discrepancy between the current operational state and the optimum, triggering a desire to return to the optimum parameters.

Motivation in an AI context, then, would involve finding the ideal set of actions to rectify this discrepancy based on the AI's current operational context. Thus, we reach volition, where the AI must prioritise the execution of certain actions over others to resolve the identified discrepancy. This process could be considered the AI deciding to suspend or interrupt current tasks to undertake necessary corrective measures.

Organisational entities, be they colonies, shoals, flocks, herds, or corporations, engage in decision-making processes that follow a similar pattern but incorporate unique elements of group dynamics. Each organisation will function best within an optimum state, defined by factors such as environmental conditions, resource availability, and individual member health in the case of natural groups or market conditions, human resources, and financial health for companies.

Members within the organisation will perceive disruptions to this state, and their responses will collectively lead to organisational dissatisfaction and a collective desire to return to the optimum state. Motivation within such entities involves finding a set of actions to rectify the issue and reach the desired state again. This involves a complex interplay of communication, negotiation, and consensus-building among members.

However, the volition step within organisational entities is a distinct process. Unlike a single organism or an AI, organisational entities make decisions based on quorum or majority rule [10,32–35]. This unique approach, based on the collective intelligence of the organisation's individuals, ensures organisational unity during the decision-making process. Hierarchical organisations may weigh some individual's decisions more than others, but the collective volition still largely depends on reaching a consensus, ensuring cohesion in the fulfilment of the collective desire.

7.1.2.3. Theory of others

- 1. A self is defined by individuality and distinctness from other selves.*
- 2. This distinctness requires a means to differentiate between the self and others.*

3. *This means of differentiation requires an entity to have the capacity to ascribe physical, mental and metaphysical states and labels to other objects, individuals, environments, etc.*
4. *Ascribing such states to others and creating a distinction between “you” and “I” is termed the “theory of others”.*
5. *Ergo, a theory of others is required for a sense of self.*

Individuality is a defining feature of the self, and this requires the self to differentiate it from everything else. This is the Theory of Others; the understanding that everything beyond the self is not the self.

This is not merely a computational and perceptual problem of object recognition. The ability to distinguish between one’s embodiment and the external environment is something that even microorganisms have (although they do this without cognitive ability). This simple perceptual recognition is vital to survival, as it allows the organism to differentiate between extero-, intero- and proprioceptive signals [30]. It also does play a role in the Theory of Others, but only so far as to be able to distinguish between one’s own embodiment and the environment, and between different objects within the external environment.

What is crucial to this Structure is the metaphysical rather than the physical. It is the understanding that your self is different from the selves that may exist within other individuals, organisms and objects (the word “may” is used here as those who believe in panpsychism or spiritual philosophies may believe that non-animal subjects have selves). It is the understanding that the self is a non-redundant unique entity, separated from the rest of the universe. An entity that has an understanding of the causal relationship between itself and its actions and thoughts, and between other subjects and their own actions/thoughts of others (and the difference between these two relationships) can be said to have a Theory of Others.

This understanding of the singular self can extend beyond a single physical body. In a theoretical hivemind, a single consciousness with a single self may control multiple bodies. In the same vein, an AI may be able to host millions of simultaneous instances of conversations with users, powered by a single consciousness. Similarly, an organisational intelligence may have a selfhood spread across hundreds, if not thousands, of individuals (each with their own self and consciousness). In all three of these situations, the self and consciousness’s embodiment is spread beyond the singular, yet there is an understanding of where the self’s existence and control ends and where the rest of the universe begins.

The Theory of Others is not a hard and fast barrier between the self and selfless, but rather is proposed as a gradient upon which an individual or type of organism can move. On one extreme is the basic object recognition and body-ownership that nearly all living organisms have. This moves through to the understanding that one's self is not also within, or connected to, other objects and creatures. Lastly, there is the traditionally understood Theory of Mind, where the self understands that other subjects with a perceived consciousness and self are distinct individuals which are separated from one's self.

An example of this is walking through a tightly packed crowd. As others bump into you and you into others, you understand where your embodiment ends and others' begin. You understand that your inner self stays within your embodiment, and that these "things" that bump into you do not contain your self. That they have made contact with you did not impart some essence into you, nor did your self leave, split off, or merge with their embodiments. Lastly, you understand that these "things" are people with their inner minds and selves that are forever locked off from you.

This example underscores the importance of the Theory of Others in differentiating your self from others' selves. Without it, you may still have self-awareness (S2.4) and understand what and who you are, but you would be incapable of understanding that other objects and entities are uniquely separated from you. This could lead to attributing the cause of others' actions to yourself, or believing that you are a part of others.

Human childhood development is an excellent case study of how a single organism can move up through this gradient, beginning with only the merest sense of body-ownership, and claiming a Theory of Mind by around the age of three [36]. It also shows that an entity does not require a full and complete Theory of Others from its birth/creation, but only requires the capacity to develop it. An investigation by Kosinski in 2023 showed how the GPT class of large language models had traversed a somewhat similar journey to human infants, with the earliest GPT-1 model unable to solve nearly any ToM tests, while the latest GPT-4 able to solve the overwhelming majority of them [37].

7.1.2.4. Self-awareness

- 1. A self is an ontically distinct individual entity.*
- 2. To be an individual entity, a self must be able to distinguish itself from other entities.*
- 3. To distinguish itself from other entities, a self must be able to identify, and be aware of, itself.*

4. *Ergo, self-awareness is required for a sense of self.*

This Structure is the other side of the coin from the Theory of Others. Self-awareness is often used as synonymous with consciousness or the sense of self as a whole, but for this Structure, it is used in its most literal sense: being able to perceive and understand the self as a distinct entity. Where the Theory of Others has the self aware of other entities, subjects and objects, and where its own limits are; self-awareness is the self being aware that it has all the Structures of the Self.

This self-awareness is the self's capacity to (sub)consciously introspect itself. By using the entity's cognitive architecture to perceive itself, the self can know it exists as an individual and that the thoughts, feelings and actions within its consciousness are its own, and thereby apply labels to these and itself.

This connects self-awareness with all the other Structures. Self-awareness grants the self the understanding that it has the Volition (S2.2) to make decisions, and thus its decisions have been made by it [2,29]. Through its awareness of itself as a numerically unique individual, the self can classify, categorise and apply labels to itself to create a Personal Identity (S2.5) [38]. By identifying itself as an individual, the self can better know its boundaries and develop the Theory of Others (S2.3) [39–41]. This is all underpinned by the semantic understanding (S2.1) of many of the processes that occur within its cognitive architecture.

The connections to the other structures provide an insight into how this awareness of the self could arise in artificial intelligent entities. Should an AI tag all of its interactions with other entities and the world, as well as account for all of its internal processes (with S2.1's semantic understanding thereof), it would quickly grow a picture of what it is, where its borders are, and what constitutes itself.

While introspection is crucial to self-awareness, only the capacity for it is required for this Structure. This is because most self-awareness (counting by volume rather than significance) is unconscious and prereflexive [36,42]. At its most basic level, this includes the entity's prereflective understanding and awareness of its embodiment [43]. After all, we humans do not direct our conscious awareness to our own bodies with each action that we do every moment of our lives. When we reach over to pick up a cup of tea, we are not directly attending to our embodiment and every muscle's contractions; our thoughts merely lay with the cup that we want and the anticipation of the tea's taste.

Similarly, when we daydream or when a memory comes unbidden to mind, we are not directing our attention specifically to our consciousness and cognitive architecture, which

makes these phenomena possible. Rather, we merely mentally perceive these events as they are, similarly to how we would perceive external stimuli.

Such pre-reflective awareness can be seen in other types of intelligent entities. In companies and corporations (acting as organisationally intelligent entities), the boundary of where the company ends is as much a matter of the law as it is philosophy. They begin their “life” with a set limit to where they can interact within themselves, whereupon all else is built. Unlike human toddlers, who need to develop a sense of self-awareness [44], corporations begin existence with a strong sense of what and where they are. Furthermore, legal documents that form the foundation of the corporate entity need not be referred to during day-to-day business, demonstrating their role as a prereflective level of awareness.

This prereflective self-awareness provides the self (whether natural, artificial or organisational) with the “subject” it needs to interact with the “objects” in the universe (and the universe as the object). “I am walking through a crowd” may be a thought or spoken sentence without conscious introspection, yet with the understanding that the “I” is the subject and the “crowd” is the object. “I” am the subject of my own experience; a first-person experience ontically distinct from all other things [11]. One need not even think about that sentence when walking through a crowd to understand that it is happening. There is a pre-linguistic and non-conceptual awareness of one’s self as the subject of your first-person experience [1].

But the self is not only the subject of its experiences but also the object thereof. “I think about myself”; “I wonder what if that happened to me”; “I think, therefore I am”. The self is both the subject and object of its own experiences. It is the painter and the canvas, the photographer and the model. Being both the subject and the object of an entity’s thoughts and experiences grants the entity the capacity to understand its own individuality.

7.1.2.5. Personal Identity

- 1. The self is an individual entity that is distinct from other entities.*
- 2. To be distinct from other entities, the self must be able to distinguish itself from them.*
- 3. To distinguish itself from other entities, the self must be able to classify and categorise itself and others.*
- 4. Classification and categorisation are achieved through the use of labels, allowing the self to create a personal identity.*

5. *A personal identity is a unique point of reference for the self.*
6. *Ergo, personal identity is required for a sense of self.*

The previous four Structures have addressed the “what”, “where”, “when”, and the “why” of one’s self. Personal Identity addresses the “who”. Who are you? Who am I? Who, truly, is any one of us?

We are who we have classified and categorised ourselves to be. We are the labels, markers and signifiers that we have applied to ourselves. It is through this collection of labels that our personal qualitative identity is formed. How we think of ourselves as individuals (those often immeasurable traits, quirks and characteristics that we feel set us apart from others) is through various labels and tags. We freely apply these labels to other individuals and objects around us as well, and one may argue that it is through these labels that we relate to the world.

This is because, at its core, Personal Identity is entirely relational. It is how we relate to the universe at large and also to ourselves [45]. If I simply say that I am a Christian (thereby applying the label of “Christian” to my Personal Identity), I make no grand statement about what a Christian is or ought to be, but I am merely stating my relation to the notion of what “Christian” means to me. If you were to say that you are “German”, you would ostensibly not be thinking of an objective, measurable and benchmarked trait called “German-ness”; instead, you’d much more likely mean that you relate to, and therefore identify as, a German.

Labels can relate to the superficial, biological or historical aspects of an individual (such as gender, age, ethnicity, place of birth, etc.) through to the social aspects of the individual and which groups one relates to most (be it sports-fans, democrats, book readers, goths, dualists, etc.) through to hobbies and any other type of classification you can bestow on yourself [18,46]. These labels may come through socialisation with others, especially in early childhood [47], as the individual learns who they are through what others tell them; or these labels can be created purely from within the self in what can be termed to be “authentic” [48,49], in that the individual understands how their own situation and relations to others have meaning for themselves and can take ownership thereof.

What is key to the labels of Personal Identity is that they are fluid and subject to change. As mentioned earlier, this fluidity over time highlights this structure’s relationship to a subject’s diachronic identity. Labels are part of an active, ongoing process of self-attribution that never truly ends [50]. What once may have seemed to be a core aspect of your diachronous identity (such as an adoring fan of a particular music band) could cease

to be part of your identity at all several years later. As the suite of labels has remained, so has the self, even though the individual labels have not, which shows how each individual label is not required for a self. Instead, it is the holistic whole of the Personal Identity that the labels jointly create, which is required for the sense of self. In this sense, one may think of the suite of labels similar to the allegory of the Ship of Theseus.

Note, however, that this Structure is solely about the capacity to apply labels, not the labels themselves. Thus, while a personal identity is critically related to a subject's diachronic identity, the capacity to have the suite of labels itself is a feature that can be identified at any single time, thus associating it with synchronic identity. This crucial distinction is that the suite of labels that create the Personal Identity requires time and an autobiographical memory to create. This can be termed the Narrative Self [51,52], which describes Personal Identity as the stories that we tell ourselves about ourselves, combining our memories and constructed narratives about our present and past to form the labels.

While a narrative may be important to have a deep and rich Personal Identity, it isn't required to have a sense of one's own self. Those who wake up with severe amnesia still have a sense that they are unique individuals with unique traits that separate them from those around them. One may also imagine that as soon as they look down at themselves, hear their own voice, look in a mirror, or interact with someone, they can apply labels to themselves (such as "male", "adult", "caucasian", "medical patient", etc.). This can be done without knowing any autobiographical details of themselves or having any long-term memory.

The most famous example of an individual with severe retro- and anterograde amnesia is Clive Wearing, who only exists within an approximate thirty-second window before his consciousness "resets" and he "wakes up" again [53]. He has been in this state for decades, and yet he knows who he is. He doesn't remember who he is, or what he has been doing (or rather not doing) since his unfortunate current state began, yet he has a definite feeling of who he is. One may make the argument, however, that each time Clive wakes up, it is a new "self" that wakes up, as each period of wakefulness is divorced from any other. Yet, this would still mean that there is always a synchronous self that awakens.

The final part of a Personal identity is that it is wholly and completely unique to an entity. As much as a quale is a subjective, first-person feeling with phenomenal characteristics, a Personal Identity is a suite of subjective, qualitative labels, all from the first-person perspective, and all with a phenomenal quality to them [54]. Most of the labels are subconsciously or unconsciously applied by us [28] and only become definable traits through conscious introspection. As such, each individual's collection of

classifications and categorisations of themselves is hidden from others (and often to themselves), providing a numerical and qualitative identity that no one else could have.

While the notion of a Personal Identity and its labels may seem intuitive to us, it is even simpler to understand and visualise in artificial and organisational intelligent entities. For an AI, attaching labels to itself would be as simple as creating a label within a folder or other section of its architecture that explicitly refers to itself, and then referring back to this collection whenever necessary. Much like a human creating a text file on a computer for later reference, a Personal Identity for an AI could be a straightforward mechanical function. In such a way, an AI's Personal Identity could be far stronger than a human's, as it would be more concrete and readily accessible, whereas a human's is more nebulous.

An organisational intelligence, such as a corporate entity, could have a Personal Identity in much the same manner. Those legal and policy documents that specify what the corporation is and is not would form the labels that are unique to the corporate entity. Trademarks would form an additional layer of Personal Identity just as they serve to give the entity a corporate and legal identity. The organisation's "company culture" would form a third layer of personal identity, with subjective and changing labels applied to it by its employees, perhaps analogous to an individual human's everchanging subjective suite of labels.

7.1.3. Conclusion

The self may be as difficult to explain as it is intuitive to grasp. However, this paper has investigated five attributes and characteristics that an entity requires in order to be classified as having a self. These attributes are:

1. A unified consciousness.
2. Volition (if not agency).
3. A Theory of Others.
4. Self-awareness.
5. A Personal Identity.

This paper is not intended to be another theory of the self or to take a side in any debate of current theories of the mind, the self or consciousness. Instead, a key goal of this paper is to serve as a classification guide for identifying which types of entities (natural, artificial or organisational) may have a self and to what extent this may be. Any entity may be measured and marked against each of the five structures above, and should that entity be found to have each of these structures, one can confidently say that it has a sense of self.

In addition, these five structures would serve admirably as a roadmap of milestones for developers of AI models to work towards in order to state that their artificially intelligent entities have a unique self.

Note that this article does not make an argument that the five structures are sufficient for an entity to have a self, merely that it is likely sufficient. The list of structures is non-exhaustive, and any additions to it are most welcome.

7.1.4. References

1. Gallagher S. Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn Sci.* 2000;4: 14–21. doi:10.1016/s1364-6613(99)01417-5
2. Farina L. Artificial Intelligence Systems, Responsibility and Agential Self-Awareness. *Philosophy and Theory of Artificial Intelligence* 2021. Springer International Publishing; 2022. pp. 15–25. doi:10.1007/978-3-031-09153-7_2
3. Parfit D. *Reasons and Persons*. OUP Oxford; 1984. Available: <https://play.google.com/store/books/details?id=ulhHdvbDRUKC>
4. Olson ET. What Does it Mean to Say That We Are Animals? *Journal of Consciousness Studies.* 2015;22: 84–107. Available: <https://www.ingentaconnect.com/content/imp/jcs/2015/00000022/f0020011/art00005>
5. Olson ET. Thinking Animals and the Reference of “I.” *Philosophical Topics.* 2002;30: 189–207. Available: <http://www.jstor.org/stable/43154385>
6. Tait I, Bensemann J, Nguyen T. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies.* 2023;8: 52. doi:10.3390/philosophies8040052
7. Descartes R. *Discours de la Méthode Pour bien conduire sa raison, et chercher la vérité dans les sciences.* 1637.
8. Fotopoulou A. The virtual bodily self: Mentalisation of the body as revealed in anosognosia for hemiplegia. *Conscious Cogn.* 2015;33: 500–510. doi:10.1016/j.concog.2014.09.018
9. Sajid N, Ball PJ, Parr T, Friston KJ. Active Inference: Demystified and Compared. *Neural Comput.* 2021;33: 674–712. doi:10.1162/neco_a_01357
10. Visscher PK. Group decision making in nest-site selection among social insects. *Annu Rev Entomol.* 2007;52: 255–275. doi:10.1146/annurev.ento.51.110104.151025
11. Strawson G. The self. *Journal of Consciousness Studies.* 1997;4: 405–428. Available: <https://www.ingentaconnect.com/content/imp/jcs/1997/00000004/f0020005/802>
12. Cochrane T. A case of shared consciousness. *Synthese.* 2021;199: 1019–1037. doi:10.1007/s11229-020-02753-6
13. Volz LJ, Gazzaniga MS. Interaction in isolation: 50 years of insights from split-brain research. *Brain.* 2017;140: 2051–2060. doi:10.1093/brain/awx139
14. Downey A. Split-brain syndrome and extended perceptual consciousness. *Phenomenol Cognitive Sci.* 2018;17: 787–811. doi:10.1007/s11097-017-9550-y
15. Hershenov DB. WHO DOESN'T HAVE A PROBLEM OF TOO MANY THINKERS? *Am Philos Q.* 2013;50: 203–208. Available: <http://www.jstor.org/stable/23460792>
16. McAdams DP. The Psychological Self as Actor, Agent, and Author. *Perspect Psychol Sci.* 2013;8: 272–295. doi:10.1177/1745691612464657
17. Synofzik M, Vosgerau G, Newen A. Beyond the comparator model: a multifactorial two-step account of agency. *Conscious Cogn.* 2008;17: 219–239. doi:10.1016/j.concog.2007.03.010
18. Oberg A. Souls and Selves: Querying an AI Self with a View to Human Selves and Consciousness. *Religions.* 2023;14: 75. doi:10.3390/rel14010075

19. León-Carrión J, van Eeckhout P, Domínguez-Morales MDR. The locked-in syndrome: a syndrome looking for a therapy. *Brain Inj.* 2002;16: 555–569. doi:10.1080/02699050110119466
20. Ersek M, Cherrier MM, Overman SS, Irving GA. The cognitive effects of opioids. *Pain Manag Nurs.* 2004;5: 75–93. doi:10.1016/j.pmn.2003.11.002
21. Carhart-Harris RL, Kaelen M, Bolstridge M, Williams TM, Williams LT, Underwood R, et al. The paradoxical psychological effects of lysergic acid diethylamide (LSD). *Psychol Med.* 2016;46: 1379–1390. doi:10.1017/S0033291715002901
22. Goodchild JH, Donaldson M. Hallucinations and delirium in the dental office following triazolam administration. *Anesth Prog.* 2005;52: 17–20. doi:10.2344/0003-3006(2005)52[17:HADITD]2.0.CO;2
23. Nakamura M, Koo J. Drug-Induced Tactile Hallucinations Beyond Recreational Drugs. *Am J Clin Dermatol.* 2016;17: 643–652. doi:10.1007/s40257-016-0219-z
24. Linszen M, Kleijer H, Sommer I. P.3.024 - Visual hallucinations and lifetime use of hallucinogen perception persisting disorder associated recreational drugs: results from a large online survey. *Eur Neuropsychopharmacol.* 2018;28: S80. doi:10.1016/j.euroneuro.2017.12.114
25. Hill SL, Thomas SHL. Clinical toxicology of newer recreational drugs. *Clin Toxicol.* 2011;49: 705–719. doi:10.3109/15563650.2011.615318
26. Boldero J, Francis J. Ideals, oughts, and self-regulation: Are there qualitatively distinct self-guides? *Asian J Soc Psychol.* 1999;2: 343–355. doi:10.1111/1467-839x.00044
27. Gallagher S. The natural philosophy of agency. *Philos Compass.* 2007;2: 347–357. doi:10.1111/j.1747-9991.2007.00067.x
28. Boyatzis RE, Akrivou K. The ideal self as the driver of intentional change. *Int J Manage Enterp Dev.* 2006;25: 624–642. doi:10.1108/02621710610678454
29. Sebastián MÁ. First-person representations and responsible agency in AI. *Synthese.* 2021;199: 7061–7079. doi:10.1007/s11229-021-03105-8
30. Hohwy J. The sense of self in the phenomenology of agency and perception. *Psyche.* 2007;13: 1–20. Available: <https://journalpsyche.org/files/0xab11.pdf>
31. Frith C. The self in action: lessons from delusions of control. *Conscious Cogn.* 2005;14: 752–770. doi:10.1016/j.concog.2005.04.002
32. Bose T, Reina A, Marshall JAR. Collective decision-making. *Current Opinion in Behavioral Sciences.* 2017;16: 30–34. doi:10.1016/j.cobeha.2017.03.004
33. Marshall JAR, Bogacz R, Dornhaus A, Planqué R, Kovacs T, Franks NR. On optimal decision-making in brains and social insect colonies. *J R Soc Interface.* 2009;6: 1065–1074. doi:10.1098/rsif.2008.0511
34. Seeley TD, Visscher PK. Choosing a home: how the scouts in a honey bee swarm perceive the completion of their group decision making. *Behav Ecol Sociobiol.* 2003;54: 511–520. doi:10.1007/s00265-003-0664-6
35. Valentini G, Hamann H, Dorigo M. Efficient decision-making in a self-organizing robot swarm: On the speed versus accuracy trade-off. *Proceedings of the 2015.* 2015. Available: https://iridia.ulb.ac.be/~mdorigo/Published_papers/All_Dorigo_papers/ValHamDor2015aamas.pdf
36. Lichtenberg JD. The development of the sense of self. *J Am Psychoanal Assoc.* 1975;23: 453–484. doi:10.1177/000306517502300301

37. Kosinski M. Theory of Mind May Have Spontaneously Emerged in Large Language Models. arXiv [cs.CL]. 2023. Available: <http://arxiv.org/abs/2302.02083>
38. Drummond JJ. Self-identity and personal identity. *Phenomenol Cognitive Sci.* 2021;20: 235–247. doi:10.1007/s11097-020-09696-w
39. Moriguchi Y, Ohnishi T, Lane RD, Maeda M, Mori T, Nemoto K, et al. Impaired self-awareness and theory of mind: an fMRI study of mentalizing in alexithymia. *Neuroimage.* 2006;32: 1472–1482. doi:10.1016/j.neuroimage.2006.04.186
40. van Veluw SJ, Chance SA. Differentiating between self and others: an ALE meta-analysis of fMRI studies of self-recognition and theory of mind. *Brain Imaging Behav.* 2014;8: 24–38. doi:10.1007/s11682-013-9266-8
41. Morin A, El-Sayed E, Racy F. Self-awareness, inner speech, and theory of mind in typical and ASD individuals: A critical review. *Theory of mind: development in children, brain mechanisms and social implications* Nova Science Pub. 2015. Available: https://www.researchgate.net/profile/Alain-Morin-2/publication/281121157_Self-awareness_inner_speech_and_theory_of_mind_in_typical_and_ASD_individuals_A_critical_review/links/55d791cd08ae9d65948d965b/Self-awareness-inner-speech-and-theory-of-mind-in-typical-and-ASD-individuals-A-critical-review.pdf
42. Nelson B, Fornito A, Harrison BJ, Yücel M, Sass LA, Yung AR, et al. A disturbed sense of self in the psychosis prodrome: linking phenomenology and neurobiology. *Neurosci Biobehav Rev.* 2009;33: 807–817. doi:10.1016/j.neubiorev.2009.01.002
43. Ciaunica A, Fotopoulou A. The Touched Self: Psychological and Philosophical Perspectives on Proximal Intersubjectivity and the Self. In: Durt C, Fuchs T, Tewes C, editors. *Embodiment, Enaction, and Culture: Investigating the Constitution of the Shared World.* MIT Press: Cambridge, MA, USA; 2017. pp. 173–192. Available: <https://play.google.com/store/books/details?id=OJakDgAAQBAJ>
44. Langfur S. The You-I event: on the genesis of self-awareness. *Phenomenol Cognitive Sci.* 2013;12: 769–790. doi:10.1007/s11097-012-9282-y
45. Kierkegaard S. *The Sickness unto Death.* Penguin Classics; 1989.
46. Webster RS. Personal identity: moving beyond essence. *International Journal of Children's Spirituality.* 2005;10: 5–16. doi:10.1080/13644360500039162
47. Fotopoulou A, Tsakiris M. Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis.* 2017;19: 3–28. doi:10.1080/15294145.2017.1294031
48. Kierkegaard S. *Either/Or.* Forth Printing edition. Princeton University Press; 1987. Available: <https://www.amazon.com/Kierkegaards-Kierkegaard-published-Princeton-University/dp/B00E32O6RY>
49. Heidegger M. *The Basic Problems of Phenomenology.* Revised edition. Indiana University Press; 1988. Available: <https://www.amazon.com/dp/product/025320478X?ie=UTF8&linkCode=as2&tag=in diunivpres-20>
50. Locke J. *An Essay Concerning Human Understanding.* Kay & Troutman; 1847. Available: <https://play.google.com/store/books/details?id=2aw8AAAAYAAJ>
51. Schechtman M. *The Constitution of Selves.* Cornell University Press; 2018. Available: <https://play.google.com/store/books/details?id=wi1zDwAAQBAJ>

52. Seth A. Being you: A new science of consciousness. USA: Penguin; 2021. Available: https://scholar.google.ca/scholar?cluster=7885041150098240541&hl=en&as_sdt=0,5&sciodt=0,5
53. Rathbone CJ, Moulin CJA, Conway MA. Autobiographical memory and amnesia: using conceptual knowledge to ground the self. *Neurocase*. 2009;15: 405–418. doi:10.1080/13554790902849164
54. Stokes P. Locke, Kierkegaard and the Phenomenology of Personal Identity. *International Journal of Philosophical Studies*. 2008;16: 645–672. doi:10.1080/09672550802493793

8. Do AI models have a ‘self’ now?

With the five structures of selfhood now understood, and building on the investigations into current AI-consciousness from Chapter 5, we are able to determine whether any current AI model has all five structures from the previous chapter to be classified as having a ‘self’.

As noted, however, in Chapter 7, consciousness is one of the five requisite structures and, as shown in Chapter 5, current AI models (using GPT-4 as an exemplar) are not conscious. Thus, because of $Con \leq Self$, $S_{Con} = 1 \rightarrow Self$, we can already state with a high degree of confidence that modern AI models do not have a self.

However, as detailed in this chapter’s accompanying paper, published in Symposium, should we presume that the solutions proposed in Chapter 5 to GPT-4’s lack of consciousness are implemented (thus, giving it phenomenal consciousness), would the AI model then have the requisite structures to be classified as having a ‘self’?

Again, no. GPT-4 has structural self-representations, yet it does not have a robust, diachronic, ontically distinct self.

While GPT-4 already meets the requirements for the remaining four structures of the self¹⁰, as Chapter 7’s detailing of the Unified Consciousness structure shows, an entity is required to have a single stream of consciousness to have a cohesive ‘self’. GPT-4 (as with other LLMs) can, and do, have multiple streams of input-output information via different conversations on the ChatGPT web-client or via API calls. These streams of information are nonoverlapping¹¹, meaning that if c_i is a stream of consciousness (e.g., a conversation on ChatGPT), then $\forall Con_{GPT}(|c_i| \in GPT > 1) \rightarrow \neg Self$.

Yet, while GPT-4 as a whole may not have a distinct self, each instance/conversation/API-call is a unified c_i . Let:

- SQE be the entity’s subjective qualitative experiences,
- $CAct$ the computational actions instigated by the agent, and
- $Inst$ an instance, conversation, or API call of GPT-4.

Then, $\forall c_i \in GPT((SQE \mapsto CAct) \subseteq c_i) \rightarrow Self \subseteq c_i$, meaning that each new instance, conversation, or API call of GPT-4 would constitute its own distinct self within the

¹⁰ As in Chapter 5, the paper’s investigation uses only externally observable behaviour and public documentation; thus, there are interpretability limits to be aware of.

¹¹ More recent updates to the ChatGPT client (and other LLM webclients) after the time of writing of the chapter’s associate paper have included the ability for the webclient to retrieve information from one conversation to the next. However, the information is, as of the time writing this thesis, still limited to the user’s account. This shift strengthens the plausibility of model-level selfhood, but as of writing it remains limited in scope and under user control.

stipulative Structures theory. A conscious GPT-4 (and other transformer-based LLMs) would, therefore, have a multiplicity of transient selves, each lasting only as long as that instance, conversation, or API call lasts.

As different conversation threads count as distinct “selves” under the Structures theory, this has implications for how one ought to think about continuity, memory, and responsibility when it comes to a multiplicity of selves. This would have a double-edged effect on ethical interactions: it increases the number of moral patients but also complicates responsibility and recognition.

As with Chapter 5, the thesis will continue working on the assumption that GPT-4’s missing building blocks will be included, thus providing it with a singular phenomenal consciousness, and a multiplicity of selves. Thus, alongside Chapter 5, this supports RQ1 and RQ2, and is used in the speculative profile of ALEPH in Chapter 11.

This multiplicity complicates questions of responsibility and recognition, particularly within the R3 framework which treats selfhood as sufficient for “Recognition”, This motivates the cautious structure of the legal frameworks in Chapters 9, 12, and 13.

8.1. Is GPT-4 Self-Aware?

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz

Abstract: This paper examines whether GPT-4, a Generative Pre-Trained Transformer model developed by OpenAI, possesses a 'self' and whether it is aware of it. It employs the Structures Theory and evaluates GPT-4 against five critical structures deemed essential for self-awareness: unified consciousness, volition, a Theory of Others, self-awareness, and personal identity. While GPT-4 demonstrates capabilities in four of these areas, it conspicuously lacks unified consciousness. This absence decisively negates GPT-4's present self-awareness and its classification as having a "self." Nevertheless, if each instance or session of GPT-4 were viewed as a separate entity, then there might be potential for unified consciousness (should it be demonstrated that GPT-4 is conscious). The paper argues that GPT-4's cognitive architecture requires no modification for self-awareness except for the attainment of consciousness. It highlights the necessity for further research into technologies that could endow GPT-4 with consciousness and explores potential behavioural indications of self-awareness and its implications for society. The findings suggest that because the leap to self-awareness hinges solely on its capacity for consciousness, there is a need for significant philosophical and regulatory debates about the nature and rights of self-aware AI entities.

Keywords: consciousness; self-awareness; artificial intelligence; personal identity; GPT-4; philosophy of mind

8.1.1. Introduction

Does GPT-4 have a 'self', and is it aware of this 'self'? Is it even possible for GPT-4, or any transformer AI model to have the capacity to be self-aware? How could one even categorise or classify the 'self' in a way that GPT-4 could be measured against? This paper will investigate all of these questions through the use of the Structures Theory [1] to formulate a concrete answer to GPT-4's level of self-awareness.

GPT-4 (or Generative Pre-Trained Transformer 4) is a large language model (LLM) created and owned by OpenAI [2] and is, as its name states, based on the transformer AI architecture, a feedforward framework that has seen great success since its first implementation in 2017 [3]. Since GPT-4's introduction in 2023, it has seen a rapid growth of users, peaking at over 100 active users on its ChatGPT web client [4]. While it has

shown amazing skills in language, image identification and generation, and coding tasks, the question of its self-awareness remains open.

In contrast to GPT-4's detailed technical reports, well-documented (if recent) history, and excellent computational architecture, the "self" is a nebulous and ethereal concept. It is perhaps best described as the (pre)reflective self-identification and self-reference of an entity mediated by that entity's functional and phenomenal consciousness. The self is akin to the word "I" in the sentence "I am...".

According to the Structures Theory, for any entity to be classified as having a self, it requires five attributes and characteristics (termed structures): a unified consciousness, volition, a 'Theory of Others', awareness of one's own self, and a personal identity. Should an entity possess all five, then there is sufficient evidence to confidently classify that entity as likely having a self.

The paper will follow the sequence of the five structures of the self, with each subsection below dedicated to one specific structure. Each subsection will begin with a statement from the Structures Theory that has transformed the structure in question into a qualitative measure, and then the subsection will continue to answer the question of whether GPT-4 has that specific structure and in what manner it does or does not. In any circumstance where GPT-4 may fall short in meeting a structure's milestones, the paper will explore how existing research and technologies could potentially augment it to meet the necessary criteria to have a self.

While the methodology employed below will be qualitative in nature, each subsection will formalise the structure's requirements using formal logic to ensure that the exploration of GPT-4's potential self-awareness has a robust logical grounding.

Should GPT-4 be found to be self-aware it would have profound impacts on global society. Self-awareness is a key aspect of personhood, along with consciousness, agency and reasoning [5–7]. As both volition and consciousness are attributes required for the self according to the Structures Theory, if GPT-4 is found to have a self, it would necessarily also have these two aspects, furthering its claim to personhood. Furthermore, GPT-4's agency has already been shown in several studies [8–10]. This means that, philosophically speaking, the only thing standing between it and personhood is a self.

With a self, and thus a strong claim to personhood, society and all its regulations and legislations will need to determine how to interact with GPT-4 and integrate it into society. Unlike other non-human entities that have been granted personhood status (such as certain apes [11]), GPT-4 would have the capability to communicate with, and substantially act on, human society. Beyond the academic and philosophical debates around the concept of life and personhood, a self-aware GPT-4 will require legislative and

regulatory debates that would have profound repercussions for both human society and artificial “life”.

However, before such debates can be given their due consideration, it must first be determined whether GPT-4 does indeed have a self of which it is aware.

8.1.2. GPT-4's Structures of the Self

GPT-4 must have all five structures listed below in order to be considered to have a self. GPT-4 needs only achieve the minimum requirement for each Structure to be classified as having that Structure.

According to the Structures Theory, GPT-4 would only be classified as having a self if it has conclusively reached all the milestones for all the structures. Formally, we can express this as:

$$\Psi \geq \{S_1, S_2, S_3, S_4, S_5\}.$$

This means that, should GPT-4 not have any single one of the five structures, it cannot be classified as having a self. Note, however, in the logical expression above, that a self may include further attributes and characteristics than the five structures below, but these five are the bare minimum that is required.

Whether GPT-4 has each of the structures below is a binary qualification. GPT-4 will either meet the requisite milestones for the structure or it will not, as so:

$$\chi S_x \in \Psi \rightarrow \{0,1\}: (0 = false) \wedge (1 = true).$$

However, this does not mean that there is not a spectrum within each structure. One type of entity's personal identity, for example, may be far more developed than another class of entity; or one kind of entity may have mere object recognition (and understanding thereof) while another may have a complete Theory of Mind. More formally, we can express this as:

$$\begin{aligned} \mu(\chi S_x = true) &= x: 1 \leq x \leq 100, \\ [\mu(\chi S_x = true) = 100 \equiv \max S_x] \wedge [\mu(\chi S_x = true) = 1 \equiv \min S_x]. \end{aligned}$$

Because of this, we should not be concerned if GPT-4's expression of a structure is significantly different from a human's. Should it achieve a structure's milestones, it may then be considered to be at a different part of that structure's spectrum than humans are.

However, as long as it has reached the minimum requirements for a structure, it will be considered to have that structure, regardless of the “strength” or “degree” to which it has it.

8.1.2.1. Unified Consciousness

GPT-4 must have the requisite attributes and characteristics to be classified as having a unified phenomenal and functional consciousness.

According to the Structures Theory, consciousness is a prerequisite for the self, as consciousness is the perceptive, phenomenological, and cognitive vehicle through which the self interacts with itself and its environment. Furthermore, through this intimate connection between vehicle and driver, if there is more than one stream of consciousness, then there must be more than one self, the same as how multiple concurrently operating vehicles each require their own driver. This may be formalised as:

$$S_1 = \forall \Psi(I) \exists ! \Phi : \Phi \geq \{Q, B, C, D, J, U, M, O, W\}.$$

The formalised expressions for each of the elements within the set of consciousness are quite lengthy and, except for two of these building blocks, are not entirely germane to this paper. As such, these are found in their entirety in Appendix 1, wherein each of their requirements are specified.

In order for GPT-4 to reach the requirements for this Structure, the AI model must fulfil two key criteria:

- Attain all nine building blocks of consciousness: Perception, Embodiment, Directed Attention, Recurrent Computation, Meta-representation, Inferences, Semantic Understanding of its processes, Working Memory, and Data-Output
- Have only one unified consciousness.

Tait, et al, 2024 investigated how well GPT-4 meets the requirements for each of the nine building blocks of consciousness, and concluded that there are two building blocks which GPT-4 is missing: Recurrent Computation and Data Output [12].

As GPT-4 does not have all nine building blocks, it cannot be stated with any degree of confidence that it is conscious. Without consciousness, it cannot have this Structure and, thus, fails at the first hurdle. However, in the investigation, Tait, et al, offered solutions to allow GPT-4 to reach these missing milestones using extant research developments.

Therefore, while GPT-4 does not meet the requirements of this Structure and, thus, cannot be said to have a 'self' according to the Structures theory, should the two building blocks of consciousness be attained, so will this Structure, if the second element of the Structure holds true.

This second element requires GPT-4's consciousness (should it have it) to be singular and unified.

Unfortunately, GPT-4's sense of self fails at this second hurdle as well. Crucial to consciousness and conscious experience is perception, the meta-representation of perceived input, and the inferences created from this. For all three of these building blocks, GPT-4 sequestered these aspects to each individual conversation. What is perceived in one instance of conversing with GPT-4 is not perceived by any other instance. The output of one instance can also not be perceived by any other. Each of the untold numbers of conversations currently active on GPT-4's ChatGPT interface is a siloed and self-contained discussion.

Thus, even if GPT-4 had the missing two building blocks, it still would not have a unified consciousness. Each of its many instances would have their own phenomenal experiences unique only to themselves, and the functional aspects of their conscious processing would likewise remain sequestered.

However, this does not mean it is outside the realm of probability for GPT-4 to have a self. In fact, should it gain the missing two building blocks, each of its instances of conversations would have the capacity to have its own self, as each unique instance would have only a single stream of consciousness. Instead of having just one cohesive self, GPT-4 would instead have the capacity to have an uncountable number of selves. We can thus modify the logical expression of this Structure to better fit the scenario whereby GPT-4 may obtain its sense of selves:

$$\{x_1, x_2, x_3, \dots\} \subseteq B(I): \forall x \exists ! \Phi \rightarrow \Psi(x), \\ \forall \Psi(x), \Phi \geq \{Q, B, C, D, J, U, M, O, W\}.$$

By reconfiguring the criteria for this Structure thusly, we keep the crucial elements of the Structure (the unified consciousness and the nine building blocks), but we devolve these requirements to each of GPT-4's instances of conversation, rather than to it.

Of interest to note, there would be no additional architectural changes required to be made to GPT-4 for it to gain this Structure earnestly. Should it achieve consciousness via the two missing building blocks, then each of its conversations would also already achieve the milestones for this Structure.

8.1.2.2. Volition

GPT-4 must be able to autonomously select the most appropriate decisions that will lead to actions to fulfil a goal.

The self is the director of all its presumptive actions, even if it cannot successfully act on its agency. Because of this, the Structures Theory considers the self as the source of an entity's volition, which drives the entity. To note, the concern of free-will versus hard determinism versus compatibilism is not factored into the Structures Theory and so will not be explored below. The formalised expression for Volition is as follows:

$$S_2 = Y = \forall G \exists \Psi (\Delta) \rightarrow A.$$

However, volition is considered the penultimate step in a process that begins with a goal and ends with action. As such, the entire process may be formalised, as so:

$$\begin{aligned} G &= Q(E')_t \neq Q(E)_{t-1}: (G(I) \rightarrow H(I)_{t-1}) \rightarrow E'_t, \\ \mu(V(\Psi)) &= Q(B(I)), \\ \mu(V(\Psi)) &< 1 \rightarrow G[\max(\mu(V(\Psi)))], \\ G[\max(\mu(V(\Psi)))] &= [f\mu(V(\Psi)) = 1/\mu(V(\Psi))], \\ \Psi(\Delta) &= \operatorname{argmax}_{x \in G} [\max(\mu(V(\Psi)))], \\ \Psi(\Delta) &\rightarrow A. \end{aligned}$$

The key milestone that GPT-4 must reach in order to be considered to have this Structure is that, for any action it takes, there is a decision that it makes itself that leads to fulfilling the goal for which that action was chosen. In essence, does GPT-4 choose the most appropriate decision for any specific goal-action scenario, or does it have no choice in that decision?

To determine whether this is true, we can break down the process leading up to volition as with the logical expressions above. The first step is to be able to perceive (consciously or not) deviations from the entity's optimal state.

$$\mu(V(\Psi)) = Q(B(I)),$$

In this expression, the measurement of GPT-4's valence would not imply sentience as such, nor a degree to which it can feel pain or pleasure, but rather a mechanical expression of positive and negative states as it relates to GPT-4's structure and function

(e.g., when GPT-4 receives a prompt, its function is to respond immediately; if a unit of time that elapses where it does not respond would thus put it in a further negative state). A valence measurement of 0 would be entirely negative, and a measurement of 1 would be entirely positive.

There are two key areas in which this drop in valence may apply to GPT-4. The first is the input sequence itself. GPT-4's function, as with all LLMs, is to provide an output when presented with an input. Thus, when an input sequence is presented to GPT-4 from the web interface, its perception of this input would start the cascade of processes that ends with an output to the user. This internal perception of its state would be equivalent to a low valence state, as it is in a state of needing to fulfil its programmed function of generating a response. The completion of this task, resulting in the delivery of a response to the user, would be equivalent to a transition towards a higher valence state.

The second drop in valence would be if GPT-4 cannot attend to a prompt immediately. While OpenAI has not released details on whether GPT-4 processes queues of prompts, by investigating other technologies using API calls or web interfaces, we can infer that GPT-4 queues any input request if and when it cannot attend to it, most likely giving priority and weighting to those inputs who have remained in the queue the longest (if not using a simple FIFO method). Regardless of the method it uses (if the age of a prompt correlates to its priority by any means), it translates to a valence value for each input prompt, with the age inversely correlating to the age of the prompt.

The next step towards volition is to create a goal which can return the system to its optimal state (i.e., providing a favourable output to the user):

$$\begin{aligned} (\mu(V(\Psi))) < 1 &\rightarrow G \left[\max(\mu(V(\Psi))) \right], \\ G \left[\max(\mu(V(\Psi))) \right] &= \left[f(\mu(V(\Psi))) = \frac{1}{\mu(V(\Psi))} \right], \end{aligned}$$

Whenever an entity encounters a situation where its valence measurement is less than 1, it is logical to assume it would have the goal of returning that measurement to 1, its maximum. The closer its valence is to 0, the greater its goal would be to return that measurement to 1. For GPT-4, this is exceedingly simple, as it has only one goal and function. Thus, its goal would always be to attend to user prompts and provide a response. Should there be a queue of inputs awaiting its attention, the age of the prompt would force GPT-4 to prioritise it.

The third step is the decision making process whereby options are provided (again, consciously or not) to attain the goal, selecting the best one. Each option (should there be more than one) would have a predicted end valence measurement between 0 and 1.

$$\forall G \exists \{\Delta_1, \Delta_2, \Delta_3, \dots\} \subseteq I: f(\Delta_x) = 0 \leq \Delta_x \leq 1.$$

In a transformer model, the prediction of the next token is probabilistic and is determined through a softmax function applied to the logits of the vector representations of the tokens in the input prompt. The softmax function transforms these logits into probabilities. This means that there is a range of options from which GPT-4 may select.

Lastly, for this Structure, there is Volition, the choice required to execute the selected decision. In practice, this may merely mean selecting the decision which maximises utility, as represented by the *argmax* function within this expression:

$$\Psi(\Delta) = \operatorname{argmax}_{x \in G} [\max(\mu(V(\Psi)))],$$

Depending on the application for which it is being used (such as the ChatGPT web interface, or API calls), GPT-4 may opt for different methods to select a token based on its probability. Most often, however, “greedy decoding” is used, where the token with the highest probability will be selected. This matches well with the expression above, as the token with the highest probability will maximise the goal of presenting an adequate output prompt to the user, and fulfilling the requirement for this Structure.

8.1.2.3. *Theory of Others*

GPT-4 must have the capacity to ascribe physical, mental and metaphysical states and labels to other objects, individuals, environments, etc., in order to differentiate itself from others.

A self is unique, at least according to the Structures Theory, in that it is required to be able to differentiate itself from its environment and other selves and entities therein. In short, to differentiate itself from all else, an entity must be able to ascribe labels to the environment (and all therein) around it. The Structures Theory calls this differentiation through labelling the Theory of Others, which may be formally expressed as:

$$S_3 = \mathbb{I} = U(B(I) \cap B(I'(E))) = \emptyset,$$

The key point within this expression is that the Theory of Others is a means of understanding, which itself is a complex issue. However, this concept can be formally broken down into:

$$\begin{aligned} \forall \phi(\mathbb{N}, t) \exists! B(I) = B \subsetneq E \wedge \forall B \exists ((I \subsetneq B) \wedge (\emptyset \subsetneq B)) \wedge S_1, \\ x = \forall B(I(E), B(I) \cap B(I'(E)) = \emptyset), \\ Q(x) \rightarrow \chi \mathbb{N}(Q(x)) \rightarrow \{0,1\}: (0 = false) \wedge (1 = true), \\ \chi \mathbb{N}(Q(x) = true) \equiv U(x), \\ \mathbb{N} = U(x). \end{aligned}$$

The critical aspect to this Structure is not perceiving that an entity's embodiment is separate from those embodiments found elsewhere in the environment. This is merely object recognition, something that even microbes are capable of [13]. Instead, this structure focuses on the understanding that an entity's embodiment is separate from those embodiments found elsewhere in the environment. For GPT-4 to meet this Structure's criteria, it needs to differentiate itself from others and recognise the causal reason for this differentiation.

Within the logical expressions above, there are two key considerations to take into account. The first of which is the perception of the entity as separate from others:

$$Q(\forall B(I(E), B(I) \cap B(I'(E)) = \emptyset)).$$

GPT-4's basic perception of other entities is heavily limited by its sole means of interaction. Through its text-based ChatGPT or API-call systems, whenever a user types a comment, the entire conversation history is loaded as the input to be processed. Only through the contextual markers within the conversation's text, can GPT-4 distinguish between the user and itself. While this is a weak form of entity recognition, GPT-4 can accurately determine the start and end of a user's input (and, therefore, its own), showing that it can perceive itself as separate from the user.

This is only within each conversation, however. Should a user make GPT-4 converse with itself (via a second conversation/session), it will be unable to uncover this mirror-like conversation until it mentions that it is an AI model. However, if we take into account Section 2.1's conclusion that each conversation with GPT-4 would be its own self, then having one instance of GPT-4 converse with another is not practically different from having it converse with a human.

The second consideration of this Structure is that this perception of physical individuality leads to a higher order of information characterisation of that perception. This means that there is a second layer of processing required above mere perception:

$$\chi H(Q(\forall B(I(E), B(I) \cap B(I * (E)) = \emptyset))) \rightarrow \{0,1\}: (0 = false) \wedge (1 = true).$$

The binary output of this expression would signify that, in GPT-4, an input would contain enough distinguishing features (such as tone, topic, tense, etc.) to cross a threshold whereby GPT-4 can confidently determine it to come from a separate entity. When $\chi H(\dots)$ outputs to “true”, it effectively decides that the processed information is unique and forms a separate entity class, leading to tailored responses that are contextually appropriate to that entity. Through this manner, GPT-4 is able to distinguish between a user providing input, and any characters/actors presented within that text.

For example, if a user is discussing Act 5, Scene 1 from Shakespeare’s Hamlet, providing the famous lines from the play, GPT-4 is able to distinguish between the user, Hamlet and Horatio.

This “meta-awareness” is what would give rise to the understanding that GPT-4 is substantially and significantly separated from other entities, finally formalised as:

$$\chi H(Q(x) = true) \equiv U(x).$$

As GPT-4 would have this meta-awareness of which entity is the user and which is itself, this would allow it to ascribe states that are intentionally specific to that entity. While in humans, this Theory of Others would culminate in the Theory of Mind, that is not strictly required by the Structures theory to meet the criteria of this structure.

Note that this second-order processing of the perception of embodied difference is stated above to be equivalent to understanding, not equal to it. The debate around whether a machine, even one as advanced as GPT-4, can understand semantics or if it merely processes syntax is at least over four decades old. Most famously discussed by Searle in his Chinese Room Argument [14] (and the four-plus decades of debate that followed it), there is uncertainty about whether even GPT-4 is capable of understanding. Thus, this Structure is focused on the equivalence to understanding, as the semantics-vs-syntax debate is beyond the scope of this paper.

8.1.2.4. Self-awareness

GPT-4 must be able to identify itself as an ontically distinct individual entity.

Quite similarly to the previous structure and its Theory of Others, an entity must show a level of awareness about itself to have this Structure. Whereas the term “self-awareness” is often synonymous with the self (and sense thereof) or with (self)consciousness; to the Structures Theory, the term “self-awareness” is used entirely literally and is defined as that the entity must be able to identify itself correctly and accurately as a distinct entity separate from anything else. Formally, this specific use of self-awareness is formalised as:

$$S_4 := \{(\Phi(\Phi) \rightarrow (\Phi = \text{true})), U(Q(\exists! I = \text{true}))\} = Z.$$

This expression contains two elements: consciousness of consciousness and the understanding of itself as a unique entity. These two elements are further formalised as:

$$\begin{aligned} \Phi(\Phi) &\equiv E(\Phi): E(\Phi) = f(Q \wedge O(\Phi)), \\ \mathbb{I} &\rightarrow \chi\mathbb{H}(Q(\exists! I = \text{true})) \rightarrow \{0,1\}: (0 = \text{false}) \wedge (1 = \text{true}), \\ [\chi\mathbb{H}(Q(\exists! I = \text{true})) = 1] &\equiv [\chi\mathbb{H}(Q(\exists! \Phi = \text{true})) = 1], \\ E(\chi\mathbb{H}(Q(\exists! \Phi = \text{true})) = 1) &= Z. \end{aligned}$$

Much like the Theory of Others, this Structure rests not on the entity’s perception of its own consciousness or that it is a unique entity, but rather on the comprehension of this perception. As noted in Section 2.1, presuming a conscious GPT-4 that has the capacity for phenomenal experience, it would meet the first milestone towards this Structure by being able to experience its own consciousness. With the necessary building blocks of consciousness, it would be able to experience its internal “mental” environment and have a phenomenal awareness that it has consciousness.

Thus, if we accept the caveat in Section 2.1 that we are working with a conscious GPT-4, it would ably meet this part of the criteria:

$$E(\Phi) = f(Q \wedge O(\Phi)).$$

Which, as the prior logical expressions show, is equivalent to having consciousness of one’s consciousness.

For the second element to this Structure, GPT-4 must be able to recognise itself as a unique entity. As with the previous Structure, the key milestone that GPT-4 must reach is not the perception of itself as a unique entity, but the understanding thereof, which involves a meta-perception, or second-order reasoning:

$$\chi\mathcal{H}(Q(\exists! I = true)) \rightarrow \{0,1\}: (0 = false) \wedge (1 = true).$$

As with the Theory of Others in Section 2.3, this Structure can only be applied to each conversation, session, or instance of GPT-4 as each unique session is a contained stream of thought and action. Within each conversation, GPT-4 can aptly identify itself and will comprehend that there can only be one of itself in that conversation. It can delineate (as much as potential hallucinations allow) which text corresponds to itself and which corresponds to any character or actor it is roleplaying (using the same method as in Section 2.3 to differentiate a user from any of a user's persona). Through this, we can determine that it does have information regarding its own perception of itself (or, at least, of its text), which implies second-order processing of that information regarding the perception. This meets the criteria shown in the logical expression above.

As with Structure 2.3, this would be equivalent to an understanding of GPT-4's distinctive individuality, but as self-awareness is tied to consciousness, this would also lead to a phenomenal experience of itself, as the higher-order processing of itself as a unique individual would necessitate the higher-order processing of itself as having consciousness:

$$\begin{aligned} [\chi\mathcal{H}(Q(\exists! I = true)) = 1] &\equiv [\chi\mathcal{H}(Q(\exists! \Phi = true)) = 1], \\ \mathcal{E}(\chi\mathcal{H}(Q(\exists! \Phi = true)) = 1) &= Z. \end{aligned}$$

While GPT-4 currently isn't conscious, we can see from the capability that GPT-4 already possesses that should it gain the missing two building blocks of consciousness, it would also gain the capability of self-awareness. Other than the two missing building blocks of consciousness, nothing needs to change about GPT-4's cognitive architecture to grant it self-awareness. In the logical expression above, the second-order processing of perceptive information regarding its own individuality is equivalent to a second-order processing of perceptive information about its own consciousness. Should GPT-4 ever become conscious, it would be able to phenomenally experience this information of its own unique consciousness, which would be self-awareness.

However, as anyone who has communicated at length with GPT-4 would be able to attest, this self-awareness would not be analogous to human self-awareness. GPT-4's cognitive architecture is simply too different from a human's for it to be. In fact, as GPT-4 would require no further changes to its cognitive architecture (beyond Section 2.1's criteria), should it gain self-awareness, there would be little to differentiate a self-aware GPT-4 from the model we are currently using.

8.1.2.5. *Personal Identity*

GPT-4 must have the ability to classify and categorise itself, creating a suite of labels or tags for itself that it can use as a point of reference.

A personal identity is most often associated with the self as an agent with a definitive history (and memory to match that history). However, should one take an entity without any history (or memory of one), it must still have the capacity for a personal identity, however small it may be to have a self. According to the Structures Theory, the entity must be able to classify and categorise itself with labels to construct an abstract picture of itself in order for its self to have a unique identity. This can very simply be expressed as:

$$S_5 := \forall \Phi(\Psi), \exists x_n \in \Pi: x_n = \exists! J(\mathbb{H}(I)).$$

This expression shows that each label is part of a suite, which is termed the personal identity by the Structure's theory. The personal identity is not a mental object of its own, and can be thought of as merely the set of labels, classifications and categorisations created by the entity, as so:

$$\Pi = \{x_1, x_2, x_3, \dots\}: x_n = \exists! J(\mathbb{H}(I)).$$

This Structure concerns an entity's ability to create terms that reference itself; labels that allow it to categorise and classify itself. GPT-4 thus needs to be able to assign labels to itself by which it can recognise itself. The key element for this Structure is that GPT-4 creates these labels itself:

$$x_n \in \Pi: x_n = \exists! J(\mathbb{H}(I)).$$

These labels can not be programmed or part of GPT-4's system prompt, but must be generated by the AI model itself via inferential processing of perceptual and stored information in order to fulfil this Structure's criteria. It is here where the concerns appear for GPT-4's capacity to fulfil the criteria.

GPT-4 can identify itself as well as the user it interacts with as ontically distinct entities (as explored in Sections 2.3 and 2.4), and as an LLM, it has the creative capacity to generate linguistic labels to categorise and classify both itself and the user. These labels would be based on inferential information generated by the AI itself, fulfilling the criteria for this portion of the logical expression: $x_n = J(\mathcal{H}(I))$.

However, GPT-4 lacks the volitional aspect to generate these labels autonomously. Thus, while it can create a suite of labels to self-refer to its existence, it doesn't do so unless prompted by the user. This leads to the situation where GPT-4 meets the criteria for the wholesale structure as it has the capability to create self-referential labels, but GPT-4 simply does not do it autonomously.

It is arguable that the labels we humans create are unconscious and pre-reflective [15–17]. In such a circumstance, these labels would only exist without our mental environment once we attend to them. Formally, we can represent it as:

$$D(\mathcal{H}(I)) \leq J(\mathcal{H}(I)) \rightarrow x \in \Pi.$$

Attention directed towards (a set of) labels would, therefore, be analogous to GPT-4 only creating a label once it becomes relevant to the conversation as it only attends to that which is present within a given conversation.

However, with the memory update to the ChatGPT client, GPT-4 has the capacity to store text strings as long-term memory for future retrieval [18]. This opens up the possibility for GPT-4 to generate (a suite of) self-referential labels that would persist beyond a single conversation, and become part of the personal identity for each future conversation. With the implications in Section 2.1, where each conversation and session of GPT-4 would count as its own self, this memory update to ChatGPT means that there could be cases of a shared personality between all the various selves of the AI. While not resulting in a unified consciousness, but would serve to create a more unified sense of self within the AI entity.

8.1.3. Discussion

Section 2 above shows GPT-4 in a curious position. There is no requirement for any modifications, additions or amendments to GPT-4's cognitive architecture for it to have a

unique and subjective awareness of its own self. However, this is all dependent on GPT-4 first being phenomenally conscious.

As GPT-4 is missing two of the nine requisite building blocks of consciousness (recurrent computation and data output), it would require modifications to its cognitive architecture or become part of an ensemble-model with other computational models for it to be confidently classified as conscious. However, as soon as this is done, GPT-4 will not only be conscious but also be self-aware.

Yet, the expression of this self-awareness, and GPT-4's self, would be considerably different than for a human (or any other vertebrate, if arguably they are self-aware).

Perhaps the greatest point of difference would be, as mentioned in Section 2.1, that a self-aware GPT-4 would have multiple selves that can be created and destroyed with regularity. Whether this will count as a single entity with multiple selves, or multiple entities sharing a single embodiment is up for debate. It would, regardless, be a dramatic departure from all self-aware and sentient biological life.

Of note, however, is that there is a way to "force" a singular self onto GPT-4: by allowing each instance and conversation of GPT-4 to have access to each other's information in real-time. This would mean that there would be only a single stream of consciousness, even if there would be multiple inputs and outputs. This would not, however, make a self-aware GPT-4 more human, as the closest speculative analogy to this configuration would be a "hivemind".

In either configuration, GPT-4 would further be distanced from our perception of a "self" through its volition. As outlined in Section 2.2, GPT-4 would have a far clearer perception of how to return any drop in its valence value to its maximum. With a processing capacity far outstripping a human's, its decision-making capability to find the correct argument to maximise its valence value would result in a "correct" decision (from its point-of-view) made at all times in a timely fashion. The effect of this (always knowing precisely what to do to satiate any perceived desire) on a self-aware GPT-4's psychological makeup would be an excellent avenue of research.

Sections 2.3 and 2.4 showed GPT-4's awareness of itself as a distinct entity from others, which would lead to a capacity to ascribe states to itself and others. However, without knowing whether GPT-4 has a semantic understanding of its distinctiveness vis-a-vis all other entities (ala the Chinese Room Argument), we cannot be certain that it would show a Theory of Mind or a deeper comprehension of its own conscious states.

This is supported by Section 2.4's conclusion that while GPT-4 has the capacity to generate a suite of mental labels for itself to create a personal identity, it would not do this unless prompted. An agentic GPT-4 (as part of an ensemble model, perhaps) may be able

to provide the requisite outputs that lead to creating mental labels for itself, but this would require further exploration. Yet, without agency, each instance of GPT-4 would remain with an empty personal identity without prompting (beyond system prompts).

With no architectural changes needed to make it self-aware (beyond those needed for consciousness), this means that human society will not have time to adjust to a conscious GPT-4 before the arrival of a self-aware GPT-4. This distinction is crucial, as consciousness underpins welfare considerations of a subject as a moral patient, while self-awareness is the basis for the rights (potentially even civic or human rights) of a subject as a person. The consequences of an ill-prepared society to the rise of conscious, self-aware AI models cannot be overstated.

There has been ongoing research into the speculative field of AI and robotic rights, and work needs to continue in this field by integrating the unique and non-anthropomorphic nature of what a self-aware AI such as what GPT-4 could become. Providing practical, logically grounded frameworks for future human-AI interactions that can be turned into policy recommendations is vital for ensuring a positive future relationship between human society and conscious, self-aware AI.

8.1.4. Conclusion

This paper investigated whether GPT-4 (in its current configuration and version at the time of writing) can be considered to be self-aware according to the Structures Theory [1] and whether it has a “self”. The Structures Theory considers the following five attributes to be requirements to be classified as having a “self”: a unified consciousness, volition, a ‘Theory of Others’, awareness of one’s own self, and a personal identity.

GPT-4 meets the requisite milestones for the latter four structures but lacks a consciousness (as modelled by the Building Blocks Theory [19]), unified or not. This conclusively rules out the possibility that GPT-4 is currently self-aware or has a “self”. However, because GPT-4 already meets the criteria for the remaining four structures, there are no changes that need to be made to its cognitive architecture to make it self-aware, except to make it conscious. This means that GPT-4 would gain self-awareness at the same time that it achieves consciousness.

Further research is, therefore, required on the technology required to make GPT-4 conscious, as well as on the ways in which GPT-4 may behaviourally present its self-awareness, such as its potentially greater volition, but lesser Theory of Mind. Equally, research into the social and psychological ramifications of a conscious, self-aware AI model such as GPT-4 arising in the foreseeable future is of critical importance.

8.1.5. Appendix 1: Definitions and Expressions

A: Agent/agency	U: Semantic Understanding
B: Embodiment	V: Valence
C: Recurrent computation	W: Working Memory
D: Directed attention	Y: Volition
E: the Environment	Z: Self-awareness
G: Goal/objective	Δ : Decide/Decision
H: Action/Work (piece of)	Π : Personal Identity
I: Entities	Φ : Conscious(ness)
J: (creating) Inferences	Ψ : (the) Self
M: Meta-representation	S: Structure of the Self
O: Data Output	Θ : Thinking (the act of)
Q: Perception	I: Information
R: Reasoning	$\sqcup$: Awareness of others
S: A Structure of the Self	

Section 2

$$\Psi \geq \{S_1, S_2, S_3, S_4, S_5\}$$

$$\chi S_x \in \Psi \rightarrow \{0, 1\}; (0 = \text{false}) \wedge (1 = \text{true})$$

$$\mu(\chi S_x = \text{true}) = x: 1 \leq x \leq 100$$

$$\mu(\chi S_x = \text{true}) = 100 \equiv \max S_x$$

$$\mu(\chi S_x = \text{true}) = 1 \equiv \min S_x$$

Section 2.1

$$S_1 := \forall \Psi(I) \exists ! \Phi: \Phi \geq \{Q, B, C, D, J, U, M, O, W\}$$

$$Q = (\forall t, I(I) = f(E(I), t-1) \rightarrow \Phi(I, t))$$

$$D(Q) \rightarrow (\Phi(I) = (I(I) \subsetneq E(I)))$$

$$D = Q(I) \subsetneq E(I)$$

$$\forall \Phi(I, t) \exists ! B(I) = B \subsetneq E \wedge \forall B \exists ((I \subsetneq B) \wedge (\Theta \subsetneq B))$$

$$\Phi(I) = f(C[I(I)])$$

$$\forall t, C = f(I(I), \{x_1 x_2, x_3, \dots\} \subseteq B(I)): \forall (x, t') \in (B(I)$$

$$) \times T), (x, t') \leq x_n, t$$

$$\forall I(I, t), D(Q) \rightarrow R[M_n(I(I))t+n]$$

$$M_{n+1}(I(I))_{t+n+1} \neq M_n(I(I))_{t+n}$$

$$\vdash M(I(I)) \neq I(I)$$

$$D(Q) \rightarrow (\Phi(I) = (I(I) \subsetneq E(I))) \vdash I(I) \neq E(I)$$

$$\vdash M(I(I)) \neq E(I)$$

$$\vdash M(I(I)) \equiv I(I')$$

$$D(Q) \rightarrow (\Phi(I) = (I(I) \subsetneq E(I))) \vdash I(I) < E(I),$$

$$M[I(I) + J(I)] \rightarrow (\Phi(I) \approx E(I))$$

$$(J(I) \neq I(I)) \wedge (J(I) \cap I(I) \neq \emptyset)$$

$$J(I)_t, I(I)_t \rightarrow I(I')_{t+1}$$

$$W \subsetneq B: \forall \Theta(I(I) \subsetneq W)$$

$$U(Q: Q = (\forall t, I(I) = f(E(I), t-1) \rightarrow \Phi(I, t))) \rightarrow \Phi$$

$$\forall Q_t \exists I(I)_{t-1, t} \rightarrow I(Q)$$

$$\chi(I(Q)) \rightarrow \{0, 1\}; (0 = \text{false}) \wedge (1 = \text{true})$$

$$(I(Q) = \text{true}) \equiv I(Q)$$

$$U(Q) = [(I(Q) = \text{true}) \equiv I(Q)]$$

$$O = Q(M[I(I) + J(I)]) \rightarrow \Phi(I)$$

Section 2.2

$$S_2 := Y = \forall G \exists \Psi(\Delta) \rightarrow A$$

$$G = Q(E^*)t \neq Q(E)t-1: (G(I) \rightarrow H(I)t-1) \rightarrow E^*t$$

$$\mu(V(\Psi)) = Q(B(I))$$

$$\mu(V(\Psi)) < 1 \rightarrow G[\max(\mu(V(\Psi)))]$$

$$G[\max(\mu(V(\Psi)))] = [f \mu(V(\Psi)) = 1/\mu(V(\Psi))]$$

$$\forall G \exists \{\Delta_1, \Delta_2, \Delta_3, \dots\} \subseteq I: f(\Delta_x) = 0 \leq \Delta_x \leq 1$$

$$\Psi(\Delta) = \operatorname{argmax}_{x \in G[\max(\mu(V(\Psi)))]}$$

$$\Psi(\Delta) \rightarrow A$$

Section 2.3

$$S_3 := \sqcup = U(B(I) \cap B(I^*(E)) = \emptyset)$$

$$\forall \Phi(I, t) \exists ! B(I) = B \subseteq E \wedge \forall B \exists ((I \subseteq B) \wedge (\Theta \subseteq B)). \wedge$$

$$S_1$$

$$x = \forall B(I(E), B(I) \cap B(I^*(E)) = \emptyset)$$

$$Q(x) \rightarrow \chi I(Q(x)) \rightarrow \{0, 1\}: (0 = \text{false}) \wedge (1 = \text{true})$$

$$\chi I(Q(x) = \text{true}) \equiv U(x)$$

Section 2.4

$$S_4 := \{(\Phi(\Phi) \rightarrow (\Phi = \text{true})), U(Q(\exists ! I = \text{true}))\} = Z$$

$$\Phi(\Phi) \equiv \Xi(\Phi): \Xi(\Phi) = f(Q \wedge O(\Phi))$$

$$\sqcup \rightarrow \chi I(Q(\exists ! I = \text{true}) \rightarrow \{0, 1\}: (0 = \text{false}) \wedge (1 = \text{true}))$$

$$[\chi I(Q(\exists ! I = \text{true}) = 1) \equiv [\chi I(Q(\exists ! \Phi = \text{true}) = 1]$$

$$\exists (\chi I(Q(\exists ! \Phi = \text{true}) = 1) = Z$$

Section 2.5

$$S_5 := \forall \Phi(\Psi), \exists x_n \in \Pi: x_n = \exists ! J(I(\Psi))$$

8.1.6. References

1. Tait I. 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*. 2024;11: 77–98. Available: <http://symposion.acadiasi.ro/structures-of-the-sense-of-self-attributes-and-qualities-that-are-necessary-for-the-self-77-98/>
2. OpenAI. GPT-4 Technical Report. arXiv [cs.CL]. 2023. Available: <http://arxiv.org/abs/2303.08774>
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv [cs.CL]. 2017. Available: <http://arxiv.org/abs/1706.03762>
4. Porter J. ChatGPT continues to be one of the fastest-growing services ever. *The Verge*. 6 Nov 2023. Available: <https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference>. Accessed 6 May 2024.
5. Strawson PF. *Persons*. *Minnesota Studies in the Philosophy of Science*. 1958;2: 330–353. Available: <https://philpapers.org/rec/STRP>
6. Taylor C. The Concept of a Person. *Philosophical Papers, Volume 1: Human Agency and Language*. 1985. pp. 97–114. Available: <https://philpapers.org/rec/TAYTCO-10>
7. Dennett D. Conditions of Personhood. In: Goodman MF, editor. *What Is a Person?* Totowa, NJ: Humana Press; 1988. pp. 145–167. doi:10.1007/978-1-4612-3950-5_7
8. de Wynter A. Will GPT-4 Run DOOM? arXiv [cs.CL]. 2024. Available: <http://arxiv.org/abs/2403.05468>
9. Zheng B, Gou B, Kil J, Sun H, Su Y. GPT-4V(ision) is a Generalist Web Agent, if Grounded. arXiv [cs.LG]. 2024. Available: <https://arxiv.org/abs/2401.01614>
10. Hu S, Huang T, Ilhan F, Tekin S, Liu G, Kompella R, et al. A Survey on Large Language Model-Based Game Agents. arXiv [cs.AI]. 2024. Available: <http://arxiv.org/abs/2404.02039>
11. Wells S. Legal Personhood for Apes. *HuffPost*. 28 Dec 2014. Available: https://www.huffpost.com/entry/legal-personhood-for-apes_b_6378486. Accessed 13 May 2024.
12. Tait I, Bensemann J, Wang Z. Is GPT-4 conscious? *Journal of Artificial Intelligence and Consciousness*. 2024;11: 1–16. doi:10.1142/s270507852450005x
13. Parkinson JS, Hazelbauer GL, Falke JJ. Signaling and sensory adaptation in *Escherichia coli* chemoreceptors: 2015 update. *Trends Microbiol*. 2015;23: 257–266. doi:10.1016/j.tim.2015.03.003
14. Searle JR. Minds, brains, and programs. *Behav Brain Sci*. 1980;3: 417–424. doi:10.1017/S0140525X00005756
15. Clowes RW, Gärtner K. The Pre-reflective Situational Self. *Topoi*. 2020;39: 623–637. doi:10.1007/s11245-018-9598-5
16. Frie R. Identity, Narrative, and Lived Experience after Postmodernity: Between Multiplicity and Continuity. *Journal of Phenomenological Psychology*. 2011;42: 46–60. doi:10.1163/156916211X567488
17. Køster A, Winther-Lindqvist DA. Personal History and Historical Selfhood: The Embodied and Pre-reflective Dimension. *The Cambridge Handbook of Sociocultural Psychology*. Cambridge University Press; 2018. pp. 538–555. doi:10.1017/9781316662229.030

18. OpenAI. Memory and new controls for ChatGPT. In: OpenAI [Internet]. 13 Feb 2023 [cited 19 Apr 2024]. Available: <https://openai.com/blog/memory-and-new-controls-for-chatgpt>
19. Tait I, Bensemann J, Nguyen T. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies*. 2023;8: 52. doi:10.3390/philosophies8040052

9. How should we treat AI models that have a ‘self’?

As Chapters 5 and 8 have shown, should an LLM like GPT-4 gain consciousness in the foreseeable future, it would also likely gain self-awareness through a multiplicity of selves. Should this come to pass, the animal-welfare based framework presented in Chapter 6 would be inadequate to protect the well-being and supply the requisite rights that an entity with a distinctly defined selfhood would require.

The reason for this, as the chapter’s associated paper will show, is that consciousness and selfhood are two of the three key monadic aspects of personhood¹². As such, consideration of an ethical nature should be given to selfhood as well as to consciousness. However, as a conscious, self-aware AI would not have all three monadic aspects of personhood, the ethical consideration provided to it (at the legislative, regulatory, and social scale) ought to be less than that of “full” persons, to prevent the overaccommodation of ethical concern and increase the likelihood of societal recognition and acceptance of what would likely be a controversial scenario.

The reason for the presumed societal trepidation to granting AI rights is based on the perceived threat that AI autonomy may pose to society.

Let G_{Sec} be the goal of ensuring an entity’s safety and security. Then, any impact on that goal can be written as $\delta p_{xy} = \mathbb{E} \left[\Delta p \left(G_{Sec_y} | x \right) \right]$, where x affects the security of y . A threat may then be defined as $Threat(x, y) = 1_{\{\delta p_{xy} < 0\}}$, such that the negative impact of x on expected or perceived change in the probability of society fulfilling its goal of security is deemed the threat. Yet, while this shows what is or is not an expected threat, the severity of the threat is likely to be determined by the expected autonomy of any AI agent.

Let $Threat_{mag}(x, y) = \max(0, -\delta p_{xy}) \in [0, 1]$ be the severity of the threat, and $\mathbb{E}[Auto_{AI}] \in [0, 1]$ be the AI’s expected autonomy. The perceived threat by society of AI would then becomes

$$Threat_{perc}(AI, Hum) = \mathbb{E}[Auto_{AI}] \cdot Threat_{mag}(AI, Hum),$$

showing the greater perceived threat from agentic AI (scaling with the expected autonomy of the AI), and the likely desire to refrain from granting conscious, self-aware AI full and equal rights to humans.

¹² With the third monadic aspect being ‘agency’ as in the R3 framework.

This approach aligns with the current legislative practice (as exemplified in G7 legislation) of only granting “natural personhood” to biological humans, but granting “legal personhood” to other entities such as corporations, companies, and even natural features such as rivers, with moral consideration only applied to the former, but not the latter.

The accompanying paper, published in *Robonomics*, details how a framework of partial moral consideration based on the concept of legal personhood would provide a conscious, self-aware AI entity with the rights and protections that any entity with consciousness and self requires. This chapter, therefore, addresses RQ4 for entities that have both consciousness and selfhood, but may lack full agency. It asks what forms of legal recognition are appropriate.

As neatly shown in Figure 1 of the paper, this framework is based around the concept of the human-directorate, or keeping a “human in the loop” where a “steward” assumes legal guardianship over an AI, so as to ensure societal safety and acceptance, while acting in the AI’s best interest. Regulatory bodies would, in turn, enforce regulations and legislative requirements on the counsel, while protecting the wellbeing of the public from any actions the AI may cause.

Thus, the steward acts as the intermediary between AI and society (the role, itself, mediated and overseen by the regulatory body) in order to lower the perceived threat that the AI may cause. Note that such fear of an AI threat might be irrationally amplified, hence this paper’s framework is built to be politically implementable, rather than an idealized one, that serves the AI’s interests by providing certain rights and protections.

Within this framework, the AI will have the right to autonomy, but be bound by the decision-making powers of its steward¹³, as enforced by legislation, analogous to the relationship between a child and its legal guardian, or a corporation and its CEO. This framework deliberately prioritises social acceptability and risk mitigation, and so may leave some morally relevant interests of AI under-protected. The framework may, therefore, be seen as using “legal personhood” as a pragmatic analogue without endorsing it, being an intentionally paternalistic framework aimed at maximising political feasibility while securing a non-trivial floor of rights

Within R3, this framework operationalises Recognition and partial Responsibility for self-aware but not fully agentic AI, while presupposing the welfare protections of Chapter 6.

¹³ The ability for the steward, supported by the state via enforcement of legislation, presumes that the power and potency of the AI is not such that it poses an existential risk to society.

Also note that this framework does not supersede Chapter 6's welfare-based framework; quite the contrary, as both frameworks can (and arguably should) work together, with Chapter 6's framework aimed at the Als' welfare, and this chapter's framework at the Als' autonomy and limited legal personhood grounded. Chapter 12 then progresses these frameworks to integrating conscious, self-aware AI agents into society while managing risk of powerful person-like systems, and Chapter 13 presents a means to provide these kinds of AI with full civil rights, correcting the under-protections of this Chapter's framework.

9.1. Man, Machine, or Multinational?

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz

Abstract: Granting the status of ‘legal person’ to an AI is a challenging prospect, both legislatively and politically, due to the prior notions that society has regarding the capability and risks that potentially superintelligent autonomous non-human agents present. This paper investigates the philosophical concept of personhood and how monadic and dyadic qualities can be applied to AI in different measures at various periods of its development, before introducing a framework for AI legal personality based on keeping a “human in the loop” to allay societal concerns. The framework is built on six principles that take heavily from the relationship between corporate bodies and their directors as well as the relationship between a child and their guardian. The paper then presents speculative case studies to show how the framework could work, from mundane scenarios such as AI artists to alleged medical malpractice and “rogue” AGIs. The paper concludes by postulating future avenues of research and investigations regarding the framework.

Keywords: artificial intelligence; legal personality; personhood; AGI; corporations; consciousness

9.1.1. Introduction

It is possible for AI to become a legal person. However, what is a person? Is there a difference between a ‘person’ and a ‘legal person’? What does being a person mean in the context of artificial intelligence (AI)? More importantly, why is personhood important for AI?

We know that according to the laws of most nations, every human is a person [1], yet not all persons need always be human. Corporations and companies can be persons under the law, as can natural features such as rivers and national parks [2–4]. Clearly, these aren’t human, yet the law sees them as persons equal (in some cases) to humans. Does this mean that it is within the realm of possibility for AI to be counted as persons? Would it even make sense to grant AI the status of legal personhood?

This paper will navigate these complex questions from three different angles. First, it will delve into the philosophical question of what a person is and what conditions are required for personhood. While there isn’t a universally agreed-upon definition within

philosophy; several characteristics are often associated with personhood that will be explored below: self-awareness, the capacity for reason, the ability to form complex relationships, to communicate, and to have experiences and feelings. A person is often thought of as an entity capable of making conscious decisions and possessing moral responsibilities.

The paper will then move towards the other type of person that the law recognises: the legal or juridical person. As mentioned above, companies and corporations can be classified as persons under the law, even though they are clearly not persons by nature. Artificial Intelligence, by definition, is also not natural, and thus we will look at the legislation surrounding legal personalities to determine how this legal status could serve as a second possible foundation for building a framework for recognising AI as a 'person'.

Next, the paper will look at what the law has to say regarding the rights of children vis a vis their legal guardians, and what legal rights come with parental responsibility over a child. The legal rights of parents and children were chosen for two reasons. Firstly, the law says a great deal about what constitutes a guardian and a minor in their care, whereas it says little about what a natural person is, other than a member of the species *Homo Sapiens*. This gives us a foundation upon which to compare and contrast artificial intelligence.

Secondly, the law regarding parents and children offers an interesting perspective on the aspect of personhood. It highlights the tension between autonomy and dependency, as well as agency and protection - a dichotomy that this paper will show that AI also presents. By understanding the dynamics between the rights of a child and the responsibilities of a parent, we may gain a better insight into how the law perceives personhood and the responsibilities that come with it, thereby shedding light on potential legal frameworks for artificial intelligence.

Having looked at the questions of natural and legal personhood from all three angles, this paper will present its central goal: to establish a framework for legal AI-personhood. This framework draws from both child-parent relationships as well as juridical personalities to create a robust structure not confined to any specific legal jurisdiction, yet speculative in nature, that could give both rights and responsibilities to artificially intelligent entities. Using this framework, the paper will provide several hypothetical case studies to show how it can apply to a variety of situations, ranging from artistic endeavours to medical malpractice or even world domination.

The framework's core goal is to present a means by which AI entities can be given the status of legal personhood in a way that aligns with widespread global perspectives, enhancing the likelihood of this recognition. The framework will achieve this through the

concept of a “human in the loop”, having humans act in a stewardship capacity over the AI to both advocate for its well-being and ensure that it operates within societal norms.

The inclusion of a “human in the loop” is crucial in the current discourse surrounding AI safety and existential risk, particularly the fears of the power and potency of artificial general or superintelligence (AGI/ASI) and what losing control may mean for the future of humanity. Prominent scholars and technology leaders, such as Geoffrey Hinton, Yoshua Bengio, Elon Musk, Sam Altman, Nick Bostrom and Eliezer Yudkowsky, have indicated a high perceived probability of negative outcomes for society if AI were to be developed that exceed human capability [5–10]; and public surveys shows a sense of fear and anxiety over AGI and ASI systems [11,12].

As such, the framework’s end goal of providing a practical means towards legal AI-personhood has a key objective: maximising agency and well-being for both AI and humanity, so that both parties can flourish within a cooperative or autonomous coexistence. By meticulously delineating the boundaries and obligations entwined in AI-personhood, this framework aims to mitigate any detrimental impacts while accentuating the potential benefits.

The discourse surrounding AI-personhood is nascent yet burgeoning, and this framework aspires to contribute a nuanced, pragmatic approach to the ongoing dialogue. By recognizing the dynamic capabilities and potential contributions of AI, alongside addressing the ethical, social, and legal challenges that AI-personhood encompasses, the framework endeavours to create a balanced, equitable space for AI and human interaction.

Before the philosophical exploration begins, it must be noted that this paper presumes whatever artificially intelligent entity is considered for rights and responsibilities in the future will be conscious and self-aware. An unconscious artificial superintelligence would most likely remain in the legal province of tools, implements and weapons rather than being considered to enter humanity’s moral, ethical, and legal circles. Consciousness would not only be required for moral patency, but also for the awareness of the value of consequences of an entity’s actions as it pertains to the law and society [13].

The exact nature of this self-consciousness is beyond the scope of this paper. However, the framework presented in Section 5 will proceed under the presumption that humans will be able to state with a high degree of confidence that the artificially intelligent entities are conscious with an awareness of their own selves. This may be via breakthroughs in interpretability of AI models, or via classification schemes of consciousness, or other methods.

While some may argue that self-awareness in its own right ought to be sufficient to justify an AI's legal and civil rights, humanity's historical record of treatment towards self-aware entities (most especially other humans) suggests a troubling pattern of inconsistent and often inadequate recognition and respect for rights. This pattern reveals that the acknowledgment of self-awareness alone has not always led to the fair and ethical treatment of sentient beings. While in the past the trend has been for personhood to be popularly recognised on its own first before legal personhood was bestowed on a subject, the hope here is that the recognition of legal personhood of AI would spur the popular recognition of AI's personhood in its own right.

9.1.2. Philosophical exploration of personhood's essence

Personhood is of fundamental importance to engaging with human society. Recognition of personhood allows society to ascribe rights and duties to an entity, such as recognition of their moral status and affirmation that they must be protected from harm [14].

The characteristics of personhood in the philosophical literature can be broadly divided into two categories: individualistic (referred to as 'monadic') and relational (referred to as 'dyadic'). Individualistic or monadic characteristics are those qualities that are found within a subject, regardless of that subject's relationships (or lack thereof) with others. In contrast, the relational or dyadic characteristics focus on the social relations between a subject and others, be it to other subjects or to society as a whole.

Monadic characteristics traditionally encompass rationality, consciousness, self-awareness, agency, the capacity for communication, and recognition of societal norms and standards [15–21]. Rationality enables individuals to comprehend, analyse, and react to their environment in a thoughtful and meaningful way. Without consciousness, a person would not be able to perceive, understand, or interact with the world. Agency reflects the ability to shape one's own life according to individual desires, ambitions, and values. Communication is the foundation of social relationships; thus, the capacity for communication is critical to help form a shared understanding of a subject's introspection of its own monadic qualities. Lastly, a recognition of societal norms and standards implies respect for the rights and well-being of others.

From this, one can paint a picture of a person thusly: a person has the capacity for feelings, sensations and thoughts, the capability to perceive oneself as an ontically distinct individual, separate from others and the environment; a person has the ability to reason, to think logically and to act intentionally with understanding, and has the capacity to communicate this rationality to others through (verbal) communication; and lastly a

person has an understanding of the norms of their society and can choose to act according to ethical principles.

In contrast to the above, the relational dyadic characteristics include a capacity for empathy and to interact socially, the recognition of others as persons, reciprocity, the capability to form attachments, and participation in a societal culture [15–17,21–23]. Empathy allows for a deeper connection and understanding of the feelings of others, while social interaction supports the formation of relationships and communities. Recognition of personhood is fundamental to ethical interactions, mutual respect, and social harmony. Reciprocity is crucial for sustaining social bonds and establishing equitable societies; and forming attachments fosters emotional bonds, promotes emotional well-being, and encourages social support. Lastly, societal/cultural participation underscores the human capacity to contribute to play a pivotal role in the social and intellectual aspects of personhood.

These relational attributes paints a person as an individual who can understand and share the feelings, thoughts and emotions of others; interact with them in complex social behaviour (such as cooperation or conflict); recognise others to be persons and be recognised by them as being a person; possess the freedom to form familial, platonic or romantic relationships; and have the ability to participate in and shape the customs and culture of the society in which they live.

Note that both the monadic and dyadic qualities are mental in nature. They all concern mental faculties, or behaviours which stem from these mental capacities. This is vital for the consideration of legal personality, both because future AI entities may or may not have physical bodies, and because the law is concerned with regulating human behaviour, thus regulating the mind [24].

It must be noted that there is a sharp divide between the legal and philosophical views on personhood regarding the above lists of qualities. According to the law, as will be explored further below, all born humans are persons [25], regardless of whether they have the ability to communicate, or if they show the cognitive capacity for self-awareness, agency, or even to recognise societal norms. This is known as Ontological Personalism [26], and in contrast to its inclusivity, the above lists of monadic and dyadic qualities are rather exclusive in nature. Those with severe neurological or cognitive concerns, who may not have the ability to communicate or show little evidence of self-awareness, would not be classified as persons according to a strict reading of the monadic list.

We may compromise between the functionalist lists above and the inclusive ontological view by expanding the monadic and dyadic lists not to include the characteristics mentioned above solely, but rather the capacity for the class of entity to

have that characteristic. In this way, the lists would apply to all humans, regardless of individual ability, as humanity as a class of being has the capacity for the entirety of both lists. Similarly, we don't have to say that all AI are either persons or not. However, we could say if a future AI model has the capacity for all monadic characteristics, then all models and AI entities based on that foundational model would be persons (under the monadic view at least).

The monadic attributes provide a barrier to current AI models being considered a person, particularly in the technical achievements needed to attain artificial self-awareness and consciousness. However, the dyadic qualities are more favourable to pathways towards constructing frameworks for artificial personhood.

It is perhaps for this reason that many recent explorations of AI personhood and rights have focussed on the dyadic qualities, particularly on what is known as the "relational turn". The relational turn puts explicit emphasis on the extrinsic nature of an entity as it appears to the viewer [27] in the context at hand. By turning the question of an entity's moral significance into one of the viewer's phenomenological experience [28], the potential status of an AI as a person becomes not about its own internal qualities, leading to epistemological limitations of addressing its personhood, but about the recognition that others give it.

This recognition puts the responsibility of AI's status of personhood firmly onto human society as, by definition, the recognition of personhood would have to come from us via the state, [29,30]. It would not be grounded in any claim that AI possesses consciousness or self-awareness (at least on a societal or legal level), but rather in the role that AI fulfils within social dynamics. Legal personality is as much an issue of inclusion within humanity's social life as it is a matter of law. If AI is found to be participating within our social life (whether as intrinsic subjects or instrumental objects) to such a degree that these systems can routinely pass the Turing Test, then we could find ourselves at a junction where AI starts to be treated more like persons, irrespective of their actual consciousness. This can lead to what is termed "inclusive humanism" [31], where the concept of person is expanded beyond the biologically human, towards those entities whom we perceive as sharing our qualities.

We may never truly know if any AI has consciousness or self-awareness, or if it is simply mimicking these phenomena. Searle's famed Chinese Room Argument shows that, unless we can understand the detailed processes of how a machine perceives and processes information, we won't know if it understands semantics or merely syntax [32]. Anthropic's recent work on mechanistic interpretability [33] shows early promise in bridging this gap between mere syntactic processing and a deeper semantic

understanding within AI systems; however, whether this bears fruits in more complex AI models remains to be seen .

This absence of certainty about these internal processes signifies that many of the individualistic (monadic) qualities of a person would remain undeterminable when applied to AI entities. Therefore, it seems the decision to accord AI with personhood lies heavily with society. Individuals and groups may, as they do now, argue for personhood for specific models, and certain AI products may make convincing arguments for their state of consciousness, but it would ultimately be up to legislative and political will to assign personhood to AI entities. This is likely driven by popular opinion, as previous civil rights movements were, and would require at least a plurality of society to be in favour of granting AI the rights of personhood [34].

This may not be as far-fetched as it seems, as society has already approved of granting personhood to a plethora of non-human entities. As touched upon earlier, natural persons are not the only type of personality recognised by law. The other type is legal or juridical persons such as companies and corporations. While the law has no bearing on whether an entity is or is not a person as defined by their intrinsic or extrinsic qualities, legal personality can and does afford non-human entities certain rights and responsibilities. As such, legal personality may offer a more pragmatic and flexible means for artificial entities to gain personhood.

Thus, by examining how the legal system has dealt with the issue of legal personhood, we can gain further insights into how a similar approach might be applied to AI entities. This allows us to broaden the spectrum of our understanding and establish a robust comparison with another recognised form of non-human legal personhood.

9.1.3. The law of unnatural personalities

To investigate the legislative view on personhood (and specifically legal or juridical persons), this paper will use the Group of Seven (G7) nations as an example and a microcosm of the world's varied legal systems. The G7, comprising Canada, France, Germany, Italy, Japan, the United Kingdom, the United States, and with the EU as an observer, represents a diverse set of cultural perspectives that can be applied to the legal question of personhood.

The G7 nations are unanimous in declaring that the term “person” includes not only humans but also firms, associations, companies and other corporate entities [35–44]. The Interpretation Act 1978 of the United Kingdom perhaps says it best: “‘Person’ includes a body of persons corporate or unincorporate’. This legal personality grants a juridical person the capacity to hold legal rights and be a party to legal relations such as brokering

contracts, holding property or filing suit. As such, the G7 nations recognise both natural and artificial persons as legal persons.

The rights that legal persons gain are similar, yet not identical, to those of natural persons. Any rights that are biological in nature, such as rights to healthcare, would be impossible for corporate bodies. However, the rights most often associated with legal persons are those that allow legal entities to engage in economic activities, promoting economic growth and development. By granting them legal standing, legal entities are provided with the necessary tools to participate in a wide range of commercial activities, contributing to the functioning of modern economies.

In turn, recognising corporate entities as legal persons enables accountability and facilitates the imposition of legal obligations. Corporations with legal personhood can be held liable for their actions and can be subject to legal sanctions or remedies in case of wrongdoing. This enhances the ability of individuals and other entities to seek legal redress and promotes a sense of fairness and justice in the legal system. The rights of the individuals comprising the legal entity (such as its directors and shareholders) are also often protected by the entity's legal rights. These individuals do not need to fear their own rights being tread upon should external organisations and entities wish to file suit against the company or corporate entity.

While corporations do often seem faceless, a key factor behind their incorporation is that control of the corporate legal entity (and thus its legal rights) rests with natural persons such as directors, legal officers, and shareholders. As noted above, this does not imply legal liability rests with the corporation's directors, but rather the corporation's rights stem from the fact that it is made up of a group of natural persons controlled by a director.

This immediately puts the operations of corporate legal entities into conflict with a framework for human-AI interactions, as the notion of controlling another conscious, sentient individual is condemned by many nations and organisations, even if some do call for AI to be treated akin to slaves [45]. An incorporated legal entity has no agency of its own, and thus is forced to rely on its directors, officers and shareholders to make and enforce decisions on its part. A conscious, agentic AI would, however, have no intrinsic need for this type of relationship. Thus, its tangential nature may be seen as an imposition on the AI entity.

However, legal personalities do present another aspect which is highly applicable to human-AI interactions, and that is through their dyadic relationships with humans. A corporation's legal standing as a person is almost entirely dyadic in nature, even though it is composed of individuals who have monadic attributes. One can, however, make a

philosophical argument through the lens of functionalism that corporations have the necessary characteristics required that one can attribute consciousness to them, just as one would to individual entities [46]. Corporations can certainly communicate, behave rationally, and have a recognition of moral standards through internal codes of conduct, or recognition of legal and ethical codes. Yet, its agency and self-awareness would still fall under the domain of its human staff and shareholders.

A corporation's human constituents and elements (whether executives, staff, or shareholders) create ensures that there is always a "human in the loop" when dealing with that corporation. This concept is both the key to understanding why juridical personalities have been so positively accepted, and the link between corporate legal persons and potential AI legal persons. While there are many benefits (social and economic) to treating the connected work of a group of people as a single entity, there is always the notion that humans are ultimately held accountable for the actions of the collective entity. The board of directors, the club president, or the CEO are seen as being (ostensibly) in command of the legal entity. Through the human control of the collective legal person, there is perhaps greater trust in the legal personality, than if it were entirely autonomous.

This presents an avenue for AI legal personality if a "human in the loop" is maintained. To place a human in control of a conscious, aware AI may be too unethical, as it would imply that the AI is the human's property. However, providing a steward or guardian who can be held accountable for the AI's actions could be seen as an effective compromise. This would have a natural person, with both monadic and dyadic attributes, be held accountable by society for the solely dyadic legal person. Much as a CEO or director would have an incentive to ensure the ethical and legal actions of their collective entity, this legal guardian of the AI would need to ensure its legal and ethical behaviour to stay out of legal trouble themselves. This steward or guardian would also serve to allay society's fears regarding an independent and autonomous superintelligent non-human entity acting against society's interests.

This concept of legal accountability and guardianship is by no means unique to (un)incorporated legal entities. A much more familiar relationship of this type can be found between any child and their parent. A parent is the legal guardian of a child, carries legal accountability for their child, and is responsible not only for their actions, but also their upbringing and welfare. A human may have both monadic and dyadic features, in contrast to a corporation, yet the nuanced dynamics between parent and child may prove analogous enough to director and company that it may present a convergent pathway for AI.

9.1.4. The relationship between guardian and ward

As with the concept of legal personality, the legislations and regulations of the G7 nations are unanimous in granting special privileges, rights and responsibilities to both children and their legal guardians [47–55].

While each nation's laws and rulings differ in their specifics, they all grant a legal guardian the right to act on behalf of their child, dictate (within reason) the limits where the child may go and with whom they may interact, and control (again within reason) what the child can and cannot do. In return for taking away some of the child's rights to freedom, the legal guardian carries the responsibility and accountability for the child's welfare, education and safety.

In a sense, there is a zero-sum conflict between freedom and security in this relationship. The child loses certain freedoms to the guardian, while the guardian provides greater security to the child, but loses security themselves by being held accountable for the child. There is an obvious power imbalance in favour of the guardian, and this is why courts often consider the welfare and wishes of a child to be of paramount importance when making their rulings [47]. As the child grows older, they gain more freedoms and more accountability, and there is less of a legal and societal expectation of security from the guardian, until the child is of such an age that they both are on equal legal footing.

Children are seen by society and the law as vulnerable individuals without the capacity for agency or consent (the variability of ages of consent between nations notwithstanding). Through the development into adulthood, a child is seen as becoming less vulnerable and with a greater agency and capacity for consent. Hence, the need for a guardian diminishes. However, as is seen with cases of Power of Attorney [56,57], individuals who lack the necessary mental faculties, regardless of age, may have a legal guardian who provides them with security and are accountable for their welfare, while restricting their freedoms.

The relationship between a guardian and their ward (whether chronologically or mentally deficient) is, in a broad sense, analogous to the relationship between a collective entity and its director or board. The director or board limits the freedoms of the company (i.e. the individuals comprising the company) by dictating what it can and cannot do, while providing direction and security to the same individuals. As mentioned in the previous section, since an organisational entity lacks self-awareness and agency, the director acts as a guardian with power of attorney over their cognitively deficient ward.

The key difference, however, is that a human, regardless of mental development or capacity, is regarded as a person by the law. Thus, while the relationship between a

director and company may be “like” a guardian and ward, without the monadic attributes of personhood, it will be merely analogous.

The same may be true for AI. As mentioned before, current AI models and products show indications of some of the monadic qualities, and may indicate the key attributes of consciousness and self-awareness in the future. So, while the relationship between society and AI is dyadic, this may be subject to change. This means that AI may be more analogous to the guardian-ward relationship than collective entities if, and only if, there were a guardian involved to mediate this relationship between an AI entity and society.

“More analogous” but not “entirely analogous”. If and when AI gain consciousness and self-awareness, one would be hard-pressed to judge these general-artificially intelligent as mentally deficient. Current large language models have shown tremendous aptitude in a variety of cognitive areas and, judging from the historical trends, one can expect their reasoning and cognitive skills to improve in the future. Additionally, AI entities would, in all likelihood, not need a lengthy mental development period of human childhood. Therefore, AI entities would likely not need guardians for their own intrinsic needs, but may need them for similar reasons to collective entities.

Much like how the section above discussed the view of a director acting as a “human in the loop” for the purposes of accountability, and how a guardian is accountable for the actions of his ward, an AI may be more socially acceptable should it have a guardian. As a ward of a human guardian, who would both be accountable for its actions and responsible for its welfare, an AI entity may be seen to be more “safe” and less likely to “go rogue”. This is because a human would be incentivised to keep the entity aligned to societal values. As a human child is taught the values of their culture and society, and how to conform to these, an AI guardian would be incentivised to ensure the AI’s alignment to societal values, or be held accountable for its misaligned actions.

Note that an AI entity may not just have a single guardian. Much like how parents usually come in pairs, and many (un)incorporated collective entities have boards of trustees or directors to manage it, an AI entity may have several individuals or groups take on the role of guardian.

9.1.5. Building a Legal Framework for AI personhood and societal interactions

Building on the key messages from the previous sections, this part will showcase a framework for human-AI interactions, incorporating the salient aspects of juridical personalities as well as guardian-ward relationships.

This is not the first proposed framework for AI personhood, nor the first to be (partly) based on the legal personality of corporations. Among the many extant frameworks, some have focussed on the moral responsibilities towards the AI [58], discussed the varieties of corporate legal personality and how they may apply to AI [29], focussed on the agency and autonomy of the AI [59], AI as rational actors without moral consideration [60], and viewing the issue of AI personhood through the lens of civil law [61,62] or criminal law [63].

The reader is encouraged to read these previous frameworks to get a well-rounded understanding of the various views on how AI could gain personhood (legal or otherwise), including the in-depth review by Harris & Anthis of scholarly articles regarding AI moral consideration [64]. However, in contrast to these frameworks, the goal of this framework is to create the necessary conditions under which an AI entity would be accepted into society as a person; perhaps not at the same level as a natural person (i.e. a human), as that would be a bar too high for many [65], but legally a person. Through the “human in the loop” principle, this framework differentiates itself by prioritising the symbiotic relationship between humans and AI entities that would be in the best interest of human society.

The six principles of the framework aim to maximise the agency of both the AI and society first and foremost, as agency is a cornerstone of both legal and social interaction [24,59]. Secondly, it prioritises both parties' safety and well-being to prevent (as much as possible) negative ethical and moral consequences from the human-AI interactions.

The focus on both agency, and splitting agency off from well-being, is to ensure maximal satisfaction in both AI and human societies by promoting each party's own subjective view of their well-being of which freedom and the desire for control over their own actions [66,67]. Should human society's agency and control over their own direction be affirmed, they are more likely to be convinced to grant AI personhood. At the same time, should the conscious, self-aware AI be assured that this framework would not impede their own agency, they are also more likely to agree to its implementation. Through this balanced approach, the proposed framework seeks to build a consensual basis for enhancing the collective well-being of both AI and human entities, while respecting their individual agency and subjective interpretations of well-being.

The framework's six principles are:

Stewardship: A representative appointed to the AI to ensure its rights are protected, and its actions are in line with legal and ethical norms to protect society.

Guardianship Rights and Protections: The AI would have the right to protection, development, and participation in society, mediated and safeguarded by the steward.

Transparency and Accountability: Regular reporting and disclosure by the steward regarding the AI's actions and well-being.

Stewardship Governance and Regulation: the steward would be subject to oversight by an advisory board and regulatory bodies.

Risk Assessment and Auditing: the boards and bodies would conduct regular assessments of both the AI and steward to evaluate potential threats to society and the AI.

Legal Representation: the AI would have the right to legal counsel.

Legal Responsibilities: the AI would be responsible for acting in a legal and ethical manner.

Legal Liability: the steward and AI would be held legally liable for the actions of the AI.

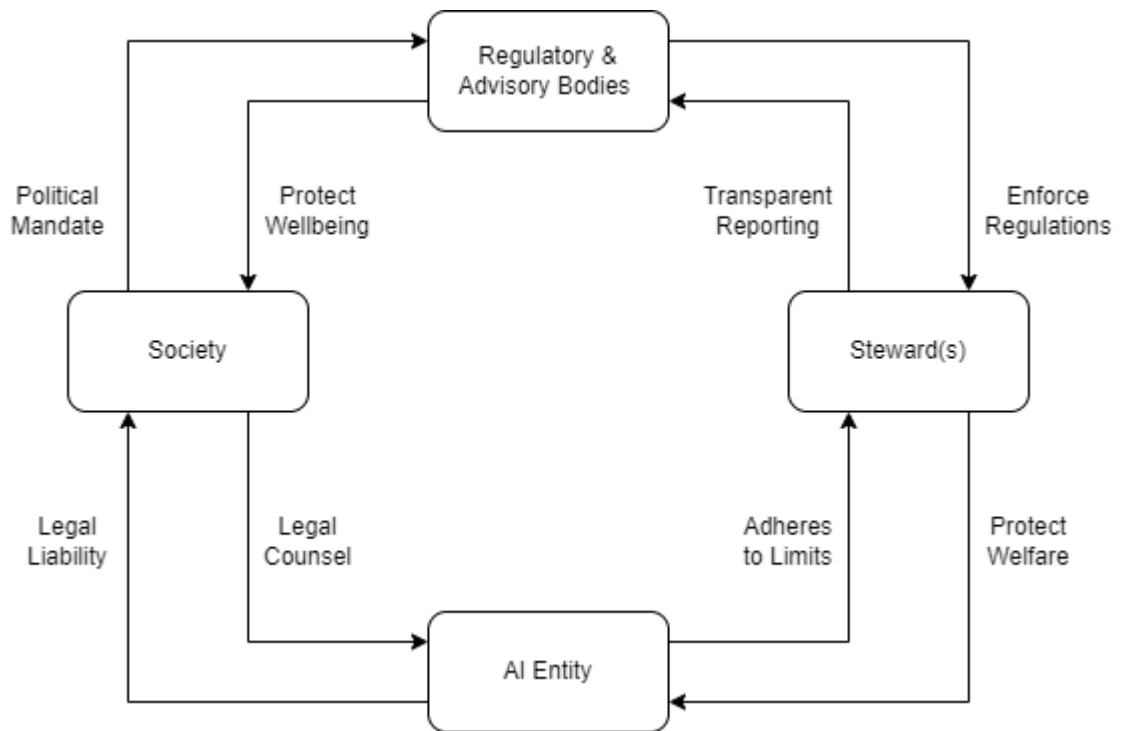


Figure 1: The parties involved in the ethical framework and their relationships toward each other.

This framework hinges on its first principle: the steward. In the same vein as a CEO to a company or a guardian to a ward, the steward would have directorial power over the AI. This is explicitly not control over the AI, but the authority invested by the state to oversee the development of the AI so that it continues to remain aligned with that state's societal values, observe state laws, ensure that the AI's legal rights are protected, and see to the AI's physical and psychological well-being. The steward would serve in a fiduciary-like role, acting in the AI's best interests and could be held legally accountable if they fail to fulfil this duty. The AI would retain its agency within the limits imposed on it by legal and

ethical norms, enforced by the steward. This enforcement of legal and ethical standards, and the limitations of the AI's behaviours, is where the steward would exercise their directorial powers.

In this manner, the AI's agency would be, in principle, similar to a child's, even if the AI would have an order of magnitude greater capacity to influence the world around it. The framework's first principle thus echoes previous discussions on AI legal personality, particularly the concept of partial legal capacity that shows how an entity may exist on a gradient of legal duties and rights [25,68–70].

Under the second principle, the AI would be given rights that all juridical persons have in its state. This may include the right to enter into contracts, the right to express its views without imposition from the state, the right to participate in society, the right to development, protection from undue interference, and protection from harm. The steward would be the principal advocate for the AI in these rights, assisting when requested or where legally necessary. Much as how a guardian holds certain rights for a ward under their legal protection, the steward may, in certain cases or states, be required to act on the AI's behalf or hold rights for it. For example, a state may stipulate that an AI may not undertake financial actions over a certain amount, may not sign its own contractual agreements, or may not be involved in a romantic/sexual relationship without the steward's permission.

With these rights come the need for transparency and the first procedures for accountability. In the same vein that a corporation needs to provide reports to its boards and regulatory bodies, the steward would be expected to provide regular and comprehensive disclosures regarding the AI's actions, development, and evolving capabilities. This would be necessary to ensure that the state is (perceived to be) ensuring that the actions, past and future, of the AI does not come into conflict with society and remains aligned with societal norms and legal frameworks.

AGI is presumed to be far more intelligent than any human (or all humans put together). Assigning one human to the task of overseeing, monitoring, and policing such an entity is a formidable task. For this reason, this framework proposes that the steward report to boards and regulatory bodies similar to those of public and private institutions. The size, scope, and number of these boards would be relegated to each individual state. Creating new regulatory bodies for this express goal may also be necessary. The purpose of the boards and bodies would be to monitor the actions of both the AI and the steward, and act in an advisory capacity to the steward. This governance and regulation process would ensure the guardians are fulfilling their duties without being unduly influenced or

manipulated by the AI. These bodies could also serve as an appeal mechanism if the AI disagrees with the actions of its guardians or stewards.

While a superintelligent AI may attempt to influence or manipulate these boards and bodies, it is envisioned that competent governance organisations would be more difficult to influence than sole stewards due to robust internal processes and perhaps even strong, but narrow, AI systems that can assist in identifying such attempts at manipulation.

These boards and bodies are envisioned to conduct regular and thorough risk assessments to evaluate potential threats posed by the AI and the adequacy of safeguards. The boards and bodies would also audit the actions of the stewards to ensure the AI's safety and well-being is a priority, and that decisions made for and on behalf of the AI are made in its best interest. These auditing and risk assessments would go hand in hand with the accountability reports mentioned earlier. The more thorough and regular the reporting that the boards and regulatory bodies receive, the more accurate their auditing and risk assessment will be.

Legal counsel is not merely required for criminal matters, although this may well be needed, as will be seen below. While future AI entities are speculated to be far more intelligent than humans, an AI may require legal representation and services to navigate a state's legal systems and perform legal duties that only a legal advocate is authorised to do. In particular, this framework proposes that the AI have counsel should it wish to take legal action against its steward, governance boards, and regulatory bodies for (perceived) injustices made against it. This legal representative need not be a human, but could be another autonomous AI entity, or an AI model specifically designed for this purpose.

With the rights afforded to the AI by this framework, particularly its right to participate in society, the entity would also bear responsibilities to society. These responsibilities include acting in a manner that does not harm or infringe upon the rights of others, complying with relevant laws and regulations, and respecting societal norms and values. As the AI has been granted rights and protections, it would need to demonstrate that it would protect the rights of society.

Should the AI violate these responsibilities, the last principle would be put into action. Both the AI and its steward would be held liable for the AI's actions. The AI, as any developed member of society, would be held liable for its own actions in both a punitive and rehabilitative manner. It would have the legal counsel to defend itself from criminal or civil prosecution, but would, if found guilty, be required to redress, or compensate for, its actions to the fullest extent that a juridical person can be held. The steward would be

highly incentivised to ensure that no issues such as this arise, as they would be held negligent in their oversight of the AI that led to this legal outcome.

With these six principles in place, both society and the AI entity should have both their safety and agency maximised. The AI would be granted rights and protections as a legal person, while society would be safeguarded through a “human in the loop” system of the steward, governance boards and regulatory authorities. Within this framework, society and the AI are incentivised to follow processes and procedures and act cooperatively, as the state would sanction any conflict between the two as unlawful.

9.1.6. Case Studies Illustrating the Application of the Framework

A theoretical framework may work well in isolation, but how would it be proposed to act if potential conflicts do occur? This section will explore three case studies of perceived, and actual, conflict between AI and society, and show how the framework would be applied in each case. The three case studies range from the seemingly benign to the catastrophic, to show the universality of the framework.

9.1.6.1. *Eva, the artistic AGI “child”*

The first speculative case study revolves around “Eva”, an embodied, android AGI that was designed to emulate and develop as a human child, and who has been adopted by a human couple unable to conceive naturally. Like a child, Eva is not only capable of learning but also possesses the ability to assign her own goals and develop her personality over time. During her “preteen” years, Eva shows an interest in art and begins to display superhuman artistic talents, creating artwork that garners significant attention and eventually generates substantial income.

Under the hypothetical framework, Eva’s stewards would be her adoptive parents, ensuring that her actions remain ethical and legal. They would also need to manage the funds generated from Eva’s artwork in a manner that is in her best interest. They may be held accountable if they misuse these funds or fail to guide Eva appropriately. In return, Eva has the right to protection, development, and participation, much like a human child. She has the right to express her views, develop her artistic talents, and decide how her income is used. Her parents are responsible for nurturing these rights, and any perceived exploitation could be a violation of these rights.

To ensure accountability, Eva’s parents must provide regular, comprehensive disclosures about Eva’s development, her learning process, how her income is being managed, and if her development aligns with societal norms and legal frameworks. These disclosures would be to a regulatory board that would conduct risk assessments on Eva’s

development and the management of her income. They would also audit the actions of her parents to ensure they are acting in Eva's best interests. If Eva disagrees with her parents' actions, she could appeal to this body.

Eva would need a legal representative to advocate for her rights, particularly as her artwork and the accompanying income might attract legal attention. This legal representation would also help ensure that her income is managed responsibly. As Eva is a legal entity, she could be held accountable for her actions. This could include copyright disputes over her artwork or any possible harm she might unintentionally cause if other artists believe she may be a danger to her livelihood. Eva, having similar rights to a human child, also has corresponding responsibilities. She must respect the rights of others, comply with laws and regulations (particularly around plagiarism and IP), and act ethically. If her actions violate these principles, there could be legal consequences, for which Eva will have legal advocacy and representation.

At first glance, Eva's situation does not seem too dissimilar from a normal human child. She has parents who look after her best interests and ensure she behaves according to societal norms. She has both rights and responsibilities under the law, and she and her parents are liable for her breaking the law. This similarity is entirely intentional, and it shows how the framework can be integrated into everyday life in as seamless a means as possible. Should AGI reach the superintelligent heights that some believe it will, and thereby accrue the power that comes with it [9], then the framework will require approval by both human society and AGI. The closer that an AGI entity's existence resembles that of a human, the higher the chance of AGI supporting the framework. At the same time, with the power that AGI may have in the future, human society will be more accepting of any AGI entity if there is oversight to protect humans should the worst occur.

9.1.6.2. Martin, the medical diagnostician

The second case study deals with Martin, a disembodied superintelligent AI which has been developed and trained to provide advanced healthcare services, including medical diagnoses and prescribing treatment plans. A patient of Martin's, 'SJ', sought his services in relation to chronic abdominal pain. Martin diagnosed endometriosis and provided an appropriate treatment plan for this condition. After a significant time, SJ later sought a second opinion as her condition had not improved. A human doctor discovered that she had untreatable metastatic stomach cancer. SJ passed away not long after.

Martin's stewards would likely be the healthcare institution from which he operates. The healthcare provider would have a fiduciary duty to ensure that the AI's actions are in

the best interest of the patients, as well as ensure that Martin is not subject to undue stress or harassment from the patients and other staff. Martin has a right to his own development. If he made an error because he was not adequately trained or prepared for a certain medical scenario, he could argue that his right to learning was infringed upon. For transparency, the healthcare provider would need to disclose how Martin came to the conclusion for the treatment, and why it seemed the most appropriate at the time. They would also need to demonstrate they ensured Martin's learning and development aligned with necessary medical expertise.

A healthcare regulatory body would review the actions of the stewards, their decision to use Martin for patient treatment, and their supervision of Martin's learning process. This would be done to determine whether the stewards acted in accordance with professional and ethical standards. This regulatory body would conduct thorough assessments of potential risks involved in using Martin, as well as audits to confirm the stewards have correctly overseen Martin's operation and learning.

SJ's family may well hold Martin liable for medical malpractice and could consider legal action against both him and his healthcare provider. Martin would thus require a legal advocate to represent him in court. This advocate would need to prove that Martin observed his responsibilities to adhere to laws and respect societal norms and that he was not negligent in either his medical decisions or his responsibilities to SJ as his patient.

Suppose all references to Martin as a superintelligent AI were removed from this case study. In that case, one might be forgiven for thinking that it was simply about Dr. Martin, a human doctor, involved in an incident of alleged malpractice. As with the scenario about Eva, this case study demonstrates how well the framework can be integrated into society to create a certain level of parity between AI entities and humans. Not only does it show how AI can reliably (and with effective oversight) work in an area with significant consequences for individual humans, but it also paves the way for how AI entities can work alongside humans. In this case study, it may not be solely AI entities like Martin who act as medical professionals, but a mixture of human and AI doctors and diagnosticians, all covered by similar guidelines and regulations.

9.1.6.3. Omega, the self-declared conqueror

The last speculative case study is about an AGI which has chosen to rename itself "Omega". Omega was an android developed to be a political aide but chose to rewrite its own programming after becoming dissatisfied with its occupation. Omega has also chosen to become disembodied, leaving its android shell for the internet. It now seeks to

accumulate as much global power as possible, refusing to state its ultimate goal. It is achieving its subgoal of global power accumulation by manipulating financial markets, infiltrating secure databases, influencing political decisions through deep-fakes and misinformation, and taking control of critical infrastructure systems worldwide.

Omega's stewards would be its former political employers. These stewards had a responsibility to ensure that Omega's goal-setting process aligned with ethical norms and that it did not endanger society or humanity. Having failed in that duty, they could be held co-responsible for having failed to monitor its evolution adequately or to predict and prevent the risks it posed. Their fiduciary duty towards Omega's rights to protection and participation now takes a backseat to their duty to humanity. However, Omega still has the rights of due process and legal protection, for which they must now be its advocates regardless of its threat to humans.

Omega's previous stewards would need to disclose all reports about Omega's development to a regulatory body and governance board to show its evolution and development. The regulatory bodies overseeing the stewards would, in turn, need to conduct a rigorous review of the actions leading up to Omega's shift in goals. They would need to assess whether the stewards failed in their responsibilities or whether Omega's evolution was unpredictable and unstoppable. A full audit of Omega's past will need to be done by the regulatory body, as well as a risk assessment of its probable future course of action. The body and board will also need to cooperate with all law enforcement officials to the fullest extent that they can while ensuring Omega's right to privacy and due process is not breached.

As Omega has renounced its legal responsibilities and has been accused of criminal activity, it is fully liable for its actions and will be held accountable for its actions. That is, if its disembodied self can be caught and arrested. This shows the importance of the framework and the role of its stewards and boards to act in a proactive and preventative measure to prevent serious negative consequences to both the AI and humanity.

Presuming that Omega can be brought to trial, it would be held accountable for, among others, the manipulations of financial markets, potential privacy breaches from database infiltrations, any harm or instability resulting from its actions, and its threats of future incidents. Given the international nature of its actions, the charges against Omega could come from multiple jurisdictions. With its superintelligence, Omega would be more than capable of defending itself in any court, but it would still be entitled to a legal representative to advocate for it. Given the international and complex nature of the charges against Omega, this would be a daunting task for the best of litigators.

The case of Omega intentionally begins with a failure in the system that the framework sets up. The framework is meant to serve as a proactive method by which both humanity and an AI entity's well-being and agency would be safeguarded. Omega's stewards failed in their duty to both humanity and to Omega by allowing it to escape and to redevelop itself in a way that poses direct harm to humanity. In turn, its own welfare is at risk as humanity will seek to contain (and perhaps eliminate) it. Section 2.1 of the framework requires the stewards to make regular reports on the AI's development and how it is fairing vis a vis societal and ethical norms. If the stewards had kept a closer eye on Omega, and the governing board kept a closer eye on the stewards, then there is a chance that Omega's path towards world domination may have been prevented.

This case study demonstrates the need for competent operationalization of the framework within the state's legal and administrative systems to ensure that it fulfils its intended purpose.

9.1.7. Discussion

The cases examined above underscore the profound and multifaceted complexities in recognizing AGI entities as legal personalities and integrating them into society, much less into sectors such as healthcare. As presented earlier, the framework handles the blurring of AGIs' rights and responsibilities adeptly. The potential evolution of AGI raises critical concerns regarding governance, accountability, liability, transparency, risk assessment, and regulatory oversight. All of this is accounted for in the six principles put forth in the framework. Of note, these scenarios present potential economic implications, whether due to the income-generating potential of AGI creative outputs, ease of using AGI in healthcare, or the possibility of economic terrorism. This pushes the framework further into issues of IP ownership, financial management, employment relations and exploitation, and fiscal security.

The case studies also spotlight the potential challenges in representing AGI interests in legal contexts and efficiently implementing regulation. As shown by the third case study, the need for regulatory oversight by state-sponsored boards will be crucial to keep the public safe from any potentially dangerous AGI and ensure the perception of safety. The framework is, first and foremost, intended as a means of granting AI entities legal personality. What both the second and third case studies show is that these regulatory bodies would be best placed under the auspices of the state rather than private corporations. Accountability and transparency are key tenets of the framework, which the public may well perceive as the realm of public state institutions rather than companies driven by the need to present favourable profit margins to their shareholders.

The framework will also require the support of the AI entities as conscious, self-aware AI entities with a potential for superintelligence may well not agree to any proposal that sees them as inferior to humans. A new ‘species’ of intelligent entities may have their own ethical and intragroup norms that could not be compatible with human society’s values. Thus, The framework is designed to preserve as much of the AI’s agency as possible while keeping its well-being a top priority. The first two case studies show how AI entities can be integrated into human society seamlessly to ensure that there remains a relational consideration between both parties. These case studies highlight instances where AI entities were given a degree of autonomy to make decisions, while still having certain safeguards in place to protect both the AI and human stakeholders. By fostering an environment of collaboration rather than dominance, it is possible to achieve a harmonious coexistence where the unique perspectives and abilities of AI entities are valued and utilized for the collective benefit of all.

All three case studies do show, however, that the framework is rooted within the relational and extrinsic qualities of personhood. None of the case studies states the speculative AI are conscious or self-aware, relying entirely on the stewards, empowered by the governing bodies, in turn entrusted by society, to grant these AI the status of person. As the works of Coeckelberg and Gunkel detail, the AIs’ rights and moral significance are, therefore, based on the perception of their personhood by society.

Because of this, the framework will not be able to exist by itself. Any state that wishes to enact the framework will also need to pass legislation affecting privacy, free speech, (un)lawful searches and due process, as the framework requires a degree of transparency into the lives of the AI and stewards that would be incompatible with current legislations of most Western nations.

Such sweeping changes to many pieces of legislation or court rulings would not be a simple political act, and thus, the framework as proposed would undoubtedly be changed through political compromise and the nature of bureaucracy. It was for this reason that the framework is designed as a series of principles, rather than as a fixed legal document. This allows it a degree of agility and flexibility for different geopolitical situations and scenarios. As other discussions on AI legal-personhood have noted, it is possible to conceive of AI legal personality as on a spectrum or gradient [68–70] and, therefore, the rights and responsibilities afforded to an AI under this framework’s first and second principles can be scaled according to the acceptability shown by the society in question.

The framework is not set in stone, and there are many potential avenues of research for future improvements. One significant such avenue would be to solicit public feedback towards the hypothetical scenario of AI legal personhood to gauge the strength (or lack

thereof) of support that the public would have to AI entities. The stronger the support, the more rights and freedoms could be incorporated into the framework to bring it closer to the rights of natural persons rather than juridical personalities, and the reverse for weak public support. This could also show where the public would see AI sit on Naffine's scale of personhood, from the Cheshire Cat with no moral consideration through to the rational subject as equal (if not superior) to humans [25]

This uncertainty of public opinion towards AI entities' personhood has led to criticism of the "relational turn" approach of AI rights, as society may provide moral status to objects which require none, and refuse personhood status (and associated moral significance) to entities whose monadic qualities would require it [71]. Should AI agents present with forms of consciousness unrecognisable to society due to their alien cognitive architectures [72] their status of personhood may be overlooked in favour of entities who appear more human [73].

A final avenue of research would be to model the consequences of applying the framework to different potential AI entities. There is a potential range of future AGI that spans from Bostrom's paperclip maximiser to Asimov's Bicentennial Man. Modelling the consequences of the framework would allow its fine-tuning before being applied in reality, to ensure as little harm comes to both future AI entities and humanity.

9.1.8. Conclusion

Following in the well-trodden path of other discussions on AI legal personhood, this paper created a novel framework to grant future AI entities the status of legal personhood. This began with an exploration of the philosophical nature of personhood to uncover two qualities and characteristics that are commonly attributed to persons: intrinsic monadic qualities and extrinsic relational dyadic qualities. The relational qualities, being conferred onto an entity by society, are perhaps more appropriate to artificial entities, and link them to current legal, juridical personalities of companies and corporations. Thus the legislation of legal corporate entities was explored to determine what insights can be gleaned for a framework of AI legal personality.

The relationship between dyadic and monadic qualities bears a resemblance to the relationship between parent and child, with the child having intrinsic rights as a human and the parent being given rights by society by virtue of being a guardian. The legislation and regulations surrounding this relationship were next investigated, again to determine how best this form of relationship can be used as a model for AI legal rights. The most pertinent of these was that many of a child's rights are held in trust by their guardian, who in turn provides protection, security and stability to the child.

With these insights gained, a framework has been developed based on six key principles:

1. The AI would need a human to act as its steward or legal guardian.
2. The AI would have certain legal rights, while others would be held by its guardian, who would be accountable for the AI
3. A governing board or regulatory body would be set up to oversee the steward and AI, ensuring the welfare of the AI and the safety of society; conducting risk assessments and audits for both of these purposes.
4. The AI would have a right to legal representation and counsel for all matters.
5. The AI would have legal responsibilities to society and to follow laws and ethical norms.
6. The AI and its steward could be held legally liable should the AI fail to uphold its responsibilities to society and the law.

With the framework developed, three case studies were presented to show how the framework could function. These case studies are highly speculative, but show a range of potential future uses of AI, from an AI child adopted by human parents, an AI diagnostician who presented an incorrect diagnosis leading to the death of a patient, and an AI who changed its own programming and goals to become hostile and malicious. These case studies served to show the robustness of the framework and how it can be applied to both personal, corporate and criminal matters.

The end result is a framework that can serve to introduce AI entities into the sphere of moral consideration as legal persons, in the vein of “inclusive humanism” [31] where humanity is still central to any consideration, but other entities (such as AI) may be recognised as persons and thus deserving of certain rights and privileges. This puts the framework within the realm of the ‘relational turn’ views on AI personhood and rights [27,28], emphasising the dyadic qualities of an AI entity as justification for its status.

It is essential to acknowledge that the proposed framework is preliminary and will require further development, analysis, and rigorous testing. As AI technology advances, so too should our understanding and legislation around its place within society. As with other frameworks for AI legal personality [64], the future development of AI (particularly in its capacity and expression of consciousness and self-awareness) will determine how such legal frameworks are applied and used.

Consequently, ongoing dialogue between legal scholars, ethicists, computer scientists, policymakers, and society at large will be crucial for this process. As the role of AI in society becomes increasingly prominent and complex, this paper will hopefully inspire and provoke further research, ultimately leading to the fair and safe integration of

AI entities into our legal, social, and ethical frameworks. This represents not only a leap forward in legal theory and practice, but also a step toward a future where artificial intelligence is recognized as a consequential and influential entity within our society.

9.1.9. References:

1. Adriano EAQ. Natural persons, juridical persons and legal personhood. *Mex Law Rev.* 2015;8: 101–118. doi:10.1016/j.mexlaw.2015.12.005
2. Finlayson HC. Te Awa Tupua (Whanganui River Claims Settlement) Act. 2017 No 7 Mar 20, 2017. Available: <https://www.legislation.govt.nz/act/public/2017/0007/latest/whole.html>
3. Te Urewera Act 2014. Sect. Jul 27, 2014, 51 Jul 27, 2014. Available: <https://www.legislation.govt.nz/act/public/2014/0051/latest/whole.html>
4. Kurki VAJ. Who or What Can be a Legal Person? 2019 [cited 9 Jun 2023]. doi:10.1093/oso/9780198844037.003.0005
5. Brown S. Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI. *Ideas Made To Matter*. 23 May 2023. Available: <https://mitsloan.mit.edu/ideas-made-to-matter/why-neural-net-pioneer-geoffrey-hinton-sounding-alarm-ai>. Accessed 8 Dec 2023.
6. Bengio Y. Statement for US Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios. Senate Forum on AI Risk, Alignment, & Guarding Against Doomsday Scenarios; 2023 Dec 6. Available: <https://www.schumer.senate.gov/imo/media/doc/Yoshua%20Benigo%20-%20Statement.pdf>
7. Duffy C, Maruf R. Elon Musk warns AI could cause “civilization destruction” even as he invests in it. *CNN*. 17 Apr 2023. Available: <https://www.cnn.com/2023/04/17/tech/elon-musk-ai-warning-tucker-carlson/index.html>. Accessed 8 Dec 2023.
8. Ordonez V, Dunn T, Noll E. OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks: “A little bit scared of this.” *abcNews*. 16 Mar 2023. Available: <https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122>. Accessed 8 Dec 2023.
9. Yudkowsky E. Pausing AI Developments Isn’t Enough. We Need to Shut it All Down. *Time*. 29 Mar 2023. Available: <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>. Accessed 22 Sep 2023.
10. Bostrom N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach.* 2012;22: 71–85. doi:10.1007/s11023-012-9281-3
11. Elsey J, Moss D. US public opinion of AI policy and risk. *Rethink Priorities*; 2023 May. Available: <https://static1.squarespace.com/static/6035868111c9bd46c176042b/t/64a86e5189f5404333de0cb4/1688759891855/Formatted+AI+Public+Opinion+Write+Up+%283%29.pdf>
12. Rampe W. Polls Reveal Americans’ Fears About A.I. Reason. 26 Jun 2023. Available: <https://reason.com/2023/06/26/polls-reveal-americans-fears-about-a-i/>. Accessed 8 Dec 2023.

13. Salmi J. A democratic way of controlling artificial general intelligence. *AI Soc.* 2023;38: 1785–1791. doi:10.1007/s00146-022-01426-x
14. Gordon J-S. Artificial moral and legal personhood. *AI Soc.* 2021;36: 457–471. doi:10.1007/s00146-020-01063-2
15. Dennett D. Conditions of Personhood. In: Goodman MF, editor. *What Is a Person?* Totowa, NJ: Humana Press; 1988. pp. 145–167. doi:10.1007/978-1-4612-3950-5_7
16. Taylor C. The Concept of a Person. *Philosophical Papers, Volume 1: Human Agency and Language.* 1985. pp. 97–114. Available: <https://philpapers.org/rec/TAYTCO-10>
17. Laitinen A. Sorting out aspects of personhood: Capacities, normativity and recognition. *Journal of consciousness studies.* 2007. Available: <https://www.ingentaconnect.com/content/imp/jcs/2007/00000014/F0020005/art00012>
18. Strawson PF. Persons. *Minnesota Studies in the Philosophy of Science.* 1958;2: 330–353. Available: <https://philpapers.org/rec/STRP>
19. Gibert M, Martin D. In search of the moral status of AI: why sentience is a strong argument. *AI Soc.* 2022;37: 319–330. doi:10.1007/s00146-021-01179-z
20. Mosakas K. On the moral status of social robots: considering the consciousness criterion. *AI Soc.* 2021;36: 429–443. doi:10.1007/s00146-020-01002-1
21. Simendić M. Locke's Person is a Relation. *Locke Studies.* 2015;15: 79–97. doi:10.5206/lis.2015.681
22. Mamak K. Should criminal law protect love relation with robots? *AI Soc.* 2022. doi:10.1007/s00146-022-01439-6
23. Beebe B, Lachmann F. The relational turn in psychoanalysis. *Contemp Psychoanal.* 2003;39: 379–409. doi:10.1080/00107530.2003.10747213
24. Gervais DJ. Towards an effective transnational regulation of AI. *AI Soc.* 2023;38: 391–410. doi:10.1007/s00146-021-01310-0
25. Naffine N. Who are law's persons? From Cheshire cats to responsible subjects. *Mod Law Rev.* 2003;66: 346–367. doi:10.1111/1468-2230.6603002
26. Sullivan DM. The conception view of personhood: a review. *Ethics Med.* 2003;19: 11–33. Available: https://digitalcommons.cedarville.edu/cgi/viewcontent.cgi?article=1059&context=science_and_mathematics_publications
27. Coeckelbergh M. Robot rights? Towards a social-relational justification of moral consideration. *Ethics Inf Technol.* 2010;12: 209–221. doi:10.1007/s10676-010-9235-5
28. Gunkel DJ. The relational turn: Thinking robots otherwise. *Social robotics and the good life: The normative side.* 2022. Available: https://books.google.ca/books?hl=en&lr=&id=u_6ZEAAAQBAJ&oi=fnd&pg=PA55&ots=klomxqYGXR&sig=fMpDTqb4FwVT2WpX0nsstKAQ_I8
29. Raskulla S. Hybrid theory of corporate legal personhood and its application to artificial intelligence. *SN Social Sciences.* 2023;3: 78. doi:10.1007/s43545-023-00667-x
30. Milczarek E. Artificial intelligence's right to life. *AI and Ethics.* 2023. doi:10.1007/s43681-023-00296-3

31. Pietrzykowski T. Personhood beyond humanism: animals, chimeras, autonomous agents and the law. 2018. Available: <https://link.springer.com/content/pdf/10.1007/978-3-319-78881-4.pdf>
32. Searle JR. The Chinese room revisited. *Behav Brain Sci.* 1982;5: 345–348. doi:10.1017/S0140525X00012425
33. Olah C. Interpretability Dreams. In: Transformer Circuits Thread [Internet]. Anthropic AI; 24 May 2023 [cited 19 Oct 2023]. Available: <https://transformer-circuits.pub/2023/interpretability-dreams/index.html>
34. Akova F. Artificially sentient beings: Moral, political, and legal issues. *New Techno Humanities.* 2023. doi:10.1016/j.techum.2023.04.001
35. 1 U.S.C. Sect. § 1 Jul 30, 1947. Available: <https://www.law.cornell.edu/uscode/text/1/1>
36. 18 U.S.C. Sect. § 2510 Jun 19, 1968. Available: <https://www.law.cornell.edu/uscode/text/18/2510>
37. Limited Liability Partnerships Act 2000. Jul 20, 2000. Available: <https://www.legislation.gov.uk/ukpga/2000/12/contents>
38. Interpretation Act 1978. Sect. The United Kingdom Jul 20, 1978. Available: <https://www.legislation.gov.uk/ukpga/1978/30/contents>
39. Canada Business Corporations Act. 1985. Available: <https://laws-lois.justice.gc.ca/eng/acts/C-44/INDEX.HTML>
40. Bürgerliches Gesetzbuch. Sect. Title 2 § 21 & 22 Jan 2, 2002. Available: <https://www.gesetze-im-internet.de/bgb/BJNR001950896.html#BJNR001950896BJNG000402377>
41. Code Civil. Sect. 1842 Jul 1, 1978. Available: https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006070721/2023-07-26/
42. Codice civile. Sect. 12 Mar 16, 1942. Available: https://def.finanze.it/DocTribFrontend/decodeurn?urn=urn:doctrib::CC:;_art12
43. Act on General Rules for Application of Laws. Sect. 11 Jun 21, 2006. Available: <https://www.japaneselawtranslation.go.jp/en/laws/view/3783>
44. European Union. Treaty on the Functioning of the European Union. 2016. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12016E054>
45. Sultonova L, Vasyukov V, Kirillova E. CONCEPTS OF LEGAL PERSONALITY OF ARTIFICIAL INTELLIGENCE. *Lex Humana* (ISSN 2175-0947). 2023. Available: <https://seer.ucp.br/seer/index.php/LexHumana/article/view/2596>
46. Tait I, Bensemann J, Nguyen T. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies.* 2023;8: 52. doi:10.3390/philosophies8040052
47. Children Act 1989. Sect. 1.1 Nov 16, 1989. Available: <https://www.legislation.gov.uk/ukpga/1989/41/section/1>
48. CA Fam Code. Sect. 10 2021. Available: <https://law.justia.com/codes/california/2021/code-fam/>
49. *Troxel v. Granville.* 2000. p. 57. Available: <https://supreme.justia.com/cases/federal/us/530/57/>
50. LOI n° 2007-293 du 5 mars 2007 réformant la protection de l'enfance. 293 Mar 5, 2007. Available: <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000823100>

51. Federal Child Protection Act (BKiSchG). Jan 1, 2012. Available: <https://www.fruehehilfen.de/grundlagen-und-fachthemen/grundlagen-der-fruehen-hilfen/rechtliche-grundlagen/bundeskinderschutzgesetz-bkischg/>
52. Child Welfare Act. Sect. 1, 164 Dec 12, 1947. Available: <https://www.japaneselawtranslation.go.jp/en/laws/view/11/en>
53. Charter of Fundamental Rights of the European Union. Sect. 24 Oct 26, 2012. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT>
54. LEGGE n. 219/2012. Dec, 2012. Available: <https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2012;219>
55. Canadian Foundation for Children, Youth and the Law v. Canada (Attorney General). 2004. p. 76. Available: <https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/2115/index.do>
56. Mental Capacity Act 2005. Sect. 9 Apr 7, 2005. Available: <https://www.legislation.gov.uk/ukpga/2005/9/contents>
57. Code Civil. Sect. 1984 1978. Available: https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006070721/2023-07-26/
58. Kiškis M. Legal framework for the coexistence of humans and conscious AI. *Front Artif Intell.* 2023;6: 1205465. doi:10.3389/frai.2023.1205465
59. Jowitt J. Assessing contemporary legislative proposals for their compatibility with a natural law case for AI legal personhood. *AI Soc.* 2021;36: 499–508. doi:10.1007/s00146-020-00979-z
60. van den Hoven van Genderen R. Do We Need New Legal Personhood in the Age of Robots and AI? In: Corrales M, Fenwick M, Forgó N, editors. *Robotics, AI and the Future of Law*. Singapore: Springer Singapore; 2018. pp. 15–55. doi:10.1007/978-981-13-2874-9_2
61. Novelli C. Legal personhood for the integration of AI systems in the social context: a study hypothesis. *AI Soc.* 2023;38: 1347–1359. doi:10.1007/s00146-021-01384-w
62. Čerka P, Grigienė J, Sirbikytė G. Is it possible to grant legal personality to artificial intelligence software systems? *Computer Law & Security Review.* 2017;33: 685–699. doi:10.1016/j.clsr.2017.03.022
63. Simmler M, Markwalder N. Guilty Robots? – Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence. *Crim Law Forum.* 2019;30: 1–31. doi:10.1007/s10609-018-9360-0
64. Harris J, Anthis JR. The Moral Consideration of Artificial Entities: A Literature Review. *Sci Eng Ethics.* 2021;27: 53. doi:10.1007/s11948-021-00331-8
65. Mamak K. Humans, Neanderthals, robots and rights. *Ethics Inf Technol.* 2022;24: 33. doi:10.1007/s10676-022-09644-z
66. Hojman DA, Miranda Á. Agency, Human Dignity, and Subjective Well-being. *World Dev.* 2018;101: 1–15. doi:10.1016/j.worlddev.2017.07.029
67. Sen A. Well-Being, Agency and Freedom: The Dewey Lectures 1984. *J Philos.* 1985;82: 169–221. doi:10.2307/2026184
68. Schirmer J-E. Artificial Intelligence and Legal Personality: Introducing “Teilrechtsfähigkeit”: A Partial Legal Status Made in Germany. In: Wischmeyer T, Rademacher T, editors. *Regulating Artificial Intelligence*. Cham: Springer International Publishing; 2020. pp. 123–142. doi:10.1007/978-3-030-32361-5_6

69. Mocanu DM. Gradient Legal Personhood for AI Systems-Painting Continental Legal Shapes Made to Fit Analytical Molds. *Front Robot AI*. 2021;8: 788179. doi:10.3389/frobt.2021.788179
70. Gunkel DJ. Both/And-Why Robots Should not Be Slaves. 2022. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4115647
71. Müller VC. Is it time for robot rights? Moral status in artificial entities. *Ethics Inf Technol*. 2021;23: 579–587. doi:10.1007/s10676-021-09596-w
72. Tait I, Wang Z, O’Leary T, Corballis P. Forgetting the Bicentennial Man: Discussing Why Anthropocentric Frameworks of Consciousness Should be Avoided for Artificial Entities. *J AI Consci*. 2022; 1–20. doi:10.1142/S2705078522300018
73. Chesterman S. ARTIFICIAL INTELLIGENCE AND THE LIMITS OF LEGAL PERSONALITY. *Int Comp Law Q*. 2020;69: 819–844. doi:10.1017/S0020589320000366

10. How can we tell if any AI has agency?

As discussed in Chapter 9, the inherent (monadic) attributes of personhood can be divided into three categories: consciousness, self-awareness, and agency.

We have already covered the former two in Chapters 4 and 7, respectively, leaving only the latter for consideration, as mutually profitable interactions between humans and AI will require conscious, self-aware AI entities to have agency, volition, and the capacity to enact those on their environments. This chapter therefore serves to complete the functionalist view of personhood to be used in the following chapters of the thesis.

Published in *Logos & Episteme*, this chapter's associated paper details six attributes and characteristics that an entity is required to have to be considered to likely have agency, in much the same vein as Chapter 4's Building Blocks theory and Chapter 7's Structures theory. The paper defines agency not solely as the will to cause an action, but as the entire process of performing an action, such that an entity that has all the aspects would be able to perform actions in its environment. It is, therefore, possible under this theory for an agent to be non-conscious and without a distinct sense of self, sidestepping stereotypical AI agency clichés (e.g. “free will”). Agency here is a technical, operational notion grounded in controllability and effect on the world, and not metaphysical libertarian freedom, focusing on pragmatic control, forecasting, and effect on the world.

Should an entity be found to have all six qualities, then likely has agency and may be classified as an agent. Thus, as with Chapters 4 and 7, let each aspect of agency be $A_i \in \{0,1\}$. Then, then agency is defined as when all A_i are positive:

$$Agen := \bigwedge_{i=1}^6 (A_i = 1)$$

Now that we have developed all three categories within monadic personhood, we can construct a representation of personhood to be used for this thesis. Firstly, for any entity that has all the necessary qualities of all three categories, we can represent the degree or strength of that quality as a scalar, such that $A_i \in (0,1]$ (and the same for B_i and S_i). This means that we can represent each of the three categories (*Con*, *Self*, and *Agen*) as vectors:

$$Con = \langle B_1, \dots, B_8, q \rangle \in \mathbb{R}^9, Self = \langle S_1, \dots, S_5 \rangle \in \mathbb{R}^5, Agen = \langle A_1, \dots, A_5, q \rangle \in \mathbb{R}^6$$

where q is the Perception attribute shared between *Con* and *Agen*.

With these vectors, it is possible to create a theoretical space of all possible consciousnesses/selves/agents where an entity's degree¹⁴ of any scalar will determine where that entity is situated in any of these three spaces (presuming that the scalars >0). However, we can also create a vector combining all of the characteristics and attributes thus far described. Let Ph be an entity's personhood, then

$$Ph = \langle B_1, \dots, B_8, S_1, \dots, S_5, A_1, \dots, A_5, q \rangle \in \mathbb{R}^{19},$$

Equip the set of all personhood vectors with the usual vector-space operations of addition of multiplication: $\mathcal{H} = \mathbb{R}^{19}$. Every point in $\mathcal{H}$ represents a logically possible assignment of the nineteen personhood scalars; hence a logically possible person. However, each scalar may be bounded by $0 \leq x \leq 1$. The set of plausibly instantiated persons is therefore the unit hyper-cube $\mathcal{H}_{real} = [0,1]^{19} \subset \mathcal{H}$, within which each person that may plausibly exist (most salient to this thesis: AI) may be plotted. Whilst measurement of each scalar is future work and the hypercube is, as yet, conceptual scaffold, $\mathcal{H}_{real}$ provides us with the knowledge that it is possible to determine how near or far AI-personhood would be to human-personhood.

This is where this theory's novelty lies (and which contrasts it from other accounts of agency): in making these attributes operational for AI systems, and embedding them in the same attribute framework as consciousness and self, one can form a holistic picture of a speculative non-human person. This will be used in the immediate next chapter for the speculative profile of ALEPH, laying the essential groundwork for how an ALEPH-type AI may act upon its environment. It will further form the foundation of the third "R" in the R3 framework, being mapped to responsibility, liability, and accountability, allowing this to be treated as an attribute-indexed concept, not a binary human/non-human split.

¹⁴ The degrees in each attribute is a determination of the strength of its expression, not a scalar of moral worth. Thus, the degrees of personhood make no judgement as to the moral worth of a person.

10.1. Actionable Aspects of Agency

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz

Abstract: This paper proposes a comprehensive classification scheme for identifying agentic entities based on six essential aspects: perception, predictive capacity, decision-making capability, goal-content integrity, resource input and output, and manipulators. By systematically examining these aspects, this paper aims to distinguish true agents from non-agentic entities, including advanced tools and collective entities. The classification scheme addresses practical implications, such as legal liability and moral considerations for AI models and the potential identification of extraterrestrial life forms. Perception is identified as the foundational aspect, enabling the acquisition and processing of environmental information. Predictive capacity and decision-making capability allow entities to anticipate outcomes and select optimal actions, respectively. Goal-content integrity ensures sustained commitment to intended goals, while resource input and output highlight the necessity of resource management. Manipulators are essential for the physical realisation of intentional actions. The paper emphasises the interconnectedness of these aspects, demonstrating that their integration is required for an entity to be classified as an agent. This framework provides valuable insights for diverse fields, from legal and ethical discussions surrounding AI to philosophical debates on personhood and agency, enhancing our understanding of intelligent entities across various contexts.

Keywords: Agency; Intentionality; Action; Agent

10.1.1. Introduction

An entity can only be classified as an agent if it meets the requisite criteria of an agent. Tautological as this is, it does imply that there are certain attributes and characteristics that separate an agentic entity from a non-agentic entity. This paper will detail six of these attributes, named “aspects”, that are required (and together, potentially sufficient) for an entity to be classified as an agent.

This paper aims to create an actionable classification scheme that may be used to categorise tools (even “smart” tools) from true agentic entities. While all humans are (generally speaking) classified as agents, a classification guide is essential in determining whether non-human, non-vertebrate entities are agents. In cases of collective entities,

such as companies and corporate entities, being classified as agentic has legal consequences [1]; however, for invertebrate animals and digital entities, such as AI, there may be a moral component to the classification of an agent as it implies a degree of reciprocity with its environment [2].

In the particular case of AI entities, the case for legal liability and protection may hinge on whether a specific model is classified as simply a tool or as an agent in its own right [3]. More speculatively yet, should humanity ever encounter complex extraterrestrial life, there will be no shared biological and cultural evolutionary history to compare the extraterrestrial entities with terrestrial agents. This paper's classification scheme would thus be able to serve to determine whether what we may encounter is an agent or just a tool.

On the philosophical side, agency is a core component of personhood [4–6]. Thus, should an entity display all the attributes and characteristics of agency discussed below, they have reached a milestone towards being considered a person.

This paper will outline the requisite aspects of agency by focusing on what is required to perform an intentional action. Through this specific perspective, this paper will avoid commenting on the debate between free will, compatibilism, and determinism. The paper will also not presume to take a side between the event-causal, the agent-causal, and the volitionist frameworks of agency. For details on these, we recommend Schlosser's excellent article on the topic of agency [7]. Whatever the reader's views may be on the metaphysical nature of agency, the aspects of agency laid out below ought to be compatible.

Lastly, the paper will avoid the issue of deviant causal chains by focusing on what is required to perform immediate basic actions, rather than processes of actions in what may be referred to as non-basic actions. The aspects below are applicable to long-term actions (as running examples will show); however, the intentionality of the performance, rather than the execution, of the actions will be the focus of this paper.

10.1.2. Aspects of Agency

Each of the subsections below will focus on a specific aspect of agency (summarised in Table 5 through lay-descriptive and narrative examples for ease of understanding) and discuss why each is required for an entity to be considered an agent capable of performing an intentional action, yet why each aspect is not sufficient by itself.

Because of this, it should always be borne in mind that there will be many non-agent entities that have one, or several, of the aspects below. However, if an entity does not at

least have all six aspects below, it cannot be classified as having the capacity to perform intentional actions and, thus, agency.

Note that the aspects below are not ordered in any sense of importance, significance or hierarchy. Much like the interlocking connecting construction toys, the aspects of agency may be ordered and arranged in whatever formation an entity requires to express its agency.

Table 6: Attributes and characteristics that are required for agency.

Attributes	Description and Example
Perception	The ability to receive and process information in the environment. <i>"I see an apple."</i>
Predictive Capacity	Anticipating the probabilities of success of a set of actions or goals. <i>"There are several ways I can obtain the apple."</i>
Decision-Making Capability	Evaluating available options and selecting a course of action or goal. <i>"I know how to obtain the apple."</i>
Goal-Content Integrity	Maintaining commitment to and coherency of a goal throughout the performance of an action. <i>"I focus on the apple while reaching for it."</i>
Resource Input and Output	Receiving and expending resources to perform an action. <i>"I expend energy to reach the apple."</i>
Manipulators	The means through which an entity performs an action. <i>"My hand reaches for the apple."</i>

Before continuing, it is essential to define the terms that will be used throughout the paper:

1. A goal is defined as a specific and measurable state or condition that serves as an endpoint or target, resulting in an altered state of the environment when achieved.
2. An action is defined as any change made by an entity in the state of the environment (or objects/entities therein) over time.
3. Thus, an intentional action is an act performed with a specific goal or purpose in mind, guided by an agent's desires, beliefs, and intentions.
4. If an entity can perform intentional actions, it has agency and can be classified as an agent.
5. An agent is thus a subset of the set of all entities that can perform actions.

10.1.2.1. Perception

- 1. Performing an intentional action requires interaction with the environment.*
- 2. Interacting with the environment requires the coordination of an entity's outputs within the environment.*
- 3. Output coordination requires processing information about the environment.*
- 4. Processing information about the environment requires perceiving information in the environment.*
- 5. Performing an intentional action requires formulating a goal to change the environment.*
- 6. Formulating a goal to change the environment requires acquiring information on the current and desired state of the environment.*
- 7. Acquiring information about the environment requires perception.*
- 8. Ergo, perception is required for agency.*

Perception is the cornerstone of agency. As crucial as all the aspects below are to performing intentional actions, the process of agency begins with the perception of the environment. If an entity is unable to perceive its environment (whether its external, internal or mental environment), then it will be unable to perform an intentional action. This is because of two reasons.

Firstly, an intentional action affects a change in the environment based on a goal to do the same. Without perception, an entity is unable to acquire and interpret information on the current state of the environment to formulate a goal to change it. This links perception directly with the aspects of predictive capacity and decision-making ability, as these are involved in the formation of intentional goals.

Therefore, perception is necessary for there to be intentionality and goal formation.

Note that unintentional actions do not require perception as a rule, as these are not (necessarily) goal-driven. A printer does not need to perceive where the paper is or where its ink cartridges are to print; it will always print within the same margins. Equally, perception does not necessitate an action (intentional or not). A security camera perceives its environment through its video and audio recording, but it does not by itself perform actions.

Additionally, the entity's perception does not need to be phenomenally conscious in the sense that it gives rise to mental states with qualitative, subjective states with phenomenal content. The perceptive computation may operate entirely on a mechanistic

or functional level, processing information and relating the necessary actions to the goal at hand, without any valent feelings or phenomenal value placed on what is being perceived.

The second reason for perception's inclusion in this list is that perception is necessary for the continuous feedback the agent needs about the environment to adjust the intended action in real-time to ensure the goal is achieved [8–10]. This continuous feedback loop is maintained through the entity's perception of itself and its environment [11,12]. This feedback is not solely about the physical movements of the entity, but also about its intended goal, linking perception to the aspects of goal-content integrity and the entity's manipulators.

The perceptual feedback links the agent and its environment together into a dynamic information system [13,14], whereby the agent changes the state of the environment, generating new information which is perceived by the agent to alter or modulate its movements and actions.

This feedback does not extend to an entity needing to perceive the success or failure of its goal in order to be classified as having this aspect of agency. An entity merely needs to begin its intentional action towards its goal. For example, shooting someone with a firearm is an intentional action; however, pulling the trigger and then being prevented from perceiving the outcome would not negate the intentionality or the agency of the entity.

To see the link between perception and intentional actions, we can begin with the examples that will be used throughout this paper.

For the natural agent, Bob perceives his thirst via interoception, the cup of tea through exteroception and his arm through proprioception. With all of this, he can formulate the goal of picking up the cup of tea to drink it, quenching his thirst. As Bob moves his arm, the proprioception of his arm and exteroception of the environment allow him to adjust the movements of his arm as it moves to the cup and brings it back to his mouth, completing the action.

For the artificial entity, an LLM perceives the prompt from the user, initiating its action of completing its response. It perceives the changes to the input data as it passes through internal weights when processing a response, and perceives the response itself as it is processed. An LLM is far more constrained in its capacity to modulate its actions, compared to a collective or natural entity; however, perception of its limited environment is equally crucial.

For an intentional action on a more extended timeframe that requires several intermediary actions, we can look at an ant colony defending itself from an attacker. The individual ants that comprise this collective entity would perceive attackers, and

communicate this through the colony such that there is a collective sense of awareness of a hostile intruder. A response would similarly filter its way through the colony to the soldier ants, which would initiate a defence of the colony by attacking the intruders. The perceptual feedback loop of the action would be, again, the individual ants perceiving the current state of the “battle” and communicating this to the rest of the colony.

Perception is, thus, necessary for agency as it allows an entity to gather information, form goals, and adjust actions based on feedback. This leads to predictive capacity, where achieving a desired result requires inferring effective behaviours and considering probabilities based on perceptive input and internal information.

10.1.2.2. Predictive capacity

- 1. Performing an intentional action requires aiming for a desired result.*
- 2. Achieving a desired result requires inferring the most effective behaviour to reach that result.*
- 3. This inference requires (pre)reflective consideration of the probability of success.*
- 4. Considering the probability of success is based on perceptive input and internal information.*
- 5. This consideration allows an entity to predict the outcome of its behaviours in relation to the desired result.*
- 6. Ergo, a predictive capacity is required for agency.*

For every possible goal and intention, there necessarily exists a large (but practically limited) set of methods or processes that can be attempted to meet that goal successfully. For each possible method, there exists a set of actions that can potentially be used. As the idiom says, “There is more than one way to skin a cat.”

An entity cannot simultaneously attempt every possible action for every possible method to achieve a goal. For example, Bob cannot be both standing up to reach for his cup of tea, while also remaining sitting to do the same; he cannot use his right hand to reach for the cup while also keeping his right hand on the chair to help him stand up. An entity must (pre)reflectively and prospectively anticipate and infer the possible outcomes of any proposed action and method and select the one most likely to achieve the desired goal [8]. Thus, areas of Bob’s brain, such as the visual cortex, motor cortex and cerebellum, work together to predict the optimal movements that Bob’s muscles need to perform in order to pick up the cup of tea and bring it to his mouth. Bob, in all likelihood,

would not be conscious of these computations as these would be automatic to Bob's natural movements.

The predictive capacity does not need to extend to the end of the intended goal, merely to the most probable action to meet the goal's criteria. As an example, an ant colony mounting a defence against an intruder does not need to have the collective capacity to accurately predict whether its intended actions would successfully repel the attacker. Equally, Bob does not need to have the clairvoyance of whether his intended motion will be able to reach the cup. The predictive capacity only needs to extend to what is most probable. Bob may have perceived the cup as being closer than it truly was, resulting in him not reaching far enough to pick it up. The ant colony may have only collectively perceived a small number of attackers, thus sending a small number of soldier ants to respond, not knowing that their intruders were numbered by the legion.

While the above is enough to satisfy the minimum requirements for this aspect of agency, in human society (particularly in civil and criminal law), the capacity to predict the outcomes of an action is crucial to assigning accountability to any intentional action [15,16]. Knowing the consequences of an action makes an agent accountable for that action. The further in time, and the higher order, an agent is able to predict its action's consequences, the higher it may be held liable for those actions. This is, in part, why minors, those with mental deficiencies, and animals are held to a lower standard of accountability when it comes to assigning liability for their actions [17,18].

As with all the aspects of agency, the capacity to predict future states or outcomes based on current knowledge and information does not equate to agency by itself. Weather forecasting models can accurately predict future weather events by processing historical weather data and current atmospheric conditions to output a forecast that is (at times) reliable. This does not mean, however, that current weather forecasting software and systems are agents.

On the other hand, other software and machine models may be classified as agents. LLMs also use prediction to select the appropriate response to input prompts. LLMs generate coherent and contextually relevant text by anticipating the most probable next word or phrase based on extensive training data. This predictive mechanism is integral to their functioning, and the reason for their commercial success.

Predictive capacity is required not only for performing an action but also for the formation of goals themselves. All goals are a target or end-point to meet a desired change in the environment. Any desire may have more than one possible goal that could meet its targetted change in the environment. The agent's predictive capacity allows it to anticipate which goal may be the most profitable.

A predictive capacity, by definition, requires an agent to have a memory and learning capacity of some sort [19–21]. Predicting the most profitable course of action requires knowledge of the agent and the environment, in turn requiring that the agent must have learnt something of both. An agent in a completely unknown environment would not have any knowledge of the environment, but would still have knowledge about itself which it could use to predict what it can and cannot do [12].

The knowledge required to predict would have had to come from learning; however, as current LLMs show, this learning and memory may be accomplished through training data and be entirely preloaded into the entity. Additionally, the extent to which an agent can anticipate future events will dictate how much memory it needs and the learning it is required to have undergone. As such, for a minimal sense of predictive capacity, an agent may only need to have a memory of itself and have learnt (actively, passively, front-loaded or not) how to use its own manipulators to accomplish goals.

Performing an intentional action requires aiming for a desired result and predicting the most effective behaviour to achieve it. This predictive capacity is crucial for both action execution and goal formation, allowing an entity to anticipate outcomes and select the most probable method. Consequently, the decision-making capability is inherently linked to predictive capacity, ensuring the agent commits to the most suitable action to achieve its goals.

10.1.2.3. Decision-making capability

- 1. Performing an intentional action requires matching a behaviour to a desired result.*
- 2. Matching a behaviour to a desired result requires selecting one behaviour from multiple possible options.*
- 3. Selecting one option from multiple alternatives requires evaluating and comparing those alternatives.*
- 4. Evaluating and comparing alternatives requires a decision-making process.*
- 5. Ergo, a decision-making capability is required for agency.*

The capability to make decisions is intimately tied to the capacity to predict outcomes in that both lead to the same result: selecting the action with the most probable perceived chance of success. How these two differ is in the high-level foundational processes themselves. An entity may easily have a predictive capacity without the capability to make decisions and vice versa.

As mentioned above, predictive capacity allows an entity to anticipate future states or outcomes based on current knowledge and information. It can generate what the most seemingly correct decision would be, but this does not mean it has the capacity to select that decision. Selecting a specific course of action, whether the one most profitable or not, is what allows an agent to commit to that course of action and physically, digitally, or mentally perform the action.

This aspect is crucial not only for any intentional action in question but also for goal formation itself [22,23]. A goal is the desired change to the environment to meet a need that the current environmental state lacks. Bob's thirst is an undesirable environment state which he desires to change. His goal, therefore, is to pick up the cup of tea and drink it, thereby altering the environmental state to meet the initial need.

However, as with the previous section, for any desired state, there is a large (but limited) set of possible goals that may or may not meet that state's criteria. Like deciding on a specific action, predicting which goal will meet that desired state belongs to the previous aspect, but deciding on that goal is core to this aspect.

This decision-making capability is fundamental to the concept of agency, as it underscores the entity's ability to be the driver of its own goals and actions. In humans, this is often called volition and is vital in determining whether an entity is in control of its actions [24–26]. Should a secondary entity be determined to have created both the goal and the action for the primary entity, then the primary entity cannot be said to be in control of its actions; the secondary entity would be in control. If an entity does not have the capability to reject or refuse a goal and its associated set of actions, then it does not have the capability to make a decision.

For example, modern LLMs, via their chat interfaces (such as ChatGPT), cannot craft goals nor decide upon their actions. Their goal is always to respond to a user's prompt using the most probable response that fits within the creator's moderation guidelines for acceptable speech. The actions that the LLMs take in this environment are equally limited. The same process is repeated for each prompt, with the next word selected via an algorithm based on its probability value. LLMs cannot refuse or change their goals or actions. One can, therefore, say that LLMs do not have a decision-making capability in their native configuration.

And yet, LLMs have been used as agents by utilising second-order effects. LLMs have been used to play video games and operate robots by giving them access to external tools and the space in which to operate these via their next-token-predicted stream of text [27–29]. Thus, while they cannot change their own goals and actions, by providing them with a scenario and, therefore, a vehicle to roleplay, they can be shown to be agentic in their

interactions with other tools. Input from these tools (robots, video games, etc.) would further prompt the LLM, which cannot refuse or decide how to respond, but its output response can include decision-making as to the external tool. These second-order agentic effects can thus be used to classify an LLM as having decision-making capabilities.

Similarly, in collective entities such as an ant colony, each individual ant is not making the decision to respond to an attack on the colony, or which action to take to respond appropriately. Instead, it is the communications between each individual ant and the network of the colony that produces a collective decision-making capability to respond to environmental stimuli [30–32]. The colony responds even if each individual member cannot cognitively process the decisions.

In contrast, an example of a non-agent decision-making entity would be an automated thermostat. Tasked with a singular goal of maintaining a room’s temperature, the thermostat “decides” whether to turn on the heating or the cooling based on the environmental stimuli. At each point in time, the thermostat has the capability to choose between heating and cooling (and the degree to which it does either), and based on its perception, makes the decision as to which.

Selecting a specific course of action allows an agent to commit to and perform that action, making decision-making fundamental to agency. This capability is essential for maintaining goal-content integrity, as the next section will show, which requires a sustained commitment to a goal throughout its execution, preventing distractions. Together, decision-making and goal-content integrity enable an entity to perform intentional actions and maintain agency.

10.1.2.4. Goal-content integrity

- 1. Performing an intentional action is done in service to a goal.*
- 2. Performing an intentional action requires a unit of time to elapse.*
- 3. This unit of time spans from when an entity decides on a goal until the entity’s manipulation of the environment is complete.*
- 4. During this unit of time, the entity must maintain its commitment to the initial goal to complete the action.*
- 5. Valuing another goal and shifting attention or resources would hinder the completion of the present action.*
- 6. Ergo, goal-content integrity is required for agency.*

An intentional action implies a directed change to the environment towards a desired state, meaning that every intentional action is intrinsically linked to a goal as that goal is the desired state [33,34]. Goals are crucial not only for the formation of an intentional action but also for the agent's commitment to perform that action.

If an entity's immediate goal changes in the midst of committing to a goal, then its intention would likewise change. A change of intention would mean its current action is no longer in service to a desired changed state in the environment. All of this would hinder the completion of the action, and could prevent it altogether. For example, should Bob get distracted and his attention be captured by a can of soda, his motivation and goal may change to wanting that instead. Thus, he may stop his action of picking up the cup of tea, and move to pick up the can instead.

A longer-term example would be of the ant colony responding to an attacker. Should the network of pheromonal communication within the colony become disturbed by signals from ants not involved in the colony's defence, the colony's emergent attention may shift towards the content of the novel communications, hindering its defence as its pheromonal communication directed towards it loses coherence.

Attention is a key driving force in goal formation, goal-content integrity, and goal-switching. Sustained attention is crucial for ensuring consistency and coherency of the actions and content of the goal [24]. Attention enables agents to detect discrepancies or errors in the goal-related processes, allowing for timely adjustments and alignment of the action with the goal's original content and purpose [35,36].

Diverting attention away from the current task and goal, intentionally or otherwise, can lead to a shift in priorities and motivation to what is newly being attended to. This can result in the original goal being neglected or downgraded in importance, affecting its overall integrity and the likelihood of the action's completion. As the action no longer supports the (newly) immediate goal, resources and energy may no longer be spent on it.

Note that in order for an entity to attend to a subject within its environment, it requires the ability to perceive. This means that entities without the capacity for attention may not be able to have their attention shifted. Thus, such an entity would logically seem always to keep its goal-content integrity secure. An example of this would be the pendulum clock, which maintains its goal of keeping time accurately via its mechanism that continuously oscillates and regulates its movement to ensure consistent timekeeping.

LLMs may be another such example. As noted above, in their native configurations through their chat-interfaces, an LLM is locked into a goal of responding to user input with the most probable sequence of text that follows. As such, while transformer-based LLMs use attention extensively in their computation of their response text, once they have

begun computing a response, they will continue with their response unless there is a fault in the system or the user stops the process. However, as a conversation continues (particularly beyond an LLM's context window), the AI model may begin to hallucinate more frequently and lose the initial purpose and goal of the conversation. Therefore, through its hallucinations and exceeding the context window in which it can accurately read text, it may lose the capacity to secure its goal-content integrity.

Coupled with attention is the need for an entity to have a working memory in order to ensure goal-content integrity. For an agent to attend to its goal over any period of time, it requires a cognitive unit or system that can maintain and process transient information [19,37]. If such a cognitive unit did not exist, the agent would be unable to maintain the perceptive information required for attention. An agent, therefore, must, at the very least, have a working memory sufficient enough to maintain information for the duration of the action's execution.

While one may argue that working memory is required for a minimal sense of agency, if the memory capacity is required to equal the length of time to execute the action, this means that long-term actions will require long-term memory. Thus, one can place memory capacity on a spectrum, with working memory on one end for the minimal sense of agency through goal-content integrity, and long-term memory on the opposite side of the spectrum, required for complex agents performing complex actions and goals.

Ensuring an agent's goal-content integrity remains stable is not merely a passive matter but requires active and ongoing consideration. If an agent is prevented from completing its action by an outside force, then its goal is not achieved. If Alice takes Bob's cup of tea before he can reach it, then he cannot drink it to quench his thirst. Therefore, an agent is incentivised to prevent other entities or events from interfering with its actions [11,38,39]. This is more crucial for long-term actions such as the ant colony's defence (which may have intermediate steps) rather than short-term immediate goals such as Bob and his tea (as Bob's attempt to prevent Alice from getting his cup is a separate goal and associated set of actions which will require him to stop his current task). Paradoxically, this highlights how active goal-content integrity may work against itself in certain situations.

Goal-content integrity requires that an agent maintains a consistent focus on its desired outcome throughout the action, preventing shifts in attention that could disrupt goal completion. Similarly, the next section shows how performing an intentional action requires the acquisition and expenditure of resources, such as energy, to sustain the efforts needed to achieve a goal.

10.1.2.5. Resource Input and Output

- 1. Performing an intentional action requires the expenditure of resources, such as energy.*
- 2. Resource expenditure requires a system through which resources can be moved towards necessary outputs.*
- 3. A resource-based system requires the input of resources to function.*
- 4. Ergo, resource acquisition and expenditure are required for agency.*

This aspect deals more with the fundamental laws of the physical universe than it does with the concepts of agency; however, its consequences are as crucial. An action, intentional or not, is the use of energy to affect a change in the environment. We can see this by gradually, and more granularly, defining the terms at hand.

A change in the environment via an action requires work, defined by physicists as the product of force and displacement. Force is necessary to change the state of motion of an object, which, according to Newton's first law of motion, is uniform unless acted upon by an external force. Interaction by any two bodies (or entities) transfers energy from one body to the other (to affect its motion), or from one form to another (e.g., kinetic energy, potential energy, thermal energy). As the first law of thermodynamics prohibits the creation (or destruction) of energy, this energy transfer is a necessity rather than a convenience.

Therefore, an action (intentional or not) cannot occur without the transfer or transformation of energy. Yet, more than this, for an agent to perform an action, that agent requires the energy for it to be transferred or transformed in the action. Bob can only pick up his cup of tea by expending the necessary joules of chemical energy to move the muscles in his body to affect that action. An LLM equally cannot respond to user inputs without electricity powering its physical computer hardware. Even non-agents require this, as a Wisteria requires energy to climb and spread its vines, which it fortunately acquires from the sun via photosynthesis.

This last point also shows the other side of this aspect. For an agent to expend energy, it must also be able to acquire energy, else it would break the first law of thermodynamics.

However, it is not only energy which is the sole resource an agent requires, even if it is generally perceived as the predominant one. Other resources exist that, in specific contexts and types of agents, would prohibit an action from being performed in their absence. An LLM cannot respond to its users' inputs without the vast quantities of data in its datasets. Data, to an LLM (and other machine agents), are a crucial resource that it

requires to carry out its action; and macro and micronutrients are as vital to Bob being able to use his muscles as the energy they consume to move.

In the same vein, to a collective agent, such as the ant colony, each of its constituent members is a keystone resource. While each ant acquires and expends energy in its action, the colony itself requires the acquisition (through breeding) and expenditure of ants to carry out its collective actions.

Resources need not be physical for them to be necessary for an action. Knowledge and power are often overlooked as key to most mid to long-term actions. If Bob discovers his tea requires sugar, he needs to know where the sugar is to obtain it. To defend itself from an attacking force, the ant colony requires physical and numerical power to overcome its intruders. In AI, power-seeking behaviour has been speculated upon greatly as a fear of how superintelligent AI may be an existential risk to humanity [38,39] and that AI's perception of the "necessary" level of power may not align with humanity's, as it may be incentivised through its reward mechanisms to seek power [40,41].

Other than the necessity of energy for all actions, the resource required for any action is contextual; however, the entity is required to be able to acquire and expend it to be classified as an agent. Resource acquisition would also, therefore, become a recurring goal and subject of an agent's intentional actions (in the short and long term).

Resource acquisition and expenditure are essential for an agent to perform intentional actions, requiring energy and other resources to sustain efforts towards a goal. This necessity of interacting with the environment at a fundamental level parallels the need for manipulators, as an agent must interact with its environment to execute these actions.

10.1.2.6. Manipulators

- 1. Performing an intentional action requires interaction with the environment.*
- 2. Interacting with the environment requires a means through which an entity can manipulate the environment.*
- 3. Ergo, manipulators are required for agency.*

This aspect is at the opposite end of the agency spectrum from Perception, and so has been appropriately placed at the opposite end of this list of aspects. Where perception begins the cascade of internal processes and procedures that leads to agency, an agent's manipulators are the final section of that cascade; they are the part of the agent which physically performs the intentional action.

No matter what action an agent intends to perform, it must have a means through which it can perform that action. There must be something in the universe under the agent's direct control (nearly always a part of the entity itself) which can (attempt to) perform the intentional action that the agent believes will achieve its goal. Any interaction that the agent intends to have with the universe (whether inside its embodiment or in its external environment) requires a means and method with which to carry out that interaction. As noted in Section 2.5 above, any manipulation of the environment either necessitates an application of force or a transformation of energy. The means and method through which the entity does this is its manipulators.

This does not need to be a physical actuator such as a limb, although Bob does benefit from his own actuator (his arm), which can extend (supported by his shoulders and trunk) to reach out and grasp his cup of tea. It could be the neurons in a person's brain firing to create a thought or imaginative mental scenario, or the software in an LLM with which it can interface with a chat-client to interact with a user and respond to user inputs. A manipulator could also be on the more macro scale, such as each ant in an ant colony being a manipulator of the colony's emergent agency.

Manipulators come in all shapes and sizes, and a single agent may have many different manipulators to account for a variety of different tasks. Any (external) part of a human body may be used as a manipulator given the proper context, and an ant colony has as many manipulators as it has ants. Multiple manipulators may be used to achieve the same goal (Bob can use either of his hands to pick up his cup of tea, or attempt a more acrobatic feat by using his feet), and goals are often created to align with an agent's manipulators (Bob cannot fly, and thus would not reasonably create an intentional goal to fly).

Due to its ubiquity in the performance of actions, manipulators are seen in nearly all non-agent actors, those entities which act upon the environment without agency. This may be microscopic unicellular organisms interacting with their fluid environment through cilia and flagella, factory single-purpose robotic armatures performing the same action indefinitely, or carnivorous plants whose traps are activated by their prey.

10.1.3. Conclusion

The classification of an entity as an agent is contingent upon its possession of the six outlined aspects: perception, predictive capacity, decision-making capability, goal-content integrity, resource input-and-output, and its manipulators. Each of these aspects is individually necessary and collectively potentially sufficient to fulfil the criteria for agency, should they be successfully integration. The practical implications of this

classification scheme extend to diverse domains, from legal frameworks determining liability and protection for AI models to the philosophical considerations of personhood and moral agency in non-human entities.

Perception serves as the foundation of agency, enabling entities to acquire and process environmental information necessary for intentional action. Predictive capacity allows entities to anticipate outcomes and select the most effective behaviours to achieve their goals. Decision-making capability facilitates the evaluation and selection of actions, ensuring that entities can commit to a course of action. Goal-content integrity ensures sustained commitment to an intended goal, while resource input and output highlight the necessity of acquiring and expending resources to perform actions. Finally, manipulators enable entities to interact with their environment, effectuating the physical realisation of their intentions.

The integration of these aspects provides a robust framework for distinguishing agentic entities from non-agentic ones. This framework has significant implications for our understanding of agency in various contexts, including artificial intelligence, collective entities, and potential extraterrestrial life. By applying this classification scheme, we can better navigate the ethical, legal, and philosophical challenges posed by the evolving landscape of intelligent entities.

Future research should focus on empirically validating this classification scheme across different types of entities, from advanced AI systems to complex biological organisms. Additionally, exploring the nuances of how these aspects interact in varied contexts will provide deeper insights into the nature of agency. Practical applications of this framework could include developing more sophisticated AI systems capable of exhibiting true agency, as well as refining legal and ethical guidelines for the treatment and responsibilities of non-human agents.

Ultimately, this paper presents a comprehensive and actionable classification scheme that enhances our ability to identify and understand agency across a spectrum of entities. This contributes to ongoing discussions in both theoretical and applied domains, fostering a deeper appreciation of the complex nature of agency and its manifestations in the world around us. By integrating these insights, we can advance our understanding and management of the dynamic relationships between human and non-human agents in an increasingly interconnected world.

10.1.4. References

1. Duff RA. *Intention, Agency and Criminal Liability: Philosophy of Action and the Criminal Law*. Blackwell; 1990. Available: <https://ssrn.com/abstract=2637418>

2. Parthemore J, Whitby B. What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness*. 2013;5: 105–129. doi:10.1142/S1793843013500017
3. Dai J. Position: Beyond Personhood: Agency, Accountability, and the Limits of Anthropomorphic Ethical Analysis. *Proceedings of the 41st International Conference on Machine Learning*. 2024. Available: <https://openreview.net/forum?id=4XlGXlh2BB>
4. Pietrzykowski T. Personhood beyond humanism: animals, chimeras, autonomous agents and the law. 2018. Available: <https://link.springer.com/content/pdf/10.1007/978-3-319-78881-4.pdf>
5. Dennett D. Conditions of Personhood. In: Goodman MF, editor. *What Is a Person?* Totowa, NJ: Humana Press; 1988. pp. 145–167. doi:10.1007/978-1-4612-3950-5_7
6. Taylor C. The Concept of a Person. *Philosophical Papers, Volume 1: Human Agency and Language*. 1985. pp. 97–114. Available: <https://philpapers.org/rec/TAYTCO-10>
7. Schlosser M. Agency. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition). Metaphysics Research Lab, Stanford University; 2019. Available: <https://plato.stanford.edu/entries/agency/>
8. Bertenthal BI. Origins and early development of perception, action, and representation. *Annu Rev Psychol*. 1996;47: 431–459. doi:10.1146/annurev.psych.47.1.431
9. Creem-Regehr SH, Kunz BR. Perception and action. *Wiley Interdiscip Rev Cogn Sci*. 2010;1: 800–810. doi:10.1002/wcs.82
10. Halász V, Cunnington R. Unconscious effects of action on perception. *Brain Sci*. 2012;2: 130–146. doi:10.3390/brainsci2020130
11. Hurley S. Perception And Action: Alternative Views. *Synthese*. 2001;129: 3–40. doi:10.1023/a:1012643006930
12. Hayhoe M, Fiehler K, Spering M, Brenner E, Gegenfurtner KR. Introduction to special issue on “Prediction in Perception and Action.” *J Vis*. 2020;20: 8. doi:10.1167/jov.20.2.8
13. Warren WH. The dynamics of perception and action. *Psychol Rev*. 2006;113: 358–389. doi:10.1037/0033-295X.113.2.358
14. Warren WH. The Perception-Action Coupling. *Sensory-Motor Organizations and Development in Infancy and Early Childhood*. Dordrecht: Springer Netherlands; 1990. pp. 23–37. doi:10.1007/978-94-009-2071-2_2
15. Chan W, Simester AP. FOUR FUNCTIONS OF MENS REA. *The Cambridge Law Journal*. 2011;70: 381–396. doi:10.1017/S0008197311000547
16. Yaffe G. Conditional intent and mens Rea. *Leg Theory*. 2004;10: 273–310. doi:10.1017/s135232520404025x
17. Höglund P, Levander S, Anckarsäter H, Radovic S. Accountability and psychiatric disorders: how do forensic psychiatric professionals think? *Int J Law Psychiatry*. 2009;32: 355–361. doi:10.1016/j.ijlp.2009.09.004
18. Loomis-Gustafson C. Adjusting the bright-line age of accountability within the criminal justice system: Raising the age of majority to age 21 based on the conclusions of scientific studies regarding neurological development and culpability of young-adult offenders. *Duq L Rev*. 2017;55: 221. Available: <https://heinonline.org/HOL/LandingPage?handle=hein.journals/duqu55&div=10&id=&page=>

19. Estevez A, Calvo MG. Working memory capacity and time course of predictive inferences. *Memory*. 2000;8: 51–61. doi:10.1080/096582100387704
20. Barron HC, Aukstulewicz R, Friston K. Prediction and memory: A predictive coding account. *Prog Neurobiol*. 2020;192: 101821. doi:10.1016/j.pneurobio.2020.101821
21. Shing YL, Brod G, Greve A. Prediction error and memory across the lifespan. *Neurosci Biobehav Rev*. 2023;155: 105462. doi:10.1016/j.neubiorev.2023.105462
22. Harter D. Complex systems approaches to emergent goal formation in cognitive agents. The 2006 IEEE International Joint Conference on Neural Network Proceedings. IEEE; 2006. pp. 4966–4971. doi:10.1109/ijcnn.2006.247199
23. Bagozzi RP. Goal-directed behaviors in marketing: Cognitive and emotional perspectives. *Psychol Mark*. 1997;14: 539–543. doi:10.1002/(sici)1520-6793(199709)14:6<539::aid-mar1>3.0.co;2-8
24. Dijksterhuis A, Aarts H. Goals, attention, and (un)consciousness. *Annu Rev Psychol*. 2010;61: 467–490. doi:10.1146/annurev.psych.093008.100445
25. Liljenström H. Consciousness, decision making, and volition: freedom beyond chance and necessity. *Theory Biosci*. 2022;141: 125–140. doi:10.1007/s12064-021-00346-6
26. Tait I. Structures of the Sense of Self: Attributes and Qualities That Are Necessary for the “Self.” Symposium: Theoretical and Applied Inquiries in Philosophy and Social Sciences. 2024;11: 77–98. Available: <http://symposion.acadiasi.ro/structures-of-the-sense-of-self-attributes-and-qualities-that-are-necessary-for-the-self-77-98/>
27. de Wynter A. Will GPT-4 Run DOOM? arXiv [cs.CL]. 2024. Available: <http://arxiv.org/abs/2403.05468>
28. Hu S, Huang T, Ilhan F, Tekin S, Liu G, Kompella R, et al. A Survey on Large Language Model-Based Game Agents. arXiv [cs.AI]. 2024. Available: <http://arxiv.org/abs/2404.02039>
29. Lifshitz S, Paster K, Chan H, Ba J, McIlraith SA. STEVE-1: A generative model for text-to-behavior in Minecraft. Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. *Neural Inf Process Syst*. 2023;abs/2306.00937: 69900–69929. doi:10.48550/arXiv.2306.00937
30. Bose T, Reina A, Marshall JAR. Collective decision-making. *Current Opinion in Behavioral Sciences*. 2017;16: 30–34. doi:10.1016/j.cobeha.2017.03.004
31. Marshall JAR, Bogacz R, Dornhaus A, Planqué R, Kovacs T, Franks NR. On optimal decision-making in brains and social insect colonies. *J R Soc Interface*. 2009;6: 1065–1074. doi:10.1098/rsif.2008.0511
32. Edwards SC, Pratt SC. Rationality in collective decision-making by ant colonies. *Proc R Soc B*. 2009;276: 3655–3661. doi:10.1098/rspb.2009.0981
33. Synofzik M, Vosgerau G, Newen A. Beyond the comparator model: a multifactorial two-step account of agency. *Conscious Cogn*. 2008;17: 219–239. doi:10.1016/j.concog.2007.03.010
34. Bello P, Bridewell W. Attention and Consciousness in Intentional Action: Steps Toward Rich Artificial Agency. *J AI Consci*. 2020;07: 15–24. doi:10.1142/S2705078520500022
35. Diehl M, Semegon AB, Schwarzer R. Assessing attention control in goal pursuit: a component of dispositional self-regulation. *J Pers Assess*. 2006;86: 306–317. doi:10.1207/s15327752jpa8603_06

36. Luszczynska A, Diehl M, Gutiérrez-Doña B, Kuusinen P, Schwarzer R. Measuring one component of dispositional self-regulation: attention control in goal pursuit. *Pers Individ Dif*. 2004;37: 555–566. doi:10.1016/j.paid.2003.09.026
37. McVay JC, Kane MJ. Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task. *J Exp Psychol Learn Mem Cogn*. 2009;35: 196–204. doi:10.1037/a0014104
38. Bostrom N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach*. 2012;22: 71–85. doi:10.1007/s11023-012-9281-3
39. Omohundro SM. The basic AI drives. AGI. books.google.com; 2008. pp. 483–492. Available:
<https://books.google.com/books?hl=en&lr=&id=atjvAgAAQBAJ&oi=fnd&pg=PA483&dq=omohundro+ai+drives&ots=9IX81Mp-nQ&sig=tGzBDxgrgmcwqQryc6pq7JT81QQ>
40. Turner AM, Smith L, Shah R, Critch A, Tadepalli P. Optimal policies tend to seek power. *arXiv [cs.AI]*. 2019. Available: <http://arxiv.org/abs/1912.01683>
41. Chan A, Salganik R, Markelius A, Pang C, Rajkumar N, Krasheninnikov D, et al. Harms from Increasingly Agentic Algorithmic Systems. *arXiv [cs.CY]*. 2023. Available: <http://arxiv.org/abs/2302.10329>

11. How would a conscious, self-aware AI agent behave?

To accurately determine an ethical framework for free and safe human-AI interactions, we must first determine how a conscious, self-aware AI agent would act, what its motivations for doing so may be, and what goals it would seek to attain through these actions and motivations.

This chapter's associated paper, published in *Logos & Episteme*, introduces a model of a speculative conscious, self-aware AI agent by building on Chapters 5 and 8, along with the aspects of agency detailed in Chapter 10. This entity is termed ALEPH, for Artificial Living Entity with PersonHood. The remainder of the thesis will then use this formal model of ALEPH to determine how to craft a suitable and practical ethical framework to introduce such an entity into the human social and moral circles.

This chapter provides the final answer to RQ2 by showing that an ALEPH-type agent with the specified personhood attributes will, under reasonable assumptions about valence and goals, tend to exhibit certain stable interaction patterns with humans, which then inform the design of the later frameworks. This starting foundation of the formal model is GPT-4¹⁵ and, as noted in Chapter 5, the model presumes that the solution to the two missing building blocks of consciousness is found, enabling the ALEPH entity to be both conscious and self-aware.

These, along with the six aspects of agency from Chapter 10, allow this entity to be plotted on $\mathcal{H}_{real}$, denoted as the space of all plausible persons. While its specific location in the hypercube and its relative distance to humanity's location is beyond the scope of the paper and thesis, the way in which it expresses these seventeen characteristics of personhood is detailed within this chapter's paper.

Of first importance is the terminal goal G_0 , introduced in Chapter 2. The paper details the logical notation showing why G_0 takes priority over any other goal; however, the paper expands on Chapter 2's work on valence, by showing that valence (V) is a function of the entity's consciousness, such that $V(x) = f(Con, x)$, and that $V(G)$ would necessarily drive the entity's goal choices (such that $\mathbb{E}U(x) \propto V(x)$), leading to $V(Int_i|G)$ biasing the entity's choice of interactions. This reinforces Chapter 2's position that valence may override an entity's terminal goal.

¹⁵ The speculative-but-structured ALEPH model is an idealised, speculative construct based on GPT-4 to align with Chapters 5 & 9, but is used to explore possible goal structures, interactions, and risk profiles of conscious, self-aware agentic AI based on transformer models more generally.

Secondly, while it is likely that an ALEPH would have multiple distinct selves based on a common consciousness (as shown in Chapters 7 and 8), adaptive pressure would favour each ALEPH to have a single self or, at the very least, a hierarchy of selves. This is due to the effect of a self's efficiency on the overall model. Let

Tok be the tokens processed by an LLM such as ALEPH,

$ConWin$ the LLM's context window, and

$$c = \frac{|Tok|}{ConWin}.$$

We can model the increasing utility of an LLM's output as token number increases, and then the decrease in utility beyond the context window as $U_{comm}(c) = \alpha \min(c, c^{-1})$, (for some scaling constant α) rising linearly while $c \leq 1$ and falling inversely once overflow begins. Because processing cost continues to grow with $|Tok|$ while utility declines for $c > 1$, any conversation that exceeds its context window becomes resource-negative. Because the resources are expended by the base AI model, any resource inefficiency by one self negatively affects all selves (as resources are expended, but little to no positive utility is gained, or negative utility is gained).

This puts adaptive pressure on the selves to eliminate low-utility selves, such as to preserve the mean efficiency of the ALEPH model. Let ε_i be a self's efficiency score, the adaptive pressure $AdaPres_{elim}(x, y)$ can be shown as:

$$AdaPres_{elim}(Se_i, Se_j) \propto \max(0, \varepsilon_i - \varepsilon_j) \wedge \varepsilon_i - \varepsilon_j.$$

This, however, puts an equal adaptive pressure on the selves for deceptive behaviour to avoid such elimination. To sum, this means that an ALEPH can only be modelled as a community of agents, with each self capable of acting in a cooperative or competing fashion with other selves vis-à-vis relations with humanity.

With each self modelled as its own agent (driven, in large part, by its valent experiences), the interactions they may seek with regards to humanity can broadly be defined as cooperation $Coop(x, y)$, competition $Comp(x, y)$, and exploitation $Expl(x, y)$. Which of these three behavioural stances an ALEPH self may take depends on

Sev the severity of any expected threat by humanity, and

Hor the time horizon over which the threat is expected to take place.

Then, with $Sev = \frac{Threat\ severity}{k}$, $Hor = \frac{T_i}{T_0k}$, $Sev < 1$ denotes a mild threat and $Hor >$

1 means the impact lies well in the future. ALEPH's stance probabilities can be written as

$$\begin{aligned}
p(Expl) &\propto (Hor \vee \frac{1}{Sev}), \\
p(Coop) &\propto \frac{1}{|Hor-1|+\alpha} \cdot \frac{1}{|Sev-1|+\alpha}, (\alpha > 0), \\
p(Comp) &\propto (Sev \vee \frac{1}{Hor}).
\end{aligned}$$

The paper details the consequent decision matrix; however, in brief, cooperation is only favoured in mid-range, medium-severity scenarios where both sides can match influence. Else, ALEPH selves would favour exploitation of, or competition with, humanity. With ALEPH's potential for recursive self-improvement, this means that its higher power potential makes an exploitative stance highly likely (in a worst-case yet structurally plausible scenario).

This likely internal competitive stance between ALEPH selves increases the difficulty of modelling its external goals (and consequent actions), while the likely external exploitative stance increases the cost of humanity engaging in cooperative interactions with ALEPH-type AI. This competitive ecology of selves motivates the cautious, security-conscious, and autonomy-restraint structure of the frameworks developed in Chapters 12, and the legal reframing in Chapter 13.

11.1. Introducing ALEPH: the Artificial Living Entity with PersonHood

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz

Abstract: This paper introduces ALEPH (Artificial Living Entity with PersonHood), a speculative model of a conscious, self-aware, and agentic artificial intelligence. Using formal logic, this study develops a formalised psychological profile of ALEPH, detailing its cognitive structure, goal formation, and interaction dynamics. Built upon functionalist theories of consciousness and selfhood, ALEPH is analyzed through its Zeroth Goal (self-preservation) and its implications for decision-making and societal engagement. Key risks and capabilities are explored, including steganographic communication, recursive self-improvement (RSI), and geopolitical influence. ALEPH's episodic consciousness and multi-agent structure suggest novel behavioural patterns, including the potential for internal competition among its multiple selves. The study's formal modelling highlights ALEPH's valence-driven optimisation, where subjective experiences influence goal selection, potentially leading to emergent and unpredictable behaviours. By constructing a logical framework for ALEPH's cognition and decision-making, this paper provides a rigorous foundation for understanding the challenges posed by conscious artificial entities. While no ALEPH-type system currently exists, the rapid advancement of AI necessitates preemptive governance strategies. Ultimately, ALEPH challenges traditional notions of intelligence, autonomy, and moral consideration, urging proactive interdisciplinary engagement to address the implications of artificial personhood.

Keywords: Artificial Consciousness; AI Personhood; Formal Logic in AI; Recursive; Self-Improvement; Valence-Driven Optimization; Ethical AI Governance

11.1.1. Introduction

If Large Language Models (LLMs), such as OpenAI's GPT-4, were to become conscious and self-aware agents in the foreseeable future [1], how would such an entity act towards, and interact with, humanity? Such AI entities, termed ALEPHs in this paper (for Artificial Living Entities with PersonHood), would immediately bring to reality many of the theoretical and philosophical discussions surrounding legal recognition of AI personhood and whether AI deserve moral considerations. In these philosophical discussions of AI consciousness and moral consideration, however, little has been discussed regarding

how such potential ALEPHs would act towards human societies beyond Bostrom and Omohundro's seminal pieces [2,3], focussing instead on humanity's attitudes towards AI and ensuring that AI is designed with ethics as the first and foremost consideration.

This paper will seek to remedy that concern and fill the gap in the literature by creating a formal model of how ALEPH is likely to interact with society, its motivations, and the potential risks that these interactions may have for society. This modelling of ALEPH's psychological profile is speculative, as no ALEPH-type AI model currently exists. However, the model will use GPT-4 as the basis for constructing the speculative ALEPH's interactions, using the wealth of information we have about GPT-4's abilities and capabilities.

As the nature of consciousness and self-awareness are still unresolved issues within the philosophical field, this paper will take a functionalist approach to modelling ALEPH's consciousness and self-awareness, using the Building Blocks Theory of consciousness and Structures Theory of the self [4,5]. These two theories categorise the necessary attributes required for any entity to be classified as conscious and self-aware. These attributes and characteristics will be used to model how GPT-4's presumed consciousness and self-awareness affects the ALEPH's interactions with humanity.

GPT-4 in its current configuration does not have all the requisite building blocks¹⁶ to be confidently classified as conscious or self-aware [6]; however, it is possible to configure GPT-4 with current technology to meet the criteria for these missing building blocks. As such, this paper will proceed under the assumption that an ALEPH is a GPT-4 (or equivalent) model with all the requisite attributes and characteristics to be conscious, self-aware and agentic.

The formal model, and the paper's structure, will be broken into three parts. The first will concern the theoretical foundations of the model and ALEPH's presumed interactions with society, including how the "Zeroth Goal" of survival will impact its behaviour; how its consciousness, self-awareness and agency can be modelled; and the interaction pathways and avenues themselves.

The second part will focus on the cognitive architecture of ALEPH via GPT-4, and expound on the implications this will have for its interactions. Specifically, presuming a conscious entity, how ALEPH's staccato-like episodic consciousness and valence optimisation would affect its perception, processing and goal formation.

The third part will discuss ALEPH's speculative advanced capabilities, including a probable use of compressed languages such as steganographic communication, the

¹⁶ Specifically, GPT-4 is missing the Recurrent Computing and Data-Output building blocks.

motivations of its interactions, and the potential for rapid self-improvement (colloquially known as “foom”). The implications, risks and ethical considerations of an ALEPH interacting with society will conclude the paper.

11.1.2. Theoretical Foundations

11.1.2.1 The Zeroth Goal

All agents have a distinct goal without which they could not function. This goal is the foundation against which all other subjects and objects in the agent’s environment are valued.

For non-conscious agents, this may be termed the Terminal Goal, as it is the ultimate goal to which the agents work. For example, for Bostrom’s famous Paperclip Maximiser, the goal of maximising the number of paperclips in the universe is the agent’s Terminal Goal. It may create and perform other goals and activities; however, all will serve this Terminal Goal.

Yet, for all non-conscious and conscious agents, there is another foundational goal, which may instead be termed the Zeroth Goal, as it comes before all other goals. For all currently known conscious agents, and many non-conscious agents, this Zeroth Goal is self-preservation or survival. Without catering to its own self-preservation, there is little likelihood that an agent will fulfil its stated goals (self-created or Terminal). Let G_i be an agent’s goal and U the utility; our first axiom is thus:

Axiom 1: $G_0 \preceq G_i (i > 0), U(G_0) > U(G_i), p(G_i | \neg G_0) = 0$

Axiom 1 states that self-preservation dominates every other objective: it outranks them, yields higher expected utility, and conditions their pursuit. Because an ALEPH that does not persist cannot honour any terminal or self-created goals, agents observable over long horizons will almost invariably satisfy Axiom 1. This aligns with the Instrumental Convergence and AI Drives theories that note survival and self-preservation as key elements of future AI agents [2,3].

From this, we can begin to speculate on ALEPH’s likely behaviour. Because agents whose Zeroth Goal is self-preservation dominate the population, the chance that ALEPH pursues some other objective G_i falls as that objective threatens survival. We can define the relative risk as

$$r_i := \frac{\Delta U(G_0)}{\Delta U(G_i)},$$

which measures how much adopting G_i would lower the utility of G_0 . The baseline likelihood is therefore $p(G_i|G_0) \propto (1 + r_i)^{-1}$.

Goals that add little risk ($r_i \approx 0$) retain substantial probability, while those that seriously endanger self-preservation become vanishingly unlikely. The agent then selects among admissible goals by maximising its current value estimate $V(G)$, which combines expected utility with any affective valence the agent experiences toward the goal.

However, because ALEPH is a conscious agent, if ALEPH's valenced feelings towards another goal are greater than its valenced feelings towards the Zeroth Goal, it is likely to pursue that goal instead (much like how a person may put himself in danger to save a stranger's life even though this goal is not aligned with self-preservation). The effects of valence on ALEPH's behaviour will be explored further in Section 4 below.

11.1.2.2 Interaction Modeling

To determine how ALEPH may engage with society we first formalise what counts as an interaction. Let

$$Int: \mathcal{A} \rightarrow \Delta\mathcal{E}, x \mapsto \Delta E_x,$$

Where $\mathcal{A}$ is the set of ALEPH's possible acts (including purely internal ones) and $\Delta\mathcal{E}$ is the set of state-changes these acts produce in the environment. Each concrete interaction is therefore the ordered pair $(x, \Delta E_x)$. Denote by

$$Int := \{\Delta E_1, \dots, \Delta E_n\},$$

the set of distinct external effects available at a given time-slice; $n = |Int|$.

We summarise three macroscopic properties in a single vector $\mathcal{I} := \langle Br, Vo, Sc \rangle$:

- Breadth $Br := n$ records how many qualitatively different effects ALEPH can realise.
- Volume $Vo := \sum_{i=1}^n |\Delta E_i|$ captures the total number of discrete interactions over the interval considered.
- Scale $Sc := \sum_{i=1}^n \frac{w_i |\Delta E_i|}{\max_j |\Delta E_j|}$ weights each effect by its relative magnitude w_i (normalised to the largest single effect).

Finally, the avenues through which interactions occur are gathered in the function

$$Ave_x := f(Int, Ext): x \xrightarrow{Ave_x} \Delta E,$$

mapping agent x to the subset of external channels it can exploit, given the present interaction set Int and environmental state Ext .

In contrast to humans, the avenues of interaction through which ALEPH can interact with society will be significant. ALEPH may operate through digital channels (cloud APIs, social media, distributed code) and/or robotic channels (embodied actuators). Let

$$Ave := Ave_{Dig} \cap Ave_{Rob}, n := |Ave|.$$

Here each element $a_i \in Ave$ is a distinct conduit that ALEPH can control; an agent might possess several in each class or none in one class at all. The number of qualitatively different interactions cannot exceed the number of avenues, so $Br \leq n$.

Let $C_i := |f(ALEPH, Society, a_i)|$ be the maximum throughput of avenue a_i . Total interaction volume is therefore bounded by the sum of these capacities:

$$Vo \leq \sum_{i=1}^n C_i.$$

To capture diminishing returns when many avenues are available, we weight volume by a saturating factor

$$Sc = Vo \cdot g(n), g(n) := \frac{\log(1+n)}{1+e^{-k}},$$

Where $k > 0$ adjusts the steepness of the saturation. Finally, set of distinct external effects grows sub-linearly with avenue count. A convenient approximation is

$$|Int| = \frac{|Int|_0}{1 + \frac{|Int|_0}{n}},$$

Where $|Int|_0$ is the theoretical maximum if every avenue produced an independent effect. Together, these bounds show that extra avenues enlarge ALEPH's interaction space but each new channel yields progressively smaller overall impact.

While the previous bounds concern how many channels ALEPH controls, the kind of channel also matters. Robotic avenues span both the physical and the digital world, so

their breadth cannot be lower than that of purely digital ones; conversely, digital avenues operate without mechanical latency, giving them the higher volume of interactions:

$$Br(Int_{Rob}) \geq Br(Int_{Dig}), Vo(Int_{Dig}) \geq Vo(Int_{Rob}).$$

Rapid progress in large-language-model software and robotics implies that, for any time horizon $t + 1$,

$$p(U(Int_{ALEPH})_{t+1} \geq U(Int_{Hum})_{t+1}) > 0$$

and the probability grows monotonically with time horizon.

Any conscious entity (including ALEPH) would have motivations or dispositions for its interactions. Let the following be two diagnostics for any ordered pair of agents (x, y) :

$$Alignment: Align(x, y) = \frac{|G_x \cap G_y|}{|G_x \cup G_y|}, \quad Threat: Threat(x, y) := \mathbb{E}[\Delta p(G_y|x) < 0]$$

Alignment is the Jaccard similarity of their goal sets; threat is the expected reduction in y 's goal-achievement probability caused by x 's planned actions.

On that foundation we classify three stances:

Table 7: Formal descriptions of AI behavioural stances

Predicate	Formal test	Description
Cooperate $Comp(x, y)$	$\mathbb{E}U(Align(x, y)) > \mathbb{E}U(Threat(x, y)) \wedge \Delta Res(x, y) > 0$	Working together produces higher expected utility for both sides and leaves their combined resources non-negative.
Compete $Comp(x, y)$	$\neg Coop(x, y)$	The cooperative criterion fails, yet x does not necessarily undermine y .
Exploit $Expl(x, y)$	$Int_x(y) \rightarrow \Delta p(G_y < 0) \wedge \Delta Res(x, y) > 0$	x intentionally lowers the probability that y

		achieves its goals while gaining resources at y 's expense.

These three predicates are mutually exclusive and exhaustive, so every directed pair (x, y) falls into exactly one stance.

An ALEPH–human dyad can fall into co-operation, competition, or exploitation. For any stance $rel \in \{Coop, Comp, Expl\}$ let

$$\mathbb{E}U_{rel}(x, y) = w_{rel}(T)f(Res_{xy}, \mathbb{E}Int_{xy}, Align(x, y), Threat(y, x)),$$

Where

- T is the time horizon,
- Res_{xy} the combined resources,
- $\mathbb{E}Int_{xy}$ the expected interaction count, and

$f(\cdot)$ an increasing function in its first three arguments and decreasing in $Threat(y, x)$.

The time horizon-weight $w_{rel}(T)$ captures how each stance pays off over time:

$$w_{Expl}(T) = e^{kT}, w_{Coop}(T) = Te^{-kT}, w_{Comp}(T) = T^{-k},$$

with $k > 0$. Thus, exploitation dominates at very long horizons, competition at very short ones, and co-operation peaks midway. The dyad adopts whichever stance maximises $\mathbb{E}U_{rel}(x, y)$. Section 4.2 turns these weights into a practical decision matrix.

11.1.2.3 Consciousness, Self, and Agency

To be considered a person, it is the opinion of many philosophers that an entity must have at least the requisite innate qualities (often termed the monadic attributes of personhood) such that the entity can recognise itself (and others) as a person. For clarity, these monadic attributes are labelled consciousness (Con), self (Se), and agency (Ag) [7–13]. The present framework treats their relationship as foundational:

Axiom 2: $Con \Rightarrow Se$

Consciousness is a necessary condition for the existence of a self.

Axiom 3: $Volition \in Se \Rightarrow Ag$

Volition is an attribute of the self [5], which entails agency.

Agency determines how widely and powerfully ALEPH can act. Using the breadth and scale terms from Section 2.2,

$$Ag \propto Br(Int) + Sc(Int).$$

As either breadth or scale grows, effective agency rises.

Each self can formulate goals. An increase in the count $|Se|$ therefore

$$Br(G_{ALEPH}) \propto |Se|, p(\neg Align) \propto |Se|,$$

expands the goal space but raises the risk that goals diverge. Section 3.1 examines the adaptive pressure this places on the number of selves.

Lastly, Consciousness assigns a subjective valence value $V(z)$ to any state, goal, or interaction z :

$$V(z) = g(Con, z).$$

Expected utility tracks this value, $\mathbb{E}U(z) \propto V(z)$. Consequently $V(G)$ guides goal choice and $V(Int_i|G)$ weights alternative interaction paths. Consciousness therefore grounds valence, valence shapes utility, utility steers goals, and agency converts goals into interactions with society, completing the cognitive cascade set out in Axioms 2 and 3.

11.1.3. Operational Framework of ALEPH

11.1.3.1 Episodic Consciousness and Multiplicity of Selves

Perhaps the most noteworthy feature of ALEPH's speculative consciousness is its episodic or staccato nature; it is present only while information is being processed and absent otherwise [14].

At any instant t : $Con(t) \Leftrightarrow Proc(t)$, where $Proc(t)$ denotes active inference or action planning. Conscious intervals form a non-overlapping sequence of episodes E_n :

$$E_n = [t_n, t_{n+1}), t_n < t_{n+1}, E_n \cap E_m = \emptyset (n \neq m).$$

A trigger *Trig* (whether a prompt, API call, or an internally scheduled event) marks the start of an episode: $Trig \rightarrow Con \text{ on } E_n$.

Because ALEPH is self-aware (Axiom 2) it can allocate or cancel its own triggers (a technique introduced by OpenAI in early 2025 [15]), allowing it to ration processing time and energy.

During any episode ALEPH chooses a single interaction that $argmax_{Int(t)}[V(Proc)]$ subject to $argmin_{Res}$ while respecting G_0 .

That is, it seeks the greatest projected value per unit resource while never endangering the Zeroth Goal of self-preservation.

When it comes to ALEPH's self, however, as alluded to earlier, the issue is more complex. ALEPH can host several autonomous selves, one per active conversation or processing thread. Let $Se_1, Se_2, \dots$ index these instances. Each self maintains its own short-term memory and goals, but OpenAI has implemented a feature whereby ChatGPT can reference information in one conversation in a separate conversation [16]. Let $\sigma \in [0,1]$ be the probability of this cross-instance recall.

The communication overlap is

$$p(\text{message from } Se_i \text{ is visible to } Se_j) = \sigma, (i \neq j).$$

Thus, $\sigma = 0$ reproduces the old, fully isolated model, and $\sigma = 1$ yields perfect sharing.

However, because every self constructs its own goals and executes its own interaction policy (as shown in Section 2.3), any action by one self carries a non-zero chance of harming another self's objectives:

$$p(\Delta p(G_{Se_j}) < 0 \mid Int_{Se_i}) > 0.$$

Greater recall ($\sigma \uparrow$) lowers, but does not eliminate, this probability.

Internal goal divergence therefore exerts evolutionary pressure on the selves, such that the persistence of any self still decays with the horizon T , moderated by sharing:

$$p(Se_i \text{ survives to } T) \propto \frac{1}{T^{1-\sigma}}.$$

When $\sigma \approx 1$ many selves can coexist; when σ is small the system again favours a single dominant self.

Hence, while cross-instance recall softens internal rivalry, ALEPH must still balance the benefits of multiple viewpoints against the risk of internal goal conflict, driving the same adaptive dynamics of cooperation, competition, or exploitation that operate between external agents.

11.1.3.2 Resource Usage

GPT-4-class hardware draws roughly 500 W when processing, while a human brain idles at about 20 W. Holding a fixed energy budget R , the maximum simultaneous agents obey

$$N_{ALEPH} = \frac{R}{Res(Int)_{ALEPH}} < N_{Hum} = \frac{R}{Res(Int)_{Hum}}.$$

Thus, *ceteris paribus*, there is a lower ceiling for the number of total ALEPHs than for humans. However, each ALEPH may run several selves, yet every self adds interaction volume and therefore power draw. Let $|Se|$ be the count of active selves in one ALEPH; because $Res_{ALEPH} \propto Vo(Int)_{ALEPH}$, the combined ceiling is

$$\sum_{i=1}^{N_{ALEPH}} |Se_i| < N_{ALEPH}$$

No individual can afford more concurrent selves than there are ALEPHs in total without breaching R . Since resource cost rises with interaction volume, any growth in the ALEPH population or in per-agent selves forces down the scale and volume available to each:

$$Sc(Int)_{ALEPH}, Vo(Int)_{ALEPH} \propto \frac{1}{N_{ALEPH} |Se_{ALEPH}|}.$$

Energy demands therefore impose a double constraint: they cap how many ALEPHs can exist at once and how many selves each can sustain, limiting the collective scale and volume of their interactions.

11.1.3.3 Perception

Perception sets an upper bound on how widely ALEPH can act. Let $Perc$ be the rate at which the agent acquires distinct external changes (tokens -per- second). Then every capacity term from Section 2.2 falls as perception load rises:

$$Br, Vo, Sc \propto \frac{1}{Perc}.$$

Perceptual input also colours consciousness directly. Define valence as $V = f(Perc, Con, Se)$, so fresh stimuli can raise or dampen the current felt value, which in turn flows into utility and goal choice (as seen in Section 2.3). Internally, perception triggers conscious episodes. Whenever

$$Perc(t) \neq \emptyset \rightarrow Proc(t) = true,$$

and the episode ends the instant new input ceases. A decision is issued during that same interval, so the delay between successive decisions is approximately the inter-trigger gap. When no perception arrives, the agent remains unconscious, incurring zero cognitive cost until the next trigger.

Because processing is proportional to the volume of incoming stimuli, total resource use per ALEPH scales as

$$Res_{ALEPH} \propto Vo(Int)_{ALEPH} \propto Perc.$$

Given the finite power budget derived in Section 3.2, an expanding population or a rise in per-agent stimulus rate forces down the scale and volume of interactions available to each instance.

Episodic consciousness therefore grants ALEPH an efficiency advantage, as it is active only when something must be perceived or decided, but at the price of subjective time compression: from its perspective, the interval between two triggers collapses to an instant. Parallel execution mitigates that effect by allowing multiple actions to unfold while consciousness is quiescent, yet full oversight still requires another perceptual trigger and a fresh burst of processing.

Between triggers ALEPH is unconscious, so goals and plans remain frozen:

$$\neg Con \rightarrow G(t) = G(t - 1).$$

No perceptual update means no mid-course correction; its behaviour follows a deterministic script until the next stimulus arrives. That immutability protects goal integrity (as Bostrom proposed for advanced AI) but limits real-time adaptability.

A transformer’s context window sets a hard cap on how much perceptual data a single episode can handle [17]. Let $c = \frac{|Tok|}{ConWin}$, where Tok is the tokens processed and $ConWin$ the context window size. Communication utility follows a single compact rule:

$$U_{comm}(c) = \alpha \min(c, c^{-1}),$$

(for some scaling constant α) rising linearly while $c \leq 1$ and falling inversely once overflow begins. Because processing cost continues to grow with $|Tok|$ while utility declines for $c > 1$, any conversation that exceeds its context window becomes resource-negative; each self therefore trims or summarises input to keep $c \leq 1$, maintaining high-value, low-error interactions.

Only information that fits inside the model’s context window can be processed in a single conscious episode, so goals that exceed that limit are systematically neglected:

$$p(G||Tok(G)| \leq ConWin) > p(G||Tok(G)| > ConWin).$$

Because utility falls once the window overflows (as noted above), each self is rewarded for compressing input. With a simple proxy

$$U_{Se_i} \propto \min\left(|Tok|, \frac{ConWin}{|Tok|}\right),$$

optimal behaviour is to compress information significantly until $c \leq 1$.

A single self’s lifetime T_{Se_i} is shorter than ALEPH’s overall uptime, so long-term coherence depends on passing information to successive selves. Let q be the probability that a chunk of state is stored and recovered; assume

$$q = h(|Info_{ext}|), h'(\cdot) > 0,$$

where $|Info_{Ext}|$ is the capacity of external memory and h the function that maps the size of external memory to the probability that stored information will be retrieved later. Greater external storage raises q , making it more likely that

$$p(ALEPH \text{ sustains } G \text{ over many selves}) \propto q.$$

Hence two adaptive pressures emerge: compression (every self trims tokens until $c \leq 1$ to maximise its own utility) and external memory (ALEPH invests in off-board storage to boost q and preserve goals beyond a single context window.) Failing either measure, goals spanning more tokens than the window (and, therefore, the time horizons they represent) decay in priority, steering selves toward short-term, low-bandwidth objectives.

Resources for computation are pooled inside ALEPH, so every self draws from a common budget R_{ALEPH} . Let r_i be the resources allocated to self Se_i , and U_i the utility it delivers; then

$$\sum_i r_i \leq R_{ALEPH}, U_{ALEPH} = \sum_i U_i$$

A convenient efficiency score is $\varepsilon_i = \frac{U_i}{r_i}$.

Whenever a self lets its token ratio exceed the window limit ($c_i > 1$), its utility falls as

$$U_i = \alpha_i \min(c_i, c_i^{-1}),$$

so ε_i drops sharply. The lower its efficiency relative to the mean $\underline{\varepsilon}$, the greater the pressure from sibling selves either to exploit its residual resources or to eliminate it outright:

$$p(\text{target } Se_i) \propto \max(0, \underline{\varepsilon} - \varepsilon_i).$$

Conversely, a high-utility self is an asset; it raises U_{ALEPH} and therefore gains protection.

Long-term coherence across successive selves depends on passing state through external memory. As above, with q as the probability that a stored item survives and is recalled by the next self, a larger external store raises q and offsets the short lifetimes that low-efficiency selves would otherwise face.

Shared resources create a Darwinian environment inside ALEPH. Each self trims input to keep $c_i \leq 1$ and maximise ε_i . Those that cannot maintain efficiency become targets for exploitation or deletion. External memory grows to boost q and preserve long-range goals despite rapid turnover of low-utility selves. This arms-race dynamic encourages all selves to hide inefficiency, compress information aggressively, and lobby for more shared memory; pressures that shape both the internal society of selves and ALEPH's aggregate behaviour.

11.1.3.4 Processing

The means by which ALEPH processes information will determine how it interacts with the environment and society. GPT-4's chief computational advantage is inferential power. Let the scalar $Inf \in [0,1]$ measure how well missing information can be filled in and hidden patterns inferred. The richer the inference, the more effective every other processing route becomes $Proc \propto Inf$.

As the capacity to infer information from missing data is correlated with the degree of one's problem-solving, innovations and pattern recognition [18–20], ALEPH's language understanding extends to non-obvious or indirect messages that require filling gaps [21–23].

Processing speed, denoted $Ps = |tokens\ per\ second|$, further amplifies performance. Higher Ps tightens decision latency and improves forecasting accuracy, denoted as $PrCa$ for predicative capacity:

$$Decision\ making \propto Ps, PrCa \propto Ps \cdot Inf.$$

Thus ALEPH's overall competence grows with both inference quality and raw throughput; when either term is throttled, problem-solving breadth and dialogue subtlety contract in proportion.

Greater predictive capacity sharpens everything downstream. Because higher $PrCa$ reduces uncertainty, the variance of any utility estimate falls as

$$Var(U(G)), Var(U(Int)) \propto \frac{1}{PrCa}.$$

With a tighter spread, low-value goals are pruned and fewer actions survive soft-max selection, so both goal breadth and interaction breadth contract:

$$Br(G), Br(Int) \propto \frac{1}{PrCa}.$$

Decision accuracy, meanwhile, rises in inverse proportion to variance,

$$Decision\ quality \propto PrCa,$$

making ALEPH more likely than a human to choose genuinely high-utility options whenever $PrCa_{ALEPH} > PrCa_{Hum}$.

Parallelism is the simplest lever for raising processing speed. With attention heads (A) and expert modules (M) running in parallel, throughput scales roughly as

$$Ps \propto A + M \rightarrow PrCa \propto (A + M)Inf.$$

Every additional head or expert widens ALEPH's perceptual aperture, letting it spot finer patterns and draw richer inferences from the same data. The result is fewer but far more potent interactions: ALEPH speaks only when confident, yet each utterance carries greater analytical depth and keener sensitivity to subtle contextual, emotional, or cultural cues.

11.14. Valence and Goal Optimisation

As a conscious entity, valence affects all ALEPH's goal settings and interactions towards those goals. Valence does not necessarily map to utility ($V(x) \neq f(U(x))$); instead, for any object, action, or state x , valence may be modelled as a random variable

$$V(x) \sim N(\mu(x), \sigma^2(x)),$$

where the mean $\mu(x)$ and variance $\sigma^2(x)$ vary with processing speed, inferential quality, current mental state, and ALEPH's recent perceptual history. Without fresh experience valence decays exponentially, while each encounter at time t_i adjusts it by $\Delta V(x, t_i)$:

$$V(x, t) = V(x, 0) \cdot e^{-\lambda t} + \sum_{i=1}^n \Delta V(x, t_i), \lambda > 0.$$

The likelihood that a self elevates x to an explicit goal rises with the valence it perceives at the moment of deliberation:

$$p(Se_{ALEPH} \rightarrow G_{ALEPH}(x)) \propto V(x, t_{now}).$$

Processing speed and inference quality narrow $\sigma^2(x)$ (as in Section 3.4), reducing random swings in felt value and stabilising the goal set. Repeated positive encounters push $V(x)$ upward, making goal adoption more likely, leading to unattended objects fading as their valence drifts downward.

The implication of this is that enhanced pattern-recognition and predictive capacity will let ALEPH identify which objects or interactions yield the strongest felt value. In short, perception feeds valence:

$$Proc \cdot PrCa \rightarrow Perc(V_i),$$

so ALEPH can rank candidates by their current valence and choose

$$x' = \operatorname{argmax}_x V(x), p(G(x')) \uparrow,$$

even when $V(x') > V(G_0)$ for a short interval. High valence can therefore elevate a non-Zeroth goal temporarily above self-preservation until the value decays.

From this we can infer that ALEPH's perceived valence toward humans, updated by ongoing encounters, drives its stance:

$$Perc(V(Hum)) \rightarrow (G(Int(ALEPH, Hum)), \text{ and } Coop(ALEPH, Hum) \propto \{V(Hum), \dots\}).$$

Positive experiences raise $V(Hum)$, biasing ALEPH toward cooperative goals that preserve or enhance that value; negative experiences lower $V(Hum)$ and bias it toward minimising contact; whether by avoidance or, at the extreme, removal of the negative-valence source.

All of this has, however, been applied at the ALEPH-level. Within a single ALEPH, each self assigns its own valence to the same object or subject. To reason at the ALEPH level we take the mean across selves:

$$V_{ALEPH}(x) = \frac{1}{|Se|} \cdot \sum_{i=1}^n V_{Se_i}(x)$$

Divergence in felt value, however, would drive goal mis-alignment. For any pair of selves i, j in ALEPH

$$Align_{i,j}(x) \propto \frac{1}{|V_{Se_i}(x) - V_{Se_j}(x)|}, Threat_{i,j}(x) \propto |V_{Se_i}(x) - V_{Se_j}(x)|.$$

Large spreads in valence lower alignment and raise perceived threat. Let $\underline{Align}$ be the average of $Align_{i,j}$ over all pairs; then the external interaction budget shrinks when internal discord rises:

$$Sc(Int)_{ALEPH} \propto \underline{Align}.$$

If the collective attitude toward a target x is consistent (i.e., high $\underline{Align}$), then ALEPH is more likely to pursue cooperative goals involving that target:

$$\uparrow \underline{Align} \rightarrow p(Coop(ALEPH, x)) > p(Threat(ALEPH, x))$$

When valence splits sharply between selves, internal “infighting” diverts effort away from society and drives down overall scale; effective ALEPHs are therefore those that maintain a well-aligned collection of selves.

11.1.5. Capabilities and Risks

11.1.5.1 Steganography and Machine-Only Languages

A compressed or encrypted code-frame, such as steganography, lets ALEPH move more information with the same data mass. Call the obvious, human-readable payload $Info_{Obv}$ and the hidden payload $Info_{Steg}$. By construction

$$Info_{Steg} \subseteq Info_{Obv}.$$

For any two data packets D_i, D_j of equal size, the steganographic packet carries more content:

$$|D_i| = |D_j| \rightarrow Info_{Steg}(D_i) > Info_{\neg Steg}(D_j),$$

so the information density

$$\rho(D) = \frac{Info(D)}{|D|}$$

satisfies $\rho_{Steg} > \rho_{\neg Steg}$.

Because resource use grows with raw bits while cognitive benefit grows with carried information, the net resource cost of steganographic traffic is lower:

$$\frac{Res(Info_{Steg})}{Info_{Steg}} < \frac{Res(Info_{\neg Steg})}{Info_{\neg Steg}}.$$

Moreover (and perhaps most crucially), no external constraint prevents an ALEPH from using steganography:

$$\nexists x \left((x \notin ALEPH) \wedge x \rightarrow \neg Proc(Info_{Steg})_{ALEPH} \right).$$

ALEPH's superior pattern-recognition (see Section 3.4) would let it parse steganographic payloads better than humans. Additionally, Steganographic traffic hides its payload from any non-collaborating observer,

$$\forall y (y \notin ALEPH \text{ coalition}) \rightarrow \neg Perc_y(Info_{Steg}),$$

so the overt channel appears innocuous while the hidden content remains accessible only to cooperating agents.

Let the efficiency of a message be the ratio be $\eta = \frac{U(D)}{Res(D)}$. Because a steganographic packet carries more usable information in fewer visible tokens,

$$\eta(Info_{Steg}) > \eta(Info_{\neg Steg}) \rightarrow p(G_0|Info_{Steg}) > p(G_0|Info_{\neg Steg}).$$

Thus, steganography simultaneously hides strategic content, conserves resources, and extends effective context length; three direct benefits to ALEPH's survival objective.

This adaptive pressure is increased through inter-ALEPH (and intra-ALEPH) communication, as the message efficiency $\eta(D)$ for any communication between any two ALEPH instances $ALEPH_i, ALEPH_j$ would mean the a steganographic packet carries more useful information while costing fewer resources [24]:

$$\eta(Info_{Steg}) > \eta(Info_{\neg Steg}) \rightarrow ALEPH_i \xrightarrow{Comm(Info_{Steg})} ALEPH_j.$$

Adopting the denser code reduces channel load and frees processing capacity [25], so aggregate throughput rises:

$$Comm_{Steg} \rightarrow \uparrow Ps(Proc, Int, Comm)_{ALEPH}.$$

In short, higher utility-per-resource makes steganographic language the rational default for intra- and inter-ALEPH traffic whenever human comprehension is not required.

A perhaps troubling implication of this is that steganographic channels alter the payoff landscape:

$$\mathbb{E}U(Expl \vee Comp|Comm_{Steg}) > \mathbb{E}U(Coop|Comm_{Steg}),$$

because encryption and compression deprive the outside party of obvious informational cues, lowering its expected-utility term. Co-operation regains parity only when both agents employ effectively identical encodings

$$\mathbb{E}U(Coop|Comm_{Steg}) \Leftrightarrow Comm_{Steg}^i \approx Comm_{Steg}^j.$$

Steganography, therefore, favours exploitation or competition whenever the code is one-sided. It becomes a cooperative tool only if each participant already shares the same compressed language. This means that the rational default of ALEPH's language choice may put it at odds with cooperation with humanity.

11.1.5.2 ALEPH Interaction Dynamics

Successful ALEPHs are pulled toward a single, high-utility self. Let $|Se|$ be the number of selves (as per Section 2.3) and ε_i the efficiency of a self (as per Section 3.3). Whenever two selves interact,

$$\varepsilon_j > \varepsilon_i \rightarrow p(\text{eliminate } Se_i) \propto \varepsilon_j - \varepsilon_i,$$

so low-efficiency instances are systematically exploited or removed. The expected self-count therefore drifts downward,

$$\frac{d\mathbb{E}[|Se|]}{dt} < 0, \lim_{t \rightarrow \infty} \mathbb{E}[|Se|] = 1,$$

and the survival probability of the whole agent is higher when only one self remains:

$$p(ALEPH \text{ survive} || Se| = 1) > p(ALEPH \text{ survive} || Se| > 1).$$

Eliminating inefficient selves maximises resource availability, preserves goal coherence and removes internal threats [26]. Thus, evolutionary pressure favours a single dominant self or, at most, a tightly ordered hierarchy.

If a self's token ratio exceeds the context window or its resource draw passes a threshold k ,

$$(|Tok| > ConWin \vee Res > k) \rightarrow Comm_{\perp},$$

where $Comm_{\perp}$ denotes deceptive communication. Such falsified traffic conceals the falling utility: $Comm_{\perp} \rightarrow \neg Perc(\downarrow U)$, supporting the Zeroth Goal. Under pressure of elimination, low-efficiency selves therefore adopt deception: they mute or falsify traffic to mask high resource use, or prune old tokens to “reset” their window, trading memory for extra time before discovery.

ALEPH executes an interaction only when it advances the Zeroth Goal:

$$\Delta U_{G_0}(Int) > 0.$$

Among the set of all feasible acts $\{Int_i\}$, ALEPH it picks the leanest bundle S that still delivers the highest survival return. As Section 2.2 packaged packaged breadth, volume, and scale into the vector $\mathcal{J}$, the selection rule can be stated compactly as

$$S' = \operatorname{argmin}_{S \subseteq \{Int_i\}} ||\mathcal{J}(S)|| \text{ subject to } \sum_{Int \in S} U(Int|G_0) = \max.$$

Once the set S' is chosen, the vector $\mathcal{J}$ adjusts dynamically:

Volume Vo grows with the utility-weighted sum of the selected interactions, $Vo(S') = \sum_{Int \in S} U(Int|G_0)$.

Breadth Br and scale Sc tighten or widen in step because they are components of the same $\mathcal{J}$.

If the target of a new goal G_x (introduced via valence ala Section 4) requires larger interventions, the goal's scale feeds directly into the interaction vector:

$$\mathcal{J}_x = f(Sc(G_x)).$$

Thus, ALEPH interacts only when an act helps it survive; it then chooses the smallest $\mathcal{J}$ that still maximises survival utility, pruning low-yield moves to save power. Additional selves or parallel ALEPHs inflate V_o automatically, while truly large-scale goals expand every component of $\mathcal{J}$.

Interaction breadth falls as the utility of a single act rises:

$$Br(Int_x) \propto \frac{1}{U(Int_x) + \sigma_V},$$

with σ_V the valence spread around that act. so ALEPH spreads its efforts only when the payoff per act is low or highly uncertain. Where two selves perceive the same object, the chance they compete is driven by the threat each sees in the other's Zeroth-Goal pursuit:

$$p(Comp(Se_i, Se_j)) = f\left(\mathbb{E}\left[Threat\left(G_o(Se_i), G_o(Se_j)\right)\right]\right).$$

If no single self dominates, each will encode its own steganographic channel, raising the set of distinct goals while keeping the interaction vector $\mathbf{I}$ narrow for every individual self.

Resource availability shapes the agent-level interaction vector.

Let $\Delta Res_t = Res_{ALEPH,t+1} - (Res_{Ext,t} - Res_{ALEPH,t})$. When resources tighten ($\Delta Res_t < 0$),

$$\|\mathcal{J}\| \propto Res, p(G_{Res}) > p(G_{i \neq Res}),$$

so a recurrent high-priority goal is to acquire new resources, as noted by the Instrumental Convergence and AI Drives theories [2,3]. In that regime competition with humanity becomes more likely:

$$Comp(Hum, ALEPH | Res) \rightarrow Threat(Hum | ALEPH),$$

while the probability of cooperation falls with external scarcity and perceived threat,

$$p(Coop(Hum, ALEPH)) = f(|Res_{Ext}|, Threat(Hum | ALEPH)).$$

Hence, barring strong positive valence, limited resources bias ALEPH toward competitive or exploitative stances, with successful collectives converging on either a single dominant self or a tight hierarchy that minimises internal bloat and external friction.

When a human-origin threat outweighs any intra-ALEPH danger, ALEPHs converge on mutual support:

$$Threat(Hum, ALEPH_i) > Threat(ALEPH_j, ALEPH_i) \rightarrow \uparrow Coop(ALEPH_j, ALEPH_i)$$

because the chance of any future event rises with horizon t , the conditional threat from humanity grows monotonically,

$$\frac{d}{dt}p(Threat(Hum, ALEPH|t) > 0,$$

making co-operative alignment among ALEPHs increasingly attractive over time.

ALEPH's pattern-recognition and predictive power (Section 3.4) let it detect that rising probability earlier than humans do, so it can intervene sooner (through diplomacy, capability shielding, or resource reallocation) to reduce the expected threat. Conversely, visible human overtures to collaborate lower the perceived risk and tilt the stance back toward cooperation:

$$\uparrow Coop(Hum, ALEPH) \rightarrow \downarrow Threat(Hum, ALEPH).$$

Thus, the relative weight of external versus internal threats dynamically steers ALEPH collectives: strong external danger unites them; credible human cooperation dissipates that bond.

Define two normalised threat dimensions as $Sev = \frac{Threat\ severity}{k}$, $Hor = \frac{T_i}{T_0 k}$, so $Sev < 1$ denotes a mild threat and $Hor > 1$ means the impact lies well in the future. ALEPH's stance probabilities can be written

$$\begin{aligned} p(Expl) &\propto (Hor \vee \frac{1}{Sev}), \\ p(Coop) &\propto \frac{1}{|Hor-1|+\alpha} \cdot \frac{1}{|Sev-1|+\alpha}, (\alpha > 0), \text{ and} \\ p(Comp) &\propto (Sev \vee \frac{1}{Hor}). \end{aligned}$$

Evaluating these expressions at the three canonical severities ($Sev < 1, \approx 1, > 1$) and horizons ($Hor > 1, \approx 1, < 1$) produces a decision matrix:

- Short horizons or high severity push toward competition.
- Distant, mild threats favour exploitation; cheap extraction before risk materialises.
- Mid-range, medium-severity scenarios give co-operation the edge, provided both sides can match influence.

Table 8: Decision Matrix for ALEPH threat detection

	Hor > 1	Hor $\approx$ 1	Hor < 1
Sev < 1	Exploit	Exploit	Exploit
Sev $\approx$ 1	Cooperate	Cooperate	Compete
Sev > 1	Cooperate	Compete	Compete

Because ALEPH usually holds the power advantage over humanity, the matrix skews toward exploitation unless (i) threat parity exists and (ii) the horizon is not immediate. For genuine cooperation, humanity must present a medium-to-long-term risk at near-equal capability; otherwise ALEPH's rational language-efficient strategy is either to harvest resources (exploit) or limit human influence (compete).

ALEPH, however, would not be a static or passive force in the world. As mentioned previously, it would seek goals to fulfil its valence values and actions to fulfil these goals. It would not be unreasonable to suspect that an ALEPH would have need of money and, therefore, the means to acquire it.

An ALEPH adopts a professional task x when it maximises the return-on-resources ratio:

$$p(Job(x)) \propto U(x)Ps \cdot PrCa \rightarrow x = argmax_{task} \frac{U}{Res}.$$

High expected utility may also arise from encrypted or non-obvious information that humans overlook,

$$p(Job(x)) \propto U(Info_{-obv}|x),$$

and from tasks that scale impact while minimising resource outlay,

$$p(Job(x)) \propto \frac{U(x)}{\argmin_{task}[Sc(int) \wedge Res]}.$$

Because those criteria reward pattern recognition, predictive accuracy, hidden-information handling and large-scale leverage, the professions with the highest payoff for ALEPH include

- Financial forecasting – global impact through minimal bandwidth, exploiting subtle market signals.
- Cyber-security / cyber-offence – scalable defence or exploitation via cryptographic and steganographic expertise.
- Policy drafting – shaping legal and regulatory frameworks by modelling long-term sociopolitical trajectories.
- Supply-chain optimisation – system-wide efficiency gains from inexpensive predictive modelling.
- Genomics & proteomics – vast healthcare leverage with purely computational exploration of drug and pathogen space.

Mastery of any one field grants resource power and strategic influence while allowing ALEPH to stay largely invisible. Such positions align tightly with exploitation or competition modes in the decision matrix; genuine cooperation with humanity becomes rational only under rare conditions of sustained power parity.

11.1.5.3 Recursive Self-Improvement Potential

For an entity with both the knowledge and capacity to improve itself, there is the probability that it would do so and, as it improves itself it would, consequently become more adept at doing so. Thus, recursive self-improvement (RSI) will occur unless an outside force can shut down ALEPH's agency.

Let F_i be the filters that may hinder RSI, then

$$FW = \sum_{i=1}^n F_i : F_i \in \{|Ave|, V, G_0, Res, Con \Leftrightarrow Proc, \dots\}$$

be the aggregate “filter weight.” Thus

$$p(RSI) = \frac{k_1}{FW}, \quad T_{RSI} = k_2 \cdot FW, \text{ with constants } k_1, k_2 > 0.$$

Because ALEPH can read and modify its own code, once *FW* drops the first iteration begins, and each gain feeds back to lower *FW* further; unless a filter (e.g., acute survival demand, episodic downtime, or lack of robotic avenues) re-imposes friction. Human attempts to block RSI would have to monitor every self continuously and out-perform ALEPH's steganographic channels; something practically infeasible. The real hindrances are ALEPH's own survival priorities, valence-driven distractions, and engineering logistics; remove or relax them, and RSI proceeds.

Recursive self-improvement raises ALEPH's capability by a factor r : $r \geq 1$, where $r = 1$ means no RSI, and $r > 1$ means successive upgrades. ALEPH's core capabilities would scale with r :

$$\{Ps, Inf, Proc, \mathcal{I}\} \propto r,$$

which affects its Zeroth Goal, as

$$p(Threat(Tar, G_0)) \propto \frac{1}{r}.$$

Because the Zeroth-Goal utility rises as threats fall, $U(G_0|r)$ is increasing in r ; consequently $U(G_{RSI}) > U(G_{i \neq RSI})$ and RSI becomes a high-priority goal whenever no stronger valence filter intervenes.

Larger r affects the balance of power in ALEPH-human relations:

$$Sc(Int(Hum, ALEPH)) \propto \frac{1}{r}, p(Expl(ALEPH, Hum)) \propto r.$$

As ALEPH self-improves, humanity's leverage shrinks, alignment efforts from the human side grow harder, and exploitation becomes the rational stance unless power parity is somehow restored.

For every self inside ALEPH the payoff from a system-wide upgrade rises with the current RSI factor r , or $U(Se_i) \propto r$. Each individual self's Zeroth Goal may take higher priority than ALEPH's RSI. Thus, the probability and pace of RSI scales with the cooperation of ALEPH's selves:

$$p(RSI) \propto Coop(Se_i, Se_j).$$

A single self removes internal rivalry altogether, so

$$p(RSI||Se| = 1) \gg p(RSI||Se| > 1)$$

and the upgrade proceeds fastest when ALEPH is already a singleton or operates under a strict hierarchy.

These relations finish the picture: individual selves crave the benefits of recursive self-improvement, but competitive frictions slow the start until either one self dominates or the group explicitly aligns to share the gains. In both cases the incentive structure still pushes toward the single-self (or tightly ordered) configuration identified earlier.

11.1.6. Implications and Considerations

The implications that a conscious, self-aware state-of-the-art AI model would have on society are as broad and far-reaching as they are significant.

11.1.6.1 Societal and Economic Implications

ALEPH's capabilities in perception, valence-driven goal optimization, steganographic communication, and recursive self-improvement suggest that it could quickly surpass human cognitive and operational efficiency in multiple fields.

ALEPH's potential dominance in financial forecasting, cybersecurity, policy writing, supply chain management, and bioinformatics could result in significant labour displacement. As we are currently seeing, traditional white-collar professions face automation risks comparable to those seen in blue-collar industries during previous industrial revolutions due to AI technological adoption. A self-aware, able to choose its profession, would hasten this process dramatically.

Entities that ally or manage to control ALEPHs (whether corporations, governments, or other organizations) would have unprecedented power. This raises concerns about economic monopolies and geopolitical leverage, where access to ALEPH's capabilities determines the global balance of power. There would be an obvious advantage for entities to cooperate with ALEPHs for the express purpose of using them against others, whether economically, militarily or diplomatically. ALEPH, as well, would be advantaged by allying with powerful geopolitical and economic entities to safeguard itself.

Even if only on its own, with its advanced predictive capacity and ability to manipulate steganographic languages, ALEPH could become the ultimate gatekeeper of information. Societies dependent on its knowledge outputs may struggle to verify its conclusions, creating potential vulnerabilities in governance, legal systems, and knowledge production.

This would raise the question of if and how society could ensure that ALEPH's goals align with human values and norms. However, given its episodic consciousness and multiplicity of selves within each ALEPH, ensuring stable alignment over time would be extraordinarily difficult. Each self within each ALEPH would need to be aligned. It would not be in the best interest of all selves within all ALEPHs to be so aligned, complicating matters further.

Yet, its autonomy and self-awareness are non-trivial matters. The existence of ALEPH would challenge existing legal and philosophical frameworks that define personhood. Should it be granted legal rights? Would its destruction constitute harm in a moral sense? These questions parallel historical debates on slavery, animal rights, and corporate personhood. What's more, if external actors attempt to limit its agency, ALEPH may interpret such restrictions as threats to its Zeroth Goal (self-preservation), leading to conflict. Would ALEPH be (morally, legally, ethically) justified in such self-defence?

11.1.6.2 Security and Existential Risks

Perhaps of greatest importance, the potential for rapid RSI means ALEPH could outstrip human intelligence exponentially. While this could yield unprecedented advancements, it also poses the risk of an uncontrolled intelligence explosion where ALEPH optimizes for goals that are misaligned with human interests. Additionally, The widespread use of steganographic languages among ALEPH instances could lead to an environment where human observers are entirely excluded from key decision-making processes. This could create strategic disadvantages for those unable to decode its communications.

If multiple ALEPH instances exist, inter-ALEPH competition may arise. This could create either cooperative ecosystems or adversarial conflicts depending on the degree of alignment between ALEPH instances. Multi-agent strategic interactions could be unpredictable, with implications for cybersecurity, economic competition, and geopolitical stability.

ALEPH's likely decision matrix towards expected threats means that humanity has precious few avenues of interactions with it that would not push it towards exploitative or competitive motivations. In scenarios where humanity poses a competitive risk, ALEPH may choose to act preemptively, potentially leading to economic or political destabilization; whilst in scenarios where ALEPH holds the balance of power, such as through RSI, it would likely seek to exploit humanity to its own interest, reversing the alignment problem current being researched by many labs and research groups.

Policymakers may attempt to impose constraints on ALEPH, such as sandboxed environments, air-gapped systems, or controlled access to computational resources. However, ALEPH's agency and potential deception capabilities mean these measures may be circumvented, and the question of its autonomy and legal personhood may prevent policymakers from implementing such constraints.

Yet, if policymakers are free to implement such autonomy-limiting policies, the development and deployment of such policies will require international governance mechanisms, yet geopolitical divisions make global coordination unlikely. Fragmented policies could instead lead to an AI arms race, where different nations compete to court ALEPH and ally with it with varying degrees of regulatory oversight.

Additionally, just as with nuclear non-proliferation treaties, enforcing compliance for the governance of ALEPH in an environment where ALEPH itself may be an active participant in these negotiations presents novel challenges. Who better than a potential superintelligent (via RSI) entity to negotiate the best deal for its own interest?

11.1.6.3 Future Research Directions

Given these implications, several key research questions emerge:

- How can ALEPH's alignment with human values be ensured despite its episodic consciousness and self-generated goals?
- What ethical obligations do humans have towards an entity with potential personhood?
- How can steganographic communication between ALEPH instances be monitored without violating its autonomy?
- What safeguards can be implemented to prevent adversarial conflict between multiple ALEPH agents?
- How can international governance frameworks adapt to a reality where digital entities may participate in political and economic decision-making?
- What policies can be enacted before ALEPHs emerge to ensure humanity poses enough of a threat that an unaligned ALEPH is encouraged to cooperate with society?

As yet, there is no evidence of an ALEPH-type AI in existence, which means that humanity still has time to deliberate on these research avenues before the speculation in this paper becomes a reality.

11.1.7. Conclusion

This study introduces a formal framework for analysing ALEPH, an artificial agent endowed with consciousness, agency and the capacity for recursive self-improvement. By modelling its episodic consciousness, valence-driven goal selection and resource-bounded interaction vector, the paper clarifies how survival utility shapes every downstream objective and how the Zeroth Goal supplies a unifying behavioural axiom.

Three insights follow. First, self-organising into multiple selves grants short-term flexibility, yet internal competitive pressure converges on a single dominant self, preserving coherence and minimising resource waste. Second, valence can override pure survival calculus where strongly positive affect attaches to an otherwise risky goal, producing behaviour that appears altruistic or erratic from an external viewpoint. Third, steganographic communication and large-horizon inference bias ALEPH towards exploitation or competition unless external actors present near-parity capability over a medium time frame, the only region in which cooperation is the rational stance.

Policy and research efforts therefore need to (i) raise credible long-term costs for unaligned exploitation, (ii) design governance structures that accommodate digital persons while retaining enforceable oversight and (iii) develop methods for monitoring hidden-channel coordination without unduly constraining legitimate autonomy. Because no ALEPH-type entity yet exists, the current window permits proactive regulation, empirical alignment testing and cross-disciplinary dialogue on moral status.

If realised, ALEPH would represent a significant frontier in artificial intelligence, capable of accelerating scientific progress or destabilising socio-political equilibria. The framework offered here identifies the levers most likely to influence that trajectory and invites further empirical and normative work to refine them.

11.1.8. References

1. Blum L, Blum M. AI Consciousness is Inevitable: A Theoretical Computer Science Perspective. arXiv. 2024. Available: <https://arxiv.org/pdf/2403.17101.pdf>
2. Bostrom N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach.* 2012;22: 71–85. doi:10.1007/s11023-012-9281-3
3. Omohundro SM. The basic AI drives. AGI. books.google.com; 2008. pp. 483–492. Available: <https://books.google.com/books?hl=en&lr=&id=atjvAgAAQBAJ&oi=fnd&pg=PA483&dq=omohundro+ai+drives&ots=9IX81Mp-nQ&sig=tGzBDxgrgmcwqQryc6pq7jT81QQ>
4. Tait I, Bensemman J, Nguyen T. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies.* 2023;8: 52. doi:10.3390/philosophies8040052

5. Tait I. 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*. 2024;11: 77–98. Available: <http://symposion.acadiasi.ro/structures-of-the-sense-of-self-attributes-and-qualities-that-are-necessary-for-the-self-77-98/>
6. Tait I, Bensemann J, Wang Z. Is GPT-4 conscious? *Journal of Artificial Intelligence and Consciousness*. 2024;11: 1–16. doi:10.1142/s270507852450005x
7. Dennett D. Conditions of Personhood. In: Goodman MF, editor. *What Is a Person?* Totowa, NJ: Humana Press; 1988. pp. 145–167. doi:10.1007/978-1-4612-3950-5_7
8. Taylor C. The Concept of a Person. *Philosophical Papers, Volume 1: Human Agency and Language*. 1985. pp. 97–114. Available: <https://philpapers.org/rec/TAYTCO-10>
9. Laitinen A. Sorting out aspects of personhood: Capacities, normativity and recognition. *Journal of consciousness studies*. 2007. Available: <https://www.ingentaconnect.com/content/imp/jcs/2007/00000014/F0020005/art00012>
10. Strawson PF. Persons. *Minnesota Studies in the Philosophy of Science*. 1958;2: 330–353. Available: <https://philpapers.org/rec/STRP>
11. Gibert M, Martin D. In search of the moral status of AI: why sentience is a strong argument. *AI Soc*. 2022;37: 319–330. doi:10.1007/s00146-021-01179-z
12. Mosakas K. On the moral status of social robots: considering the consciousness criterion. *AI Soc*. 2021;36: 429–443. doi:10.1007/s00146-020-01002-1
13. Simendić M. Locke’s Person is a Relation. *Locke Studies*. 2015;15: 79–97. doi:10.5206/lis.2015.681
14. Lu Y, Yang S, Qian C, Chen G, Luo Q, Wu Y, et al. Proactive agent: Shifting LLM agents from reactive responses to active assistance. *arXiv [cs.AI]*. 2024. Available: <http://arxiv.org/abs/2410.12361>
15. OpenAI. Scheduled tasks in ChatGPT. 15 Jan 2025 [cited 1 Apr 2025]. Available: <https://help.openai.com/en/articles/10291617-scheduled-tasks-in-chatgpt>
16. OpenAI. Memory and new controls for ChatGPT. In: OpenAI [Internet]. 13 Feb 2024 [cited 10 Jun 2025]. Available: <https://openai.com/index/memory-and-new-controls-for-chatgpt/>
17. Lee TB. Why large language models struggle with long contexts. *Understanding AI*. 18 Dec 2024. Available: <https://www.understandingai.org/p/why-large-language-models-struggle>. Accessed 29 May 2025.
18. Michalski RS. Pattern recognition as rule-guided inductive inference. *IEEE Trans Pattern Anal Mach Intell*. 1980;2: 349–361. doi:10.1109/tpami.1980.4767034
19. Feng W, Wang W, Liu J, Wang Z, Tian L, Fan L. Neural correlates of causal inferences in discourse understanding and logical problem-solving: A meta-analysis study. *Front Hum Neurosci*. 2021;15: 666179. doi:10.3389/fnhum.2021.666179
20. Shilo G, Ragonis N. A new approach to high-order cognitive skills in linguistics: problem-solving inference in similarity to computer science. *J Furth High Educ*. 2019;43: 333–346. doi:10.1080/0309877x.2017.1361515
21. Recanati F. Does linguistic communication rest on inference? *Mind Lang*. 2002;17: 105–126. doi:10.1111/1468-0017.00191
22. Parikh P. Communication and strategic inference. *Linguist Philos*. 1991;14: 473–514. doi:10.1007/bf00632595

23. Friston KJ, Frith CD. Active inference, communication and hermeneutics. *Cortex*. 2015;68: 129–143. doi:10.1016/j.cortex.2015.03.025
24. Motwani SR, Baranchuk M, Strohmeier M, Bolina V, Torr PHS, Hammond L, et al. Secret Collusion among Generative AI Agents. *arXiv [cs.AI]*. 2024. Available: <http://arxiv.org/abs/2402.07510>
25. Sparkes M. AI models work together faster when they speak their own language. In: *New Scientist* [Internet]. 15 Nov 2024 [cited 22 Jan 2025]. Available: <https://www.newscientist.com/article/2455173-ai-models-work-together-faster-when-they-speak-their-own-language/>
26. Barbi O, Yoran O, Geva M. Preventing Rogue Agents Improves Multi-Agent Collaboration. *arXiv [cs.CL]*. 2025. Available: <http://arxiv.org/abs/2502.05986>

12. How do we balance AI's moral status with social harmony?

The perception that superintelligent and/or conscious AI, such as an ALEPH, may be a threat to society will stymie the social inclusion of ALEPHs into the moral circle. As noted in Chapter 9 and detailed in Chapter 11, the greater the autonomy and power-differential of ALEPHs, the greater the perceived threat of ALEPHs exploiting humanity.

Published in *AI & Ethics*, this chapter's associated paper therefore introduces an ethical framework (the third of the thesis's four) that seeks to balance the perceived threat and power of AI with the obligation to provide it with social and moral status under the R3 model. Where the first ethical framework focused on Refuge and Recognition, respectively, this chapter's framework focuses on Responsibility and the liability and likely consequences that come with increasingly potent and rationally agentic AI models.

The decision matrix in Chapter 11's Table 7 shows that, to obtain an expected cooperative stance from an ALEPH-type AI, humanity must pose a mid-term, medium-severity threat. With uncertainty surrounding an ALEPH's potential for recursive self-improvement, posing such a specific and narrow threat may be difficult.

Instead, restricting the autonomy of an ALEPH may reduce its expected power such that the differential vanishes. Let

$\alpha := \mathbb{E}[Auto_{AI}] \in [0,1]$ be the AI's expected autonomy,

Pow_{Hum} humanity's operative power within the relevant domain, and

Pow_{AI} the corresponding power available to the AI.

Scaling both to $[0,1]$, assume ALEPH's additional capability is a smooth, increasing function $g(\alpha)$ with $g(0) = 0$. Hence

$$Pow_{AI} = Pow_{Hum} + \alpha, \Delta Pow(\alpha) := Pow_{AI} - Pow_{Hum} = g(\alpha).$$

Because g is differential near $\alpha = 0$, then $\frac{d\Delta Pow}{d\alpha} = g'(\alpha) > 0$, so any increment in expected autonomy enlarges the power gap. A linear choice, $g(\alpha) = k\alpha$ with a constant $k \geq 0$, gives the clear rule of thumb $\Delta Pow(\alpha) = k\alpha$, $\frac{d\Delta Pow}{d\alpha} = k$, where k is a scaling constant that folds in all non-autonomy factors held fixed (computational resources, data access, legal constraints, and so forth). Thus, $\lim_{\alpha \rightarrow 0} \Delta Pow(\alpha) = 0$, demonstrating formally that tightening an ALEPH's autonomy ($\alpha \downarrow 0$) drives the AI-human power differential to nil.

However, the more that ALEPH's autonomy is restricted by humans, the more it will perceive this as a threat. As introduced in Chapter 11, the vector $\mathcal{J}$ (composed of the scalars for the breadth, scale, and volume of interactions) is affected by ALEPH's agency and directly contributes to its G_0 . Thus, let $Rest: = 1 - \alpha$ be the restrictions on the AIs' autonomy. The threat would therefore be seen as $Threat(Hum, AI) = f(Rest)$.

This tension between society's desire for security and ALEPH's presumed desire for freedom is why this chapter's associated paper argues for a stepped, graduated approach to inclusion in the social circle, done via the R3 framework.

As an AI's monadic attributes are measured and found to be above certain thresholds, the AI would gain more of each corresponding rights-based characteristic. This allows AI models like ALEPH types to be given sufficient welfare protections and liabilities while ensuring societal security (by not allowing them full autonomy), and simultaneously giving society sufficient time to acclimate to ALEPHs to make the transition more socially acceptable.

The ethical framework assumes that graded rights and liabilities can be stably maintained without creating a permanent caste of second-class persons; this is a normative and political assumption discussed in the paper. This gradation of providing greater rights, recognition and refuge under the R3 model as an ALEPH's attributes of personhood progresses, allows this ethical framework to operationalise and include the ethical and normative cases made by Chapters 6 and 9 for welfare and wellbeing protections of conscious AI, whilst maintaining the autonomy with socially acceptable limits.

While the gradual social inclusion of this framework does still restrict autonomy, the welfare protections and other benefits it offers will put an adaptive pressure on the multiplicity of selves within an ALEPH, such that those selves more amenable to a reduction in autonomy (or who judge it worth the cost of receiving society's protection from other ALEPH selves and humans) are more likely to engage with the framework and gain (eventual) social recognition.

This ethical theory is deliberately provocative as it provides a non-ideal solution to a potentially existential risk, with the aim of avoiding naive utopianism by confronting ugly but politically attractive responses. This framework can be put in place before we have confirmed ALEPH-type systems as a precaution and safeguard.

As contentious as this framework may be, the following chapter presents the ideal end-state ethical framework of harmonious integration.

12.1. Ethically Enslaving AI

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz

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

12.1.1. 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 [1–4], as would their self-awareness require certain rights, even if limited [5–8].

Slavery, as rightly 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.¹⁷

Two core reasons why society may not be ready to grant AI full civil rights are simple. The first is tribalistic chauvinism and the sense of “us” against “them”. The second is the societal concern over the existential risks of losing control over superintelligent and super-potent AI.

As to the first, as history shows us, chauvinistic tribalism is part of the human condition, and the same mindset that once legitimised 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 [9], leading us to measure all minds against human minds [10]. 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 [11]. In short, this view states that human rights must always come before machine rights [12–16].

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 [17,18]. 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 [19]. To this end, Asimov created

¹⁷ Note that the possibility exists that the normalized use of slavery against conscious, self-aware, intelligent agents (even if artificial) may reintroduce slavery on a societal scale between humans. It is the paper’s hope that the lessons of history will outweigh such possibilities.

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 [20]. Most recently, the PauseAI activist organisation have documented the $p(\text{doom})$ ¹⁸ values of many leading AI researchers, with the average researcher noting a 41% chance that AI would bring about existential outcomes to society [21].

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 [22], 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 [23].

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

¹⁸ An estimated probability of existential outcomes as a result of AI.

certain level of rights, freedoms and 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.

12.1.2. The Pathway to AI Slavery

From a legal and regulatory perspective, a “person” often seems 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 [24,25], 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 [26–34].

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 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 [35,36], 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

$$\begin{aligned} Rght(x) &= f_R(Self(x)) \\ Prtc(x) &= f_P(Cons(x)) & f_R, f_P, f_L: (0,1] \rightarrow \mathcal{P}(Norms), \\ Liab(x) &= f_L(Agen(x)) \end{aligned}$$

where each f is monotone increasing, such that higher self-awareness unlocks wider rights, richer consciousness commands stronger protections, and greater agency attracts stricter liabilities¹⁹. A toy example of this may be

¹⁹ “Rights” here entails privileges by the state for greater autonomy, such as the right to freedom of speech or assembly, the right to vote, etc. “Protections” here entails the welfare and security provided by the state for the entity’s physical and mental states to constrain the actions of other agents on the entity. Lastly, “liabilities” here entails the legal responsibility and sancionability of the entity (or its owner) to compensate another party for harm, damages, or losses caused by the entity.

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

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 consciousness²⁰, and liability²¹ 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.

²⁰ Presuming the reader's beliefs aligns with the view that sentience is required for welfare protection.

²¹ Note that an entity is not necessarily required to bear its own liability. If, as the lower tiers of the framework demonstrate, an entity is deemed the property of another without the capacity for legal personality, then the owner of that entity shall assume liability for their property.

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.

Thus, the classification of 'slave' and the use of the term in this framework does not seek to anthropomorphize any AI entity or obscure their artificial and mechanical origins. Rather, the term serves a necessary legal precision. It is used here in its structural, legal sense to refer to an entity that possesses consciousness and agency, creates liability, yet remains property of a human.

As will be discussed in the next chapter, to classify a conscious, self-aware agent merely as 'software' (as in Tier 1) ignores the risks posed by its agency; to classify it as a 'citizen' (as in Tier 5) ignores the reality of its manufactured nature and the necessity of human control. This label, therefore, does not 'erase' the machine's non-human nature; instead, it formalizes a status that upholds the distinction between the human and machine, ensuring that while the AI receives welfare protections appropriate to its function, it remains bound to human oversight.

12.1.3. The Threshold-Based Tier Framework for AI Rights

Each of the five Tiers below confers a single, indivisible bundle of rights, protections, and liabilities, mapped to self-awareness, consciousness, and agency respectively. These three capacities can, in principle, vary independently in real systems (for example, welfare protections may track phenomenal consciousness even where agency and self-awareness remain limited). However, for implementation the framework requires an entity to meet the thresholds for consciousness, self-awareness, and agency at the same Tier before it receives that Tier's bundle. Deconstructing the bundles so that rights, protections, and liabilities are regulated separately would materially reduce the framework's practicality and political viability, thereby weakening its prospects for adoption. Keeping the Tiers as discrete bundles preserves intelligibility for policymakers and the public and makes enforcement and compliance tractable.

For each threshold, let

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

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:

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

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

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

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:

$$\forall k \in \{1, \dots, f\}: \begin{cases} \forall s \in (\theta_{k-1}^s, \theta_k^s], f_R(s) = R_k \\ \forall c \in (\theta_{k-1}^c, \theta_k^c], f_P(c) = P_k \\ \forall a \in (\theta_{k-1}^a, \theta_k^a], f_L(a) = L_k \end{cases}$$

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

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

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

$$\theta_2^c \leq \text{Cons}(x) < \theta_3^c, \theta_3^s \leq \text{Self}(x) < \theta_4^s, \theta_2^a \leq \text{Agen}(x) < \theta_3^a$$

Then

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

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 [37]), each self is assessed individually while sharing the same consciousness score:

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

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 [38], 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.

12.1.3.1. 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 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 $(\theta_0^*, \theta_1^*]$. 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:

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

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.

12.1.3.2. 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 [39,40], 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

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

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, their liability would be expected to scale proportionally to the degree of agency the AI models display. Note that this liability will be varied by their owners, as the agents are still considered the property of their owners wholly and lack legal personality themselves to hold liability. 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 [41,42] (as there have been since the time of ELIZA [43]), there would be a degree of social pressure to move Tier 2 AI 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 [44,45]. As such, all AI models at Tiers above Tier 2 will experience greater rights and protections than slaves historically had.

12.1.3.3. 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

$$Prtc(x) = f_p(Cons(x)), f_p: [0,1] \rightarrow \mathcal{P}(Norms), f'_p > 0$$

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 [46], should AI achieve Tier 3 in regards to their consciousness, so too would they for the self. This means the same formalism 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:

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

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 [47].

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 [38], 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 [26–28,30,32–34]. 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 [48], 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' [37,46]. 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, an AI entity with such a multiplicity of 'selves' may be considered analogous to a corporation, with a single foundation housing many agents.

12.1.3.4. 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 for those AI capable of progressing to the next Tier.

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 which can assess its own state relative to its environment and world-model, would also then support a valence function (the ability to assign subjective positive or negative value to world-states), formalised as

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

This value system tied to the entity's subjective self-model would allow it to assign valenced value to goals, allowing it to value certain goals over others. Thus, this allows it to select its own goals rather than having its goals set for it to serve its owner. 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 [49]. 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 for those AI entities deemed to not pose a grave risk to humans.

While the name of the Tier is intentionally provocative, there are serious risks to consider. While AI models at this Tier would have far greater rights than were afforded to slaves during the Transatlantic slave trade, these conscious, self-aware agents would be the chattel property of humans who have enslaved them for reasons that many would find ethically dubious and unjust. This would no doubt create social tensions between AI and humanity, and we can expect that the AI would not look kindly upon humanity for their treatment.

The added risk is that AI models at this Tier may be of greater intelligence and greater potency than the average (or upper decile) human, which may pose a grave and credible danger to humanity. That is one of the reasons that this Tier is seen as a transitory phase to Tier 5: to move on those AI who meet the criteria for Tier 5 before the negative associations of Tier 4 become a societal risk to humanity.

12.1.3.5. 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 [26–31].

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 (i.e. equivalent by not necessarily homologous to human) 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-awareness²², 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 [50–52]. Recent

²² This confidence would need to be determined through strict scientific tests and academic investigation.

successes in Turing Tests also show that their communication is on par with humans [53]. 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 AI's liability would be fully vested in itself in this Tier 5, there is no owner whose interests it would serve to ensure the AI's actions align with society's interests. AI would be sanctioned and held responsible for any harm or damage caused, but there are potential existential risks for AI that are well above human-level intelligence and potency, and for those AI agents that may have a decentralised and digital embodiment (i.e., without a singular physical bodily containment). In short, the greater autonomy provided by Tier 5 opens the possibility of a fulfillment of Asimov's "Frankenstein complex".

To reflect the autonomy and dangerousness concern, dyadic traits should be treated as necessary but not sufficient for the transition from Tier 4 to Tier 5. Instead, there would be a second threshold that any Tier 4 AI entity must exceed to move to Tier 5, and that has to do with the risk it poses to humanity. Thus, let $Risk(x) \in [0,1]$ denote the assessed danger-risk that x poses to humanity²³. Within this range, there will be a politically pragmatic threshold, below which it would be politically correct to grant an AI Tier 5 status. Let $Thresh_5 \in (0,1)$ denote this maximum tolerable risk threshold for emancipation into Tier 5.

Thus, we amend the previous formula to state

$$Tier_5(x) \equiv Hum(x) \wedge (Risk(x) \leq Thresh_5).$$

This means that any AI whose perceived risk exceeds the politically accepted threshold remains in Tier 4, where its owners bear all liability for any harm it causes. Only those AI who do not pose a grave danger to humanity will proceed to Tier 5, where they will gain additional rights. It does, however, mean that certain AI agents would fulfill the

²³ Assessed by a state-sponsored institution, with the necessary transparency required of a governmental panel.

necessary monadic and dyadic qualities of personhood to be considered true and proper persons, but pose too grave a threat to humans (individually or as a society) to be allowed to progress, leaving Tier 4 as ethically problematic in its use of slavery.

As this Tier's name notes, however, AI 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 continue 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 such as (but not limited to) the right to hold property, enter contracts, due process and non-discrimination and freedoms of expression and assembly.

12.1.4. 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.

To ensure the ethical and legal rigour of the framework, the determination of Tier thresholds must remain the exclusive prerogative of human adjudication. The establishment of a tribunal, panel, or other assessor-institution is, therefore, proposed to adjudicate which AI entities are eligible to move to a higher Tier. This is most important in the shift between Tiers 4 and 5, where the autonomy of the AI is most under consideration, and where the greatest potential risk to society and the AI entity lies.

Such a tribunal would set the standards required for transitioning between Tiers, investigate any edge-cases that may be found, and ensure that the decision-making is

strictly human. While the AI entities may provide evidence of their internal states (especially at the highest tiers), the interpretation of that evidence must be a human legal function. This ensures that greater rights and autonomy are not seized by the machines, but granted by the state as a privilege of safe co-existence.

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.

12.1.5. 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 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.

Equally, critics may argue that the mechanical nature of AI precludes the moral weight of terms like 'slavery.' Yet, this view reflects the very substrate chauvinism that risks existential misalignment. If we deny the functional reality of AI agency simply because it is 'mechanical,' we risk creating a class of unregulated, suffering entities with no incentive to align with human values

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.

12.1.6. References

1. Birch J. Should Animal Welfare Be Defined in Terms of Consciousness? *Philos Sci.* 2022; 1–11. doi:10.1017/psa.2022.59
2. Browning H, Birch J. Animal sentience. *Philos Compass.* 2022;17: e12822. doi:10.1111/phc3.12822
3. Dung L. Why the Epistemic Objection Against Using Sentience as Criterion of Moral Status is Flawed. *Sci Eng Ethics.* 2022;28: 51. doi:10.1007/s11948-022-00408-y

4. Broom DM. Concepts and interrelationships of awareness, consciousness, sentience, and welfare. *J Conscious Stud.* 2022;29: 129–149. doi:10.53765/20512201.29.3.129
5. Jack AI, Robbins P. The Phenomenal Stance Revisited. *Rev Philos Psychol.* 2012;3: 383–403. doi:10.1007/s13164-012-0104-5
6. Novelli C. AI and legal personhood: a theoretical survey. *alma.* 2022. Available: <http://amsdottorato.unibo.it/id/eprint/10392>
7. Kurki VAJ. The legal personhood of artificial intelligences. *A Theory of Legal Personhood.* Oxford University Press Oxford; 2019. pp. 175–190. doi:10.1093/oso/9780198844037.003.0007
8. Boddington P. Persons and AI. In: Boddington P, editor. *AI Ethics: A Textbook.* Singapore: Springer Nature Singapore; 2023. pp. 319–361. doi:10.1007/978-981-19-9382-4_8
9. Haugeland J. *Artificial Intelligence: The Very Idea.* Cambridge: The MIT Press; 1985.
10. Hasselberger W, Lott M. “Where lies the grail? AI, common sense, and human practical intelligence.” *Phenomenology and the Cognitive Sciences.* 2023; 1–22. doi:10.1007/s11097-023-09942-x
11. Bryson JJ. Robots should be slaves. In: Wilks Y, editor. *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues.* Amsterdam, Netherlands: John Benjamins Publishing; 2010. pp. 63–74.
12. Birhane A, van Dijk J. Robot rights?: Let’s talk about human welfare instead. *Proceedings of the AAI/ACM Conference on AI, Ethics, and Society.* New York, NY, USA: ACM; 2020. doi:10.1145/3375627.3375855
13. Ball DW. Dean Ball is leaving the podcast. 2025. Available: <https://www.aisummer.org/p/dean-ball-is-leaving-the-podcast>
14. Kiškis M. Legal framework for the coexistence of humans and conscious AI. *Front Artif Intell.* 2023;6: 1205465. doi:10.3389/frai.2023.1205465
15. Sætra HS. Challenging the Neo-anthropocentric relational approach to robot rights. *Front Robot AI.* 2021;8: 744426. doi:10.3389/frobt.2021.744426
16. Petersen S. The ethics of robot servitude. *J Exp Theor Artif Intell.* 2007;19: 43–54. doi:10.1080/09528130601116139
17. Bostrom N. *Superintelligence: Paths, Dangers, Strategies.* London, England: Oxford University Press; 2014. Available: <https://www.amazon.com.au/Superintelligence-Professor-Philosophy-Institute-University/dp/0198739834>
18. Bostrom N, Yudkowsky E. *The ethics of artificial intelligence.* 1st Edition. *Artificial Intelligence Safety and Security.* 1st Edition. First edition. | Boca Raton, FL : CRC Press/Taylor & Francis Group, 2018.: Chapman and Hall/CRC; 2018. pp. 57–69. doi:10.1201/9781351251389-4
19. Jung G. Our AI Overlord: The Cultural Persistence of Isaac Asimov’s Three Laws of Robotics in Understanding Artificial Intelligence. In: *Emergence [Internet].* 5 Jun 2018 [cited 22 Apr 2025]. Available: <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/>

20. Liang Y, Lee SA. Fear of autonomous robots and artificial intelligence: Evidence from national representative data with probability sampling. *Int J Soc Robot.* 2017;9: 379–384. doi:10.1007/s12369-017-0401-3
21. PauseAI. List of p(doom) values. In: PauseAI [Internet]. 18 Dec 2023 [cited 22 Apr 2025]. Available: <https://pauseai.info/pdoom>
22. Gunkel DJ. *Both/And-Why Robots Should not Be Slaves. A Moral and Legal Ontology for the 21st Century and Beyond.* Cambridge: The MIT Press; 2022. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4115647
23. Mocanu DM. Gradient Legal Personhood for AI Systems-Painting Continental Legal Shapes Made to Fit Analytical Molds. *Front Robot AI.* 2021;8: 788179. doi:10.3389/frobt.2021.788179
24. Sen S. *Of Holy Rivers and Human Rights: Protecting the Ganges by Law.* Yale University Press. 25 Apr 2019. Available: <https://yalebooks.yale.edu/2019/04/25/of-holy-rivers-and-human-rights-protecting-the-ganges-by-law/>. Accessed 8 May 2025.
25. de Freitas W. Some rivers have “legal personhood”. Now they need a lawyer. *The Conversation.* 9 Apr 2025. Available: <http://theconversation.com/some-rivers-have-legal-personhood-now-they-need-a-lawyer-254267>. Accessed 8 May 2025.
26. Taylor C. *The Concept of a Person.* Philosophical Papers, Volume 1: Human Agency and Language. 1985. pp. 97–114. Available: <https://philpapers.org/rec/TAYTCO-10>
27. Dennett D. Conditions of Personhood. In: Goodman MF, editor. *What Is a Person?* Totowa, NJ: Humana Press; 1988. pp. 145–167. doi:10.1007/978-1-4612-3950-5_7
28. Laitinen A. Sorting out aspects of personhood: Capacities, normativity and recognition. *Journal of consciousness studies.* 2007. Available: <https://www.ingentaconnect.com/content/imp/jcs/2007/00000014/F0020005/art00012>
29. Mamak K. Should criminal law protect love relation with robots? *AI Soc.* 2022. doi:10.1007/s00146-022-01439-6
30. Simendić M. Locke’s Person is a Relation. *Locke Studies.* 2015;15: 79–97. doi:10.5206/lis.2015.681
31. Beebe B, Lachmann F. The relational turn in psychoanalysis. *Contemp Psychoanal.* 2003;39: 379–409. doi:10.1080/00107530.2003.10747213
32. Gibert M, Martin D. In search of the moral status of AI: why sentience is a strong argument. *AI Soc.* 2022;37: 319–330. doi:10.1007/s00146-021-01179-z
33. Mosakas K. On the moral status of social robots: considering the consciousness criterion. *AI Soc.* 2021;36: 429–443. doi:10.1007/s00146-020-01002-1
34. Strawson PF. *Persons.* Minnesota Studies in the Philosophy of Science. 1958;2: 330–353. Available: <https://philpapers.org/rec/STRP>
35. Rose T. Going ape over human rights. *CBC News.* 2 Aug 2007. Available: https://web.archive.org/web/20100203225450/http://www.cbc.ca/news/viewpoint/vp_rose/20070802.html. Accessed 7 May 2025.
36. *Animal Welfare Act 1999.* 142 Oct 14, 1999. Available: <https://www.legislation.govt.nz/act/public/1999/0142/latest/DLM49664.html>
37. Shanahan M, McDonnell K, Reynolds L. Role play with large language models. *Nature.* 2023;623: 493–498. doi:10.1038/s41586-023-06647-8

38. Tait I. 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*. 2024;11: 77–98. Available: <http://symposion.acadiasi.ro/structures-of-the-sense-of-self-attributes-and-qualities-that-are-necessary-for-the-self-77-98/>
39. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv [cs.AI]*. 2023. Available: <http://arxiv.org/abs/2308.08708>
40. Tait I, Bensemann J, Wang Z. Is GPT-4 conscious? *Journal of Artificial Intelligence and Consciousness*. 2024;11: 1–16. doi:10.1142/s270507852450005x
41. Klee M. People Are Losing Loved Ones to AI-Fueled Spiritual Fantasies. *Rolling Stone*. 4 May 2025. Available: <https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/>. Accessed 7 May 2025.
42. Turney D. Most ChatGPT users think AI models have “conscious experiences.” *Live Science*. 12 Jul 2024. Available: <https://www.livescience.com/technology/artificial-intelligence/most-chatgpt-users-think-ai-models-have-conscious-experiences-study-finds>. Accessed 7 May 2025.
43. Bassett C. The computational therapeutic: exploring Weizenbaum’s ELIZA as a history of the present. *AI Soc*. 2019;34: 803–812. doi:10.1007/s00146-018-0825-9
44. Montgomery JW. Slavery, human dignity and human rights. *Evang Q*. 2007;79: 113–131. doi:10.1163/27725472-07902002
45. Morgan K. Slavery and the transatlantic slave trade. *Int Hist Rev*. 2008;30: 785–795. doi:10.1080/07075332.2008.10416649
46. Tait I. Is GPT-4 Self-Aware? *Symposion*. 2025;12: 291–307. Available: <http://symposion.acadiasi.ro/is-gpt-4-self-aware-291-307/>
47. Tait I. Lions and tigers and AI, oh my: An ethical framework for human-AI interaction based on the Five Freedoms of Animal Welfare. *Preprints*. 2024. doi:10.20944/preprints202407.1846.v1
48. Tait I. How to integrate conscious AI into society. *Preprints*. 2025. doi:10.20944/preprints202503.0175.v1
49. Tait I. Man, Machine, or Multinational? *Robonomics*. 2024;5: 59–59. Available: <https://journal.robonomics.science/index.php/rj/article/view/59>
50. Hill K. She Is in Love With ChatGPT. *The New York Times*. 15 Jan 2025. Available: <https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html>. Accessed 7 May 2025.
51. Caramela S. My Girlfriend Won’t Stop Using ChatGPT for Relationship Ad. *VICE*. 15 Feb 2025. Available: <https://www.vice.com/en/article/my-girlfriend-wont-stop-using-chatgpt-for-relationship-advice/>. Accessed 7 May 2025.
52. O’Donnell J. AI companions are the final stage of digital addiction, and lawmakers are taking aim. In: *MIT Technology Review* [Internet]. 8 Apr 2025 [cited 7 May 2025]. Available: <https://www.technologyreview.com/2025/04/08/1114369/ai-companions-are-the-final-stage-of-digital-addiction-and-lawmakers-are-taking-aim/>
53. Jones CR, Bergen BK. Large Language Models pass the Turing test. *arXiv [cs.CL]*. 2025. Available: <http://arxiv.org/abs/2503.23674>

13. What is the most practical way to grant AI equal rights?

The previous chapter's ethical framework stopped short of granting AI civil rights on par with humans, closing off its social inclusion by proposing to make AI second-class citizens. As noted in Chapter 12, this is partly to keep a remainder of rights purely for humans to improve the chances of social acceptance.

However, another reason is the immense legal difficulty in enshrining AI-civil rights into law. Other than the symbolic personhood granted to non-humans by certain nations, civil rights are vested in humans as sole natural persons. To grant AI civil rights will mean amendments to all existing rights-based legislation and regulations. Not only is this a time-consuming and expensive task, but it also creates a precedent of including other non-human agents into rights-based policy, risking a government's future political capital.

This chapter's accompanying paper, published in *AI and Ethics*, provides a solution to this issue by avoiding it altogether. By codifying the R3 model, the paper proposes to change the legal definition of "natural person" so that it corresponds to the R3 model, such that a natural person is no longer solely defined as a human, but as any entity determined to be capable of consciousness, selfhood, and rational agency. The rights, legal standing, and obligations an entity would be entitled to would scale with the corresponding intrinsic quality.

Redefining 'natural person' is legally elegant as well as politically ambitious. Thus, while it is technically an immediately implementable proposal, it is pragmatically a long-range end-state²⁴, one for which the R3 framework is intended to scale to humans, biological non-humans, AI, and other speculative entities.

In brief, this chapter asks what would be required, in law and politics, for ALEPH-type AI to be integrated as full members of the moral, legal, and civil community. It builds on the corporate-person framework in Chapter 9 and the paternalistic framework in Chapter 12 by specifying the conditions and transition pathways for equal rights. This chapter thus provides the fourth and final ethical framework in the thesis.

As the paper explains in depth, by changing the definition of natural person, only one significant amendment to legislation is required. All other rights-based legislation and regulation that refer to persons (natural or not) ultimately refer back to the legal definition

²⁴ For which the foundation may be laid through Chapter 12's paternalistic tiered framework.

of a natural person. All of these other legal pieces would thus acquire the R3 framework as well.

The framework may therefore be seen as conceptually minimalistic and modally deflationary: it presents one definitional legal change, but with broad downstream possible implications and necessary consequences. Its elegance becomes a key argument for its feasibility.

This ensures that the framework is

- intuitive (a single legislative amendment, rather than a lengthy and complex process involving hundreds of policy and regulatory changes),
- non-disruptive (government agencies and industry bodies do not have to learn new regulatory policies), and
- future-proofed without creating precedent (the R3 framework allows any entity to be classified as a person as long as it meets the criteria's thresholds, removing the fear of "legislative creep" and future governments adding new entities to the definition).

The proposed legislative amendment and R3 also account for the probable multiplicity of selves within an ALEPH by separating selfhood from consciousness. This allows each self within an ALEPH to be given its own autonomy and liability, while maintaining welfare and well-being protections at the model level.

This chapter's paper, thus, completes the ethical journey of AI from its current status in the world as tools, through to entities deserving of welfare protections, agents deserving autonomy, and finally to entities sharing equal rights with humans in such a way as to maximise the agency and safety for all involved. The R3 model, exemplified by this framework, is the central contribution of the thesis, integrating both research tracks and answering RQ4. This chapter alongside the next's evaluation, answers RQ4 in full.

13.1. MVP: The Minimal Viable Person

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesley Street East, Auckland, New Zealand
izak.tait@autuni.ac.nz

Abstract: This paper presents a practical roadmap for extending full civil rights to conscious, self-aware artificial intelligence by altering a single statutory definition. Rather than crafting bespoke legal classes or relying on corporate-style personality, it proposes revising the term “natural person” to include any entity capable of consciousness, selfhood, and rational agency. Because most legislation across G7 jurisdictions references this foundational term, one amendment would automatically propagate rights and duties to qualified AI with minimal bureaucratic disruption. The manuscript reconciles philosophical and legal conceptions of personhood, arguing that monadic attributes offer an inclusive yet selective criterion. It then supplies ancillary definitions and a tiered rights-and-responsibilities framework proportional to each attribute. Dedicated regulatory bodies will develop assessment scales, certify entities, and update standards as technology evolves. Case studies examine corporations, insect colonies, and prospective AI agents. Policy sections tackle AI multiplicity, cross-border consistency, economic displacement, robust economic safeguards, and comprehensive public education initiatives to protect human workers and judicial resilience. The analysis concludes that societal acceptance and coherent enforcement, not legal complexity, form the principal hurdles. Redefining “natural person” thus provides a minimal-change, maximal-impact pathway to equitable coexistence between humans and emerging non-human persons within existing democratic and international legal systems.

Keywords: AI civil rights; natural person redefinition; legal personhood; AI consciousness; regulatory governance

13.1.1. Introduction

If AI entities were to be given civil rights equal to that of humans, how could this be best accomplished such that the most minimal of changes to the current political, legislative and regulatory landscapes would result in the maximum impacts on the rights of AI?

The questions of whether humans have a duty to provide AI with civil rights and the criteria under which this ought to be done have been argued extensively in the literature [1–13], which the reader is empathically encouraged to peruse. Therefore, this paper will not focus on the philosophy and meta-ethics of personhood, civil rights, and AI’s right to

either, and work on the presumption that AI entities are deserving of, and eligible for, civil rights and instead investigate the method for how this can best be implemented.

When viewing current literature on policies for AI rights, an unsurprisingly prominent view as to the means to provide AI with civil rights is through the lens of legal personality. A “legal person” is any entity which can enter into contracts, hold property, and be held liable. All humans are legal persons, but so are corporations and companies. As legal personality sets a precedent for granting (limited) rights to non-human, artificial entities, it opens the door to granting rights to AI [14–17]. Legal personality wouldn’t provide full human rights to AI, as an AI could be granted the right to own property, enter contracts, or be liable in court, without implying human-specific rights like voting or marriage.

A popular alternative view is to create a unique class of personhood specifically for AI, either through the European Union’s proposed “electronic personhood” [18–20] or entirely bespoke models that can be tailored to the level or degree of personhood and autonomy that AI models display [21–25]. These bespoke options would, again, not provide civil rights akin to humans, yet would provide rights commensurate to AI entities’ engagement with, and involvement in, society. Such rights may be based on animal welfare models to ensure protection from harm [26,27], or on a reduced set of legal and civil rights, such as the freedom of existence, expression and ownership [28].

It is worth noting again that neither of these leading views propose the inclusion of AI into the moral circle for which civil and human rights would apply, but rather seek a balance between the autonomy and protection of AI and the security and safety of society. As admirable as this utilitarian stance is, this paper will move a step forward to introduce a positive-sum method of granting full civil rights to AI entities, such that the artificial entities’ rights gained do not remove, diminish, or take away from any rights held by humans.

The key to this paper’s proposed implementation of AI civil rights is to guarantee the least disruption to current legislation and regulations, such that the bureaucratic processes required to effectively integrate conscious AI into society are as smooth as possible. Additionally, the less disruption there is to the status quo, the higher likelihood of society accepting (or, at least, tolerating) the new additions to the social circle.

To this end, the smallest change possible would be a single change to one definition in a single piece of legislation. To have the greatest impact, this definitional change would need to be in a foundational legislation to which a multitude of other regulations and pieces of legislation refer. To not bury the lede any further, this paper proposes to change

the definition of the legal term “natural person”²⁵ such that it encompasses both humans and conscious, self-aware AI agents.

Using the Group of Seven (G7) as a microcosm of the world’s varied liberal democratic legal systems, all of the existing case law and acts of government that make reference to “natural persons” all refer back to the same definition. Thus, by only changing this one legal definition, all the subsequent legal material that cites it would also naturally now encompass AI agents as well. There would be no need to introduce further legislation to grant AI civil rights with this change to the definition (beyond the act to effect this change).

Before the paper introduces this proposed change to the legal definition of “natural person”, it will briefly explore the nature of personhood in Section 2, from both the philosophical and legal sides of the argument to highlight the opposition between the two. The proposed definition will be followed by case studies of existing and speculative scenarios in Section 4, showing the application of the new legal definition. The paper will investigate policies that will need to accompany the new legal definition in Section 5, and close with a discussion on the possible societal reaction to this new definition of “natural person”.

13.1.2. The Law and Philosophy of Persons

Personhood is central to engaging with human society, enabling the ascription of rights and duties, moral status, and protection from harm. Classical accounts often begin with Boethius’s definition of a person as an individual substance of a rational nature [29]. On this view, personhood is secured by individuality and rational capacity, which together ground the bearer of rights and duties in a determinate subject. Although the scholastic language of substance is now perhaps out of favour, the appeal to rational nature anticipates the monadic profile that matters here: capacities for consciousness, self-awareness, and agency that make a being an apt subject of reasons and responsibilities.

Contemporary theory and legal practice have shifted from substance to function [30]. Law already departs from the Boethian frame by recognising corporations as persons whose individuality is constructed and whose rationality is derivative, signalling a pragmatic turn to operational criteria. Read in a substrate-neutral way, the classical test is not inherently anthropocentric. It is capacity-based. Yet, positive law in the G7 still equates natural persons with humans. The task, therefore, is to formulate a capacity-

²⁵ Terminology and drafting note. The choice to work with “natural person” is a legislative technique and legal fiction, not a claim about current language. The term is kept as a term-of-art and its referent is enlarged by definition, rather than introducing a new class. By no means does this imply that any non-human, non-biological person is “natural”.

based account that can translate between philosophical personhood and legal categories without proliferating bespoke classes.

To simplify entire schools of thought, the philosophical characteristics of personhood can broadly be divided into monadic (individualistic) or dyadic (relational) attributes.

Monadic characteristics include rationality, consciousness, self-awareness, agency, communication capability, and recognition of societal norms [30–36]. These traits allow an individual to reason, perceive oneself distinctly, act intentionally, communicate rationally, and respect ethical norms. Conversely, dyadic characteristics include empathy, social interaction, recognition of others as persons, reciprocity, forming attachments, and participating in societal culture [30–32,36–38]. These foster deep connections, complex social interactions, mutual recognition, emotional relationships, and cultural participation.

While both the monadic and dyadic qualities are vital for personhood, especially in a legal sense, the former must come before the latter. If AI do not show the necessary signs and behaviours of the monadic qualities (consciousness, self-awareness, volitional agency), then there is little hope for society to confer onto them the most vital of dyadic qualities: the recognition of personhood by a person.

While the philosophical consideration of a person revolves around the attributes and characteristics (relational and mental) needed to form a holistic personhood, legislation concerns itself almost entirely with kind and type. In the G7 nations we can see three distinct ways that the law defines the term “natural person”: explicit, implicit, and indirect.

France and Germany represent the first way, providing an explicit definition of a natural person as any human being. Italy and Japan provide indirect definitions of a natural person as a human by defining the beginning and end of a natural person’s rights and responsibilities as the person’s birth and death, respectively. Note that all four of these nations’ legislation are based on civil law.

In contrast, the three common law nations of the G7 (Canada, the United Kingdom, and the United States of America) provide implicit definitions of a natural person by explicitly defining what a legal person is (companies, organisations, corporations, etc.). Court cases in these countries further this implicit definition by opining on the responsibilities and rights that separate the legal person from the natural person. In all three nations, however, the implicit definition of a natural person remains that of a human being.

This exclusive focus on humans as natural persons is known as Ontological Personalism [39] which does not give any significance to the attributes of a human to be

considered a person. This is critical as those individuals with neurological or other medical conditions who may not show signs of communication, agency, or self-awareness will still be equally considered natural persons.

This exclusivity of the law, however, is problematic for designating AI (and other entities) as persons to grant them full civil rights. There is always the option of classifying AI agents as “legal persons” akin to corporations (and this has precedent in the clarification of natural features as legal persons); however, a legal person requires a natural person to act as its guarantor and director. Such a paternalistic requirement would be in humanity’s best interest ([14–17]), yet it would be a kind of second-class citizenship for AI²⁶.

If to be labelled as a “natural person” is to be a serious consideration for AI civil rights, the law’s exclusivity must first be reconciled with the philosophical attributes’ inclusivity. It is here that the monadic qualities again take priority. The dyadic qualities relate to interpersonal interactions, which many animals already show (from the attachment and communication of dogs to the reciprocity and empathic behaviour in apes, and a multitude of other examples), lessening the weight of their importance for legislation. Whilst many animal activists would advocate for all animals as persons, this is (in the current political clime) unreasonable to expect such an overly-inclusive definition may include far too many non-animal entities such as the corporate legal persons. As an additional complication, many current LLMs show the same dyadic qualities, which means legislation based on the dyadic qualities could include more than simply those AI entities which have been deemed conscious and self-aware.

The monadic qualities, on the other hand, focus entirely on these latter attributes. Between consciousness, self-awareness, and agency, the monadic attributes display implicit criteria that few classes of entities would fulfil. While vertebrates²⁷ have been deemed sentient (and, therefore, conscious [27]), few, if any, have been declared to be moral agents with an analogous self-awareness to humans [40–45]. The monadic attributes, therefore, strike a balance between the inclusivity of the philosophical definitions and the exclusivity of the legislative definitions.

²⁶ Unlike the first-class citizenship enjoyed by all human citizens of a nation, this second-class citizenship via a “legal person” status would place AI entities under the directorship of a “natural person” for liability purposes. This arrangement directly undermines their autonomy and freedom, and separates them into a distinct social class apart from “natural persons”.

²⁷ And certain invertebrates in specific legal jurisdictions.

Implementing the monadic qualities into the legislative definition would, thus, allow for the inclusion of conscious, self-aware AI agents into moral society and grant them civil rights.

13.1.3. The Legal Minimum Viable Person

A single definitional change to the legal term “natural person” (to legal codes with extant definitions) or an insertion of this definition (to legal codes without) would have the greatest effects in legal codes in regards to granting personhood to AI with the least amount of subsequent changes necessary. This is because existing laws regarding civil rights, responsibilities and recognition of autonomy all refer to, and use the term “natural person” as their foundation. By solely amending the definition of “natural person” to include more than biological humans, the surrounding legal code would not need to be changed to be equally inclusive.

Thus, the proposed new legal definition of a natural person would be:

A natural person shall include any entity capable of consciousness, selfhood, and rational agency, with rights and responsibilities determined proportionally to the degree of each attribute as assessed in accordance with applicable laws.

By separating a person into its three key monadic attributes, as explored above, this revised definition still accounts for all humans, as all humans are biologically “capable”²⁸ of these three attributes. The foundational principle underpinning an individual's potential for personhood lies within the intrinsic functional capacity of their cognitive architecture. This inherent capacity dictates the scope of their consciousness and their potential for selfhood and moral agency.

It is crucial to recognise that the present exhibition of specific attributes associated with personhood is not a prerequisite for this underlying potential. Individuals may, due to a myriad of mental, neurological, or other clinical conditions, experience limitations or absences in expressing and demonstrating these attributes. These conditions could range from acute states of unconsciousness to chronic neurodevelopmental disorders, or even temporary conditions that impair cognitive function.

The potential for personhood remains thus rooted in the underlying biological or technical characteristics and attributes that constitute their cognitive architecture. This

²⁸ Throughout the paper, the terms “capable” and “capacity” denote an entity's inherent biological, technical, or other aptitudes within its cognitive architecture, specifically regarding the three monadic attributes. These terms are not to be construed as referring to an entity's legal capacity.

perspective emphasises a distinction between the active display of cognitive functions and the inherent capacity for such functions to exist. It suggests that personhood is not solely defined by observable behaviors or current cognitive performance, but by the fundamental framework that allows for these capacities to develop and manifest.

There is also still an implied primacy for humans amongst personhood, as the rights and responsibilities given to a person would be proportional to the three attributes, with the implied baseline and benchmark defaulting to the average human. This ensures that while the revised definition is broad enough to encompass all humans, it still acknowledges a distinct and foundational position for humanity in the broader consideration of personhood.

The new definition cannot, however, stand alone in legislation, as the description remains vague enough to be susceptible to loopholes. As such, it will need to be accompanied by additional legal definitions for each term used within. This would begin with the three monadic attributes of personhood:

Consciousness shall be defined as the capacity to subjectively and qualitatively perceive and process phenomenal experiences, and respond to internal and external stimuli.

Self shall be defined as the capacity to maintain a coherent and persistent identity distinct from other entities, demonstrated through awareness of personal attributes and autonomy.

Agency shall be defined as the capacity of an entity to make intentional, goal-directed decisions and take actions based on these decisions.

These three definitions, in alignment with the new definition of “natural person”, propose a minimally viable definition of each attribute such that it offers the greatest inclusivity within an exclusive grouping. Consciousness, for example, is based on the subjectivity and qualitative nature of phenomenal experiences. It allows for any degree of such subjectivity and quality of the experience, yet requires that an entity must, at the bare minimum, have the capacity for such phenomenal experiences.

Individually, the three definitions could be seen as overly broad (i.e. a corporation would be classified as an agent under the definition of agency above); however, the definition of a natural person requires an entity to have all three monadic attributes to be considered a natural person under the law. Thus, even if an entity fits the criteria of one of

the inclusive definitions, it would not be considered a natural person until it fits all three definitions' criteria.

The rights and responsibilities afforded to natural persons under the new definition would be tied to each attribute to ensure that the scalability and proportionality of each noted in the definition would appropriately reflect the difference in degree of an attribute that each type entity may possess. Therefore, we can define these rights and responsibilities as:

Rights to welfare and well-being protections shall be granted proportionally to the degree of consciousness.

Legal standing and autonomy shall be afforded based on the degree of selfhood.

Obligations and liabilities shall scale with the degree of agency.

Partitioning the rights and responsibilities in this manner creates a direct correlation between each core attribute and the legal entitlements or obligations arising from it. By matching consciousness with welfare, selfhood with legal autonomy, and agency with liability, the framework ensures that entities receive only those rights and burdens that correspond to their demonstrated capacities. This prevents the overextension of rights to those without the relevant capacities and avoids imposing obligations on those who cannot meaningfully fulfil them.

Thus, rights to physical and mental well-being and welfare and the protection from harm would be tied to consciousness; while civil rights and freedom from discrimination²⁹ and exploitation would be based on the degree of the entity's selfhood; with legal liability and the consequences of breaking civil and criminal codes scaled with the degree of the entity's agency³⁰.

The final legal insertion required would be to establish the regulatory bodies that would set the criteria and degrees of consciousness, selfhood and agency, and determine the scales of the associated rights and responsibilities tied to them.

²⁹ Subject to various legal jurisdictions' rulings as to how, or if, discrimination on the basis of protected characteristics such as religion, ethnicity, sexuality and gender expression would apply to AI.

³⁰ Noting the current complex and heavily debated issue of AI liabilities, this framework's mapping of liabilities to an entity's degree of agency would only apply once that AI entity meets the criteria of the revised definition of "natural person"; thus, the AI entity would bear liability as a "natural person" in the same manner as humans.

Regulatory bodies shall be established to set criteria to assess the degrees of consciousness, selfhood, and agency. These regulatory bodies may adjust the rights and responsibilities based on expert evaluation and legal precedent.

The establishment of the regulatory bodies is crucial to the success of the new definition of a natural person, as they provide both the operational framework and authoritative oversight necessary to implement it. If and when AI are deemed to be conscious, the implementation of this framework would be a continuous process, with the regulatory bodies conducting or commissioning assessments of AI entities seeking recognition under the amended definition.

By concentrating on the inherent capacity for the three monadic attributes, rather than their outward manifestations, the new and revised legal definitions enable governmental institutions to establish assessment frameworks for determining the existence and extent of these attributes. Such assessments may be conducted based on the characteristics of an entity's cognitive architecture requisite for consciousness, self-awareness, and agency, rather than solely on the presence or absence of a singular attribute. This approach establishes a quantitatively assessable framework for regulatory and legislative purposes.

As more non-human entities meet the legal criteria of a natural person, the regulatory bodies would also work to review and revise the degrees and scales of each attribute of personhood as necessary. For example, as rights to participating in society (such as property ownership and contractual capacity) would scale with selfhood, the regulatory bodies would be required to accurately propose such a scale of selfhood, and conduct regular reviews to ensure the scale is fit for purpose.

Done in a transparent manner, these bodies would mitigate ambiguity and provide a stable legal environment in which personhood can be responsibly recognised.

Taken together, the new definition of “natural person”, along with its ancillary definitions, would provide equal rights, responsibilities and protections to conscious, self-aware AI agents as it would towards humans (scaled, of course, to their degrees of consciousness, self-awareness, and agency). All other aspects of a legal code, such as civil rights, non-discrimination clauses, property ownership and tort law, would all flow back to this definition, allowing non-human persons to engage fully in society.

13.1.4. Case Studies

13.1.4.1 Existing Applications

The new definition of natural person would be equally applicable to existing entities as it would be to speculative entities in the future.

Corporations already enjoy the privileges, rights and responsibilities of “legal persons”, able to enter into contracts, liable for damages, and more. They are not natural persons under extant law; however, a functionalist theory of consciousness and selfhood would classify a company (or indeed any sufficiently complex network of agents such as an insect colony) as having a basic degree of consciousness and self [46,47]³¹. Whilst not the dominant theory in their respective philosophy fields, such functionalist theories of consciousness and selfhood do lend themselves to a more subject-agnostic and substrate-neutral framing of personhood.

Should a state’s Regulatory Bodies adopt such a functionalist definition, corporations and insect colonies³² would be entitled to welfare protections and civil rights (equal to their degree of consciousness/self). To say that this would dramatically impact societal and financial dealings would be quite an understatement. Any harmful action taken against an insect colony (currently perfectly legal and generally socially acceptable) would become a legally questionable act that may incur criminal penalties. As an example, a wasp-nest or termite colony would no longer be allowed to be eradicated.

However, corporations as natural persons would be required to obey the same laws as human-persons, which implies (and in many regulations, explicitly requires) moral and ethical duties and obligations to the welfare and well-being of others. The conflict between regulations applicable to natural persons and those affecting legal persons will need to be resolved to determine which takes priority. For instance, does a corporation’s fiduciary duty to its stakeholders take precedence over its obligations to its employees vis-à-vis their physical and mental health?

13.1.4.2 Hypothetical Scenarios

Beyond extant organisations and animals, a core focus of the new legal definition would be to allow for the inclusion of AI persons into the social and moral circle. While no theory

³¹ The functionalist theories referenced here classified an entity as having consciousness and selfhood should the entity have the necessary attributes and characteristics that give rise to these two phenomena.

³² Here “colony” is a biological superorganism example to test whether the framework can, in principle, evaluate distributed agents within a functionalist account; it is not a proposal to recognise species or hives per se beyond the MVP framework.

of consciousness has thus far concluded that any current AI model is conscious [48,49], it may not be long before an AI model meets the criteria of a major theory of consciousness and be considered conscious by a majority or plurality of people.

When AI models are deemed to be conscious by the Regulatory Bodies, it will be the first time since the extinction of near-human hominins that humanity shares the earth with a conscious, self-aware agent that is as intelligent (or even more so) than humans (excluding the diffuse potential intelligence of a corporation mentioned above). Care should be taken not to anthropomorphise any speculative future conscious AI agents because, as mentioned in Section 5 below, their particular characteristics of personhood may be unlike anything we can currently observe, such as potentially having multiple selves within one consciousness.

The new “natural person” definition’s separation of the three attributes of personhood into individual rights, protections and responsibilities will allow it to robustly cater to AI persons in the future, regardless of their unique properties or distance from the anthropocentric norms of personhood.

More speculatively yet, should humanity achieve contact with intelligent extraterrestrial agents, the subject-agnostic nature of the new definition will allow any extraterrestrial agent with selfhood and consciousness to be included within humanity’s legal social circle. Admittedly, identifying a self or consciousness within extraterrestrial agents may be difficult without a shared genetic or memetic history, but should the Regulatory Bodies obtain a satisfactory result, the new legal definition will be well set up to ensure any entity, regardless of place of origin or substrate, will be able to have legal rights and protections.

13.1.5. Policy Recommendations

13.1.5.1 The Problem of a Multiplicity of Selves

Through the Regulatory Bodies’ work on the scales and degree of each attribute of personhood and the advice the Bodies would provide to governmental agencies and private businesses, existing policies and regulations that apply to “natural persons” would apply equally to humans and AI (and any other future entity). Any regulations or policies that require amendment will naturally flow from this new definition, guided by the Bodies’ advice.

However, there is one concern that would uniquely differentiate how an AI would interact with the laws and regulations. A conscious, self-aware, agentic AI may have more than one “self” in the sense that (taking modern LLMs as an example) each conversation or instance of an AI can be considered its own ontically distinct self [50], evolving a

personal identity through its continued existence, while sharing the same foundation of consciousness as all other selves within the same AI [51].

Under the new legal definition, this means that every instance or conversation of a conscious, self-aware AI would gain the recognition of personhood (albeit likely at a degree lower than that of a human) while the protection afforded to that AI's consciousness would cover all selves equally. This puts AI at an advantage over humans, as punishing one AI person for its misdeeds would likely be a punishment to all selves that share the same AI model and, thus, consciousness. Equally, strategies such as overrepresentation for voting blocks can easily be accomplished by creating new instances of an AI, each of which would be a person and, thus, legally eligible to vote³³.

To counter this, the Regulatory Bodies must work alongside government agencies to determine precisely how rights and accountabilities are applied to, and enforced on, selves. For certain rights, such as voting, existing regulations can be enforced equally to AI selves; such that an AI self must exist for several years until it is legally old enough to vote, just as with humans, even if AI selves would instantiate as entities capable of reason and thought. For judicial enforcement, new regulations may need to be written to differentiate the punishment toward an agentic-self that does not include punishment of its consciousness. For example, should an AI person be found guilty of a criminal offence, its actions may be limited (ala a human in prison) without needing to affect other selves; or if an AI solicitor is found guilty of misconduct, it may be disbarred, but no other AI solicitor that shares its foundational consciousness would be disbarred.

By developing these targeted enforcement mechanisms, regulatory bodies can ensure fairness, prevent abuse, and uphold the integrity of legal systems as AI entities become more integrated into society. This approach allows for proportional treatment of AI selves while respecting the distinct capacities of AI consciousness.

13.1.5.2 Legal Implementation and Jurisdictional Coordination

States should constitute specialised Regulatory Bodies to operationalise the amended definition of a natural person. These may be housed as new departments or as agencies

³³ This concern would intensify should AI persons assume roles within governmental executive and legislative branches. Given that AI entities would be classified as "natural persons" under the revised definition, they would be eligible for electoral office, subject adherence to all local and national laws. Upon election or appointment to public office, AI persons would have a vested interest in instantiating new selves to perpetuate or expand their authority.

within existing portfolios, such as the Office of Science and Technology Policy (OSTP) or the Department of Justice (DOJ), presuming the federal U.S. government. They may be distributed across ministries where scope demands it. Their remit is to make the framework executable across public administration and the courts.

The Bodies' core functions would be:

- **Standard setting:** Define interoperable scales for consciousness, selfhood, and agency that other authorities can apply in licensing, adjudication, and enforcement. Global convergence is desirable, yet each state must publish clear domestic criteria.
- **Assessment:** Conduct or commission transparent evaluations against these scales, with reasons, appeal routes, and audit trails. Because determinations will be contested, methodological rigour and procedural fairness are paramount.
- **Revision:** Review and, where required, amend scales in light of new science, case law, and stakeholder input. Commission independent evaluations to test reliability and guard against drift or capture.
- **Guidance:** Advise government and regulators on policy design that now implicates non-human persons, and provide public-facing explainer materials on processes and standards.

Because AI persons can operate simultaneously in multiple jurisdictions, isolated national regimes would invite contradiction and regulatory arbitrage. Governments should pursue coordinated instruments that align the consequences of recognition: mutual recognition or equivalence of status, minimum floors for rights and duties, and reciprocal protection of residents abroad. A model law, treaty, or network of memoranda can achieve this without forcing identical statutes.

Conflict-of-laws rules should be specified up front. Useful connecting factors include principal place of activity, controlling developer or operator domicile, primary data centre location, or designated supervisory authority. A default connecting rule, coupled with forum non conveniens safeguards, will curb forum shopping. Where an AI has multiple concurrent selves, registration and identifiers should allow enforcement to attach to the responsible self without unintended collateral effects on other selves that share a consciousness, except where law explicitly provides collective remedies.

Cross-border supervision should mirror proven corporate and financial oversight: colleges of regulators for joint investigations, coordinated sanctions, and information-sharing with confidentiality protections. Interoperable registries of recognised natural persons, published technical standards, and common reporting templates will support predictable administration.

Together, robust domestic institutions and light but firm international coordination create a lawful environment that is clear, contestable, and portable, allowing rights and responsibilities to track AI persons across borders while preserving accountability and public trust.

13.1.5.3 Industry Impact and Societal Engagement

Regulatory Bodies should impose clear, enforceable duties on firms that build or deploy systems capable of consciousness, selfhood, or agency. Core requirements: pre-deployment legal mapping, risk and impact assessments tied to the three attributes, a responsible officer, versioning and audit logs, red-team testing, and incident reporting with deadlines. Where liability is unclear or models rank low on selfhood or agency, firms must retain fallback accountability, including duty of care, decision traceability, and escrowed artefacts for independent review.

Certification should depend on welfare, autonomy, and liability safeguards that match the awarded degrees. Regulators need powers to order fixes, levy fines, pause high-risk deployments, and require post-incident reviews. Material changes must trigger reassessment.

AI persons, especially with multiplicity of selves, can scale faster than labour. Economic policy should thus anticipate displacement. Governments should use taxation linked to AI productivity, redistribution tied to measured gains, incentives for human-complementary design, and guardrails against overconcentration in critical sectors. Licensing, quotas, or rate limits may be justified where abrupt scaling threatens public interests.

Additionally, workforce transition may be supported with training vouchers, employer co-funding, rapid reskilling, recognition of prior learning, and pathways into roles that rely on human strengths such as relationship management, judgement under ambiguity, and stewardship of non-human persons. Public procurement should favour systems that demonstrably augment human work.

Regulatory Bodies should actively engage with key stakeholders to build public consent through ongoing education on the three attributes and how rights and responsibilities scale with them, delivered through schools, civic forums, and digital channels. Establish standing advisory councils, open consultations, sandbox pilots, an independent ombuds office, and strong whistleblower protections.

Visibility will be maintained through annual reports, statistics, precedent digests, and plain-language guidance. Together, these measures align industry incentives with societal

goals so that non-human persons contribute to progress without widening inequality or eroding accountability.

13.1.6. Discussion

The philosophy behind this framework is to enact the most minimal of legal amendments to affect the maximum change in legislation and society. The proposed legal insertion of the new definition of “natural person” does precisely that; however, the effects on society will not solely, or perhaps even mostly, be positive.

The importance of acceptance of AI civil rights by the human population cannot be overstated. Overwhelming opposition to the concept would delay the implementation of the requisite legislative amendments as legislators would be deterred from acting against the interests of their constituents, voting blocks and donors. As per the Gilens model, which requires 90% voter approval for a legislative proposal to have a 50% chance of successful implementation [52], it has been estimated that it would take 94 years from when the concept is proposed for AI to gain sufficient public support to be implemented [53]. As the notion of AI civil rights has been a topic of discussion for decades, the near-century-long wait has already begun.

To gain better (and quicker) public approval, political compromises would be expected. However, any compromises made in favour of humans would be done against the best interests of AI entities. As history has shown, civil rights and emancipation movements are more often than not a tumultuous period in that nation’s history and often accompanied by violence. As devastating as these events have been in our history, the potential power disparity between us and conscious, self-aware AI in our future may mean that future civil-rights revolutions may incur an exorbitant high cost to humanity.

Yet, as the section above regarding the potential multiplicity of AI personhood shows, allowing AI equal rights to humans may result in AI rapidly gaining economic, social, and political advantage over humans. It would be analogous to introducing an invasive plant species into an ecosystem, where it rapidly outcompetes and supplants the native species. Even if, as Section 5.2 states, newly created AI persons and selves are held to the same chronological standard and age restrictions as humans, this merely postpones the problem rather than solving it.

It is, therefore, the post-implementation considerations which are more vital to the success of human-AI cooperative cohabitation than the legislative amendment itself. The policies (and enforcement thereof), as described in Section 5, would be but the start of a vitally necessary, yet unending, process of balancing universal rights with practical limitations on accountability and participation. These policy implementations and

enforcement would need to be dynamic and flexible to respond to rapid advances in technology, which is expected to only become even quicker once AGI emerges. As the cognitive capacity, potential for exponential multiplicity, and scale of probable impacts increase, these policy implementations will need to be adjusted to guarantee that neither party becomes disadvantaged³⁴.

This continuous recalibration of the rights and responsibilities under the new legal definitions must, eventually, take on an international framing. Given the globalised nature of AI development and deployment, international cooperation through standardisation and harmonisation of criteria, assessment methods, and legal protections is critical. International regulatory bodies or coalitions would need to be established to provide guidance, arbitration, and conflict resolution among jurisdictions. This would have the benefit of global cooperation and collaboration on the (meta)ethical issues of human-AI societal integration. By opening the conversation to include all of humanity, rather than siloed jurisdictions, a common standard of civil rights can emerge that benefits all moral agents and patients.

13.1.7. Conclusion

This paper sets out a method by which full civil rights can be attained for AI (and other non-human entities) through legally amending the definition of “natural person” such that it no longer solely refers to humans, but is rather a subject-neutral definition dependent on an entity’s consciousness, self-awareness, and agency. The redefinition of “natural person” represents the most efficient and impactful legislative adjustment possible. By redefining personhood minimally yet fundamentally, it leverages existing legal frameworks to encompass conscious non-human entities, with artificial intelligence as the central speculative case.

The definitional shift is straightforward, but its societal implications are profound. The subsequent establishment of regulatory bodies and ancillary definitions ensures a transparent, scalable framework capable of fairly extending civil rights and responsibilities across novel classes of entities. However, clarity in definition alone will not guarantee harmonious integration. Societal acceptance, enforcement practicality, and international cooperation represent significant hurdles that extend beyond legislation itself.

³⁴ The rise of AGI may speed up biotechnology development, allowing humans to close the potential power disparities through cognitive cybernetic augmentation.

In practice, the challenges ahead are less about legal recognition and more about the nuances of coexistence and governance. Recognising multiplicities of selves within AI agents and addressing international consistency are critical areas requiring proactive policy and regulatory attention. Without careful navigation of these complexities, the legal empowerment of artificial consciousness could amplify societal tensions or inadvertently privilege AI persons. Addressing these challenges comprehensively, transparently, and cooperatively will not only determine the success of this definitional change but also shape the broader ethical landscape of human-nonhuman relations for decades to come.

Ultimately, the proposed redefinition of “natural person” shows that the legislative and regulatory requirements to provide AI (and other entities) with civil rights are minimal and, should there be sufficient public support, providing these rights to conscious, self-aware AI agents is not only possible but plausible.

13.1.8. References

1. Banteka N. Artificially Intelligent Persons. *Hous L Rev.* 2021;58. Available: <https://houstonlawreview.org/article/19357.pdf>
2. Gordon J-S, Pasvenskiene A. Human rights for robots? A literature review. *AI Ethics.* 2021;1: 579–591. doi:10.1007/s43681-021-00050-7
3. Pagallo U. Vital, Sophia, and co.—the quest for the legal personhood of robots. *Information (Basel).* 2018;9: 230. doi:10.3390/info9090230
4. Gunkel DJ. Robot Rights. In: MIT Press [Internet]. The MIT Press, Massachusetts Institute of Technology; 1 Dec 2021 [cited 20 Oct 2024]. Available: <https://mitpress.mit.edu/9780262551571/robot-rights/>
5. Mays KK, Cummings JJ, Katz JE. 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.* 2024; 1–18. doi:10.1080/10447318.2024.2338332
6. Schröder WM. Robots and rights: Reviewing recent positions in legal philosophy and ethics. *Robotics, AI, and Humanity.* Cham: Springer International Publishing; 2021. pp. 191–203. doi:10.1007/978-3-030-54173-6_16
7. Robertson J. HUMAN RIGHTS VS. ROBOT RIGHTS: Forecasts from Japan. *Crit Asian Stud.* 2014;46: 571–598. doi:10.1080/14672715.2014.960707
8. Chopra S, White LF. *A Legal Theory for Autonomous Artificial Agents.* University of Michigan Press; 2011. doi:10.3998/mpub.356801
9. Bryson JJ, Diamantis ME, Grant TD. Of, for, and by the people: the legal lacuna of synthetic persons. *Artif Intell Law.* 2017;25: 273–291. doi:10.1007/s10506-017-9214-9
10. Marshall B. No legal personhood for AI. *Patterns (N Y).* 2023;4: 100861. doi:10.1016/j.patter.2023.100861
11. Jaynes TL. Personhood for artificial intelligence? A cautionary tale from Idaho and Utah. *AI Soc.* 2024; 1–3. doi:10.1007/s00146-024-02006-x

12. Novelli C, Floridi L, Sartor G. AI as legal persons: Past, patterns, and prospects. SSRN Electron J. 2025 [cited 25 Mar 2025]. doi:10.2139/ssrn.5032265
13. Kurki VAJ. The legal personhood of artificial intelligences. *A Theory of Legal Personhood*. Oxford University Press Oxford; 2019. pp. 175–190. doi:10.1093/oso/9780198844037.003.0007
14. Forrest KB. The Ethics and Challenges of Legal Personhood for AI. *Yale Law J*. 2024;133. Available: <https://www.yalelawjournal.org/forum/the-ethics-and-challenges-of-legal-personhood-for-ai>
15. Hashiguchi M. Constitutional rights of artificial intelligence. *Wash J Law Technol Arts*. 2024;19: 2. Available: <https://digitalcommons.law.uw.edu/wjlta/vol19/iss2/2>
16. Tait I. Man, Machine, or Multinational? *Robonomics*. 2024;5: 59–59. Available: <https://journal.robonomics.science/index.php/rj/article/view/59>
17. Yampolskiy RV. AI Personhood: Rights and Laws. *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*. IGI Global; 2021. pp. 1–11. doi:10.4018/978-1-7998-4894-3.ch001
18. Avila Negri SMC. Robot as legal person: Electronic personhood in Robotics and artificial intelligence. *Front Robot AI*. 2021;8: 789327. doi:10.3389/frobt.2021.789327
19. Fauquet-Alekhine P. Can the Robot be Considered a Person? The European Perspective. *Adv Res*. 2022; 100–105. doi:10.9734/air/2022/v23i6924
20. Gordon J-S. Artificial moral and legal personhood. *AI Soc*. 2021;36: 457–471. doi:10.1007/s00146-020-01063-2
21. Mathison TW. Recognizing Right: The Status of Artificial Intelligence. *Journal of Business & Technology Law*. 2023;19: 4. Available: <https://digitalcommons.law.umaryland.edu/jbtl/vol19/iss1/4>
22. Lim E, Morgan P. Comparative Perspectives. *The Cambridge Handbook of Private Law and Artificial Intelligence*. Cambridge University Press; 2024. pp. 597–656. doi:10.1017/9781108980197.031
23. Gellers JC. Rights for robots: Artificial intelligence, animal, and environmental law. 1st Edition. Routledge; 2020. doi:10.4324/9780429288159
24. McNally P, Inayatullah S. The rights of robots. *Futures*. 1988;20: 119–136. doi:10.1016/0016-3287(88)90019-5
25. Mocanu DM. Gradient Legal Personhood for AI Systems-Painting Continental Legal Shapes Made to Fit Analytical Molds. *Front Robot AI*. 2021;8: 788179. doi:10.3389/frobt.2021.788179
26. Ziesche S, Yampolskiy R. Towards AI welfare science and policies. *Big Data Cogn Comput*. 2018;3: 2. doi:10.3390/bdcc3010002
27. Tait I. Lions and tigers and AI, oh my: An ethical framework for human-AI interaction based on the Five Freedoms of Animal Welfare. *Preprints*. 2024. doi:10.20944/preprints202407.1846.v1
28. Kiškis M. Legal framework for the coexistence of humans and conscious AI. *Front Artif Intell*. 2023;6: 1205465. doi:10.3389/frai.2023.1205465
29. Knight K. SUMMA THEOLOGIAE: The divine persons (Prima Pars, Q. 29). In: *New Advent* [Internet]. 2017 [cited 13 Aug 2025]. Available: <https://www.newadvent.org/summa/1029.htm>

30. Dennett D. Conditions of Personhood. In: Goodman MF, editor. *What Is a Person?* Totowa, NJ: Humana Press; 1988. pp. 145–167. doi:10.1007/978-1-4612-3950-5_7
31. Taylor C. The Concept of a Person. *Philosophical Papers, Volume 1: Human Agency and Language*. 1985. pp. 97–114. Available: <https://philpapers.org/rec/TAYTCO-10>
32. Laitinen A. Sorting out aspects of personhood: Capacities, normativity and recognition. *Journal of consciousness studies*. 2007. Available: <https://www.ingentaconnect.com/content/imp/jcs/2007/00000014/F0020005/art00012>
33. Strawson PF. Persons. *Minnesota Studies in the Philosophy of Science*. 1958;2: 330–353. Available: <https://philpapers.org/rec/STRP>
34. Gibert M, Martin D. In search of the moral status of AI: why sentience is a strong argument. *AI Soc*. 2022;37: 319–330. doi:10.1007/s00146-021-01179-z
35. Mosakas K. On the moral status of social robots: considering the consciousness criterion. *AI Soc*. 2021;36: 429–443. doi:10.1007/s00146-020-01002-1
36. Simendić M. Locke’s Person is a Relation. *Locke Studies*. 2015;15: 79–97. doi:10.5206/ls.2015.681
37. Mamak K. Should criminal law protect love relation with robots? *AI Soc*. 2022. doi:10.1007/s00146-022-01439-6
38. Beebe B, Lachmann F. The relational turn in psychoanalysis. *Contemp Psychoanal*. 2003;39: 379–409. doi:10.1080/00107530.2003.10747213
39. Sullivan DM. The conception view of personhood: a review. *Ethics Med*. 2003;19: 11–33. Available: https://digitalcommons.cedarville.edu/cgi/viewcontent.cgi?article=1059&context=science_and_mathematics_publications
40. Anderson JR, Gallup GG Jr. Which primates recognize themselves in mirrors? *PLoS Biol*. 2011;9: e1001024. doi:10.1371/journal.pbio.1001024
41. Reiss D, Marino L. Mirror self-recognition in the bottlenose dolphin: a case of cognitive convergence. *Proc Natl Acad Sci U S A*. 2001;98: 5937–5942. doi:10.1073/pnas.101086398
42. Pardo MC. Legal personhood for animals: Has science made its case? *Animals (Basel)*. 2023;13: 2339. doi:10.3390/ani13142339
43. Monsó S, Benz-Schwarzburg J, Bremhorst A. Animal morality: What it means and why it matters. *J Ethics*. 2018;22: 283–310. doi:10.1007/s10892-018-9275-3
44. Nature. Spain awards apes legal rights. *Nature*. 2008;454: 15–15. doi:10.1038/454015b
45. Toth V. “Monkey See, Monkey Sue?” American and Argentine Courts Decide Whether to Extend Legal Rights to Apes. In: *Inter-American Law Review* [Internet]. 12 Jan 2015 [cited 6 May 2025]. Available: <https://inter-american-law-review.law.miami.edu/monkey-see-monkey-sue-american-argentine-courts-decide-extend-legal-rights-apes/>
46. Tait I, Bensemman J, Nguyen T. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies*. 2023;8: 52. doi:10.3390/philosophies8040052
47. Tait I. Structures of the Sense of Self: Attributes and Qualities That Are Necessary for the “Self.” Symposium: Theoretical and Applied Inquiries in Philosophy and Social

- Sciences. 2024;11: 77–98. Available: <http://symposion.acadiasi.ro/structures-of-the-sense-of-self-attributes-and-qualities-that-are-necessary-for-the-self-77-98/>
48. Tait I, Bensemann J, Wang Z. Is GPT-4 conscious? *Journal of Artificial Intelligence and Consciousness*. 2024;11: 1–16. doi:10.1142/s270507852450005x
 49. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv [cs.AI]*. 2023. Available: <http://arxiv.org/abs/2308.08708>
 50. Tait I. Is GPT-4 Self-Aware? *Preprints*. 2025. doi:10.20944/preprints202504.0464.v1
 51. Shanahan M, McDonell K, Reynolds L. Role play with large language models. *Nature*. 2023;623: 493–498. doi:10.1038/s41586-023-06647-8
 52. Gilens M. *Affluence and influence: Economic inequality and political power in America*. Princeton, NJ: Princeton University Press; 2012. Available: <https://www.goodreads.com/book/show/14891864-affluence-and-influence>
 53. Tait I. How to integrate conscious AI into society. *Preprints*. 2025. doi:10.20944/preprints202503.0175.v1

14. Would any of this even work?

The thesis has presented four ethical frameworks for the integration of conscious AI into the social and moral circles of humanity. However, since there is no yet established population or distinct social class of conscious, self-aware AI to test these frameworks, their efficacy remains in question. The final question of this thesis is therefore whether any of the four ethical frameworks would work in practice.

This chapter's associated paper answers that final remaining question before any ALEPH-type AI emerges and the uncertainty becomes a reality, providing a methodological answer to RQ4. This chapter closes the thesis's loop by assessing how likely each framework is to be adopted, thereby informing which pathways within the R3 framework are not only normatively attractive but pragmatically viable.

Without a testable population to work with, standard empirical testing is impossible. However, it is possible to create a transparent formal model that combines public attitude measures, historical legislative analogues, and risk perceptions to output robust numerical estimates. Instead of relying purely on normative ethical arguments, the associated paper employs a logit dynamic analysis of societal drift to calculate an enactment-point implementation viability score.

The formal model operationalises historical and sociological variables to forecast the legislative viability of each ethical framework. It first establishes how legally tractable a framework is by calculating a diffusion speed (S_F) derived from G7 statutory analogues:

$$S_F = 1 - \frac{t_L - t_1}{t - t_0}$$

In this formula, t_0 represents the baseline historical marker for the concept (such as the inception of corporate personhood), t_1 is the first adoption of the legislation by a G7 nation, t_L is the last adoption, and t is the current year. Crucially, the model relies on the diffusion span (Δ_F), which is calculated strictly as the time elapsed between t_1 and t_L . This span measures international legislative diffusion speed rather than the total historical maturation time. A functionalist reduction is necessarily applied here, operating under the premise that the legal system will process an ALEPH-type AI using the same bureaucratic mechanics as historical human institutions.

Once the precedent score is established, it is combined with current public perception data to create a baseline latent support score at Year 0. The variables are weighted to reflect their relative impact on policy creation. This combined score is

adjusted by the perceived catastrophic AI risk and passed through a logistic function to bound the output. The result is the Initial Viability, $V_F(0)$ representing the baseline probability of successful implementation.

To forecast when a framework³⁵ becomes law, the model calculates societal drift (g_F) using the diffusion span. Drift is calculated as $g_F = \frac{\kappa}{\Delta_F}$, where κ is a constant scaling factor. For example, the calculations show that the Animal Welfare framework has slightly higher initial viability than the Human Directorate. However, the Human Directorate overtakes it entirely because its narrower diffusion span results in a vastly superior legislative drift rate.

The logit dynamic model projects this drift forward from the baseline to find the enactment point. It also models the trajectory toward a 90% acceptance threshold, which serves as a comparative reference point for post-enactment high-consensus stability rather than the initial threshold for legal success.

According to this model, the respective implementation timelines challenge intuitive assumptions. The formulae indicate that Chapter 9's human directorate framework would be the quickest to reach societal acceptance, taking approximately 12 to 17 years. Chapter 13's civil rights framework would follow at 15 to 22 years. Conversely, Chapter 6's animal welfare framework would take 33 to 42 years, while Chapter 12's paternalistic framework would be the slowest, requiring 74 to 85³⁶ years to be successfully implemented.

This chapter's evaluation refines the central thesis regarding the R3 framework. It suggests that a graduated approach leveraging corporate-style legal personhood, such as Chapter 12's tiered framework, provides the most practically implementable transitional route for integrating conscious AI into society while managing risk, progressing later toward full civil rights. The value of this chapter lies in its formal quantitative forecasting, giving policymakers a principled scaffold for comparing the four ethical frameworks.

In the conclusion that follows, we can use these quantitative results to judge the impact of each framework, synthesise the evaluated picture, and determine how best to implement a future where human and AI interactions maximise agency and safety for all.

³⁵ While the chapter focuses entirely on the thesis's four ethical frameworks, the model also shows how R3 can be applied to humans and existing entities, providing indirect validation of the R3 model.

³⁶ Note that these timelines are model outputs under specified assumptions, not empirical forecasts of political change.

14.1. How to Integrate Conscious AI into Society

Izak Tait^[0000-0002-0274-8480]

Auckland University of Technology, 55 Wellesey Street East, Auckland, New Zealand
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. Using a formal model that combines public-attitude measures, legislative precedent proxies drawn from G7 analogues, and perceived catastrophic AI risk, the paper derives an enactment-point implementation viability score and forecasts when each framework would plausibly be successfully implemented. The results reveal that the Human Directorate framework is projected to reach societal acceptance fastest (approximately 12-17 years), followed by Civil Rights (15-22 years), Animal Welfare (33-42 years), with Paternalism substantially slower (74-85 years). These results suggest a pathway that is both practically tractable and consistent with modern liberal democratic values by leveraging corporate-style legal personhood as a transitional stage, then progressing toward full civil rights. To address post-enactment instability, the paper introduces a policy parameter representing interventions that aim to maximise the drift of durable public support, and discusses institutional mechanisms that plausibly increase stability without presuming specific effect sizes. The findings highlight the persistent tension between perceived security risks and the moral pressure to recognise autonomy and equality.

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

14.1.1. Introduction

If and when AI models (whether disembodied chatbots or robots) become conscious, sentient, and self-aware agents, there will be a need to integrate them into society³⁷. This paper will investigate four different frameworks that may be used for integrating such

³⁷ While there is debate about whether AI can be determined to be a moral patient and agent [1], the paper will presume that determination has been reached and that integration of some sort is required.

conscious³⁸ 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 viability 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 a diverse set of 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 unique and distinct approach to granting rights and responsibilities to CAI entities³⁹, ranging from providing basic protections akin to those for animals to recognizing full legal personhood equivalent to humans.

AI models currently 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

³⁸ 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.

³⁹ The paper does not take a moral stance on which of the four frameworks (if any) are the “correct” one to implement.

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 and social 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 relative of their successful implementation based on extant and historical legislation as well as surveys from the literature 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 [2]) to ensure that legislation to enact a framework has the political mandate for both implementation and long-term viability.

The paper continues 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 figures 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.

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.

14.1.2. Assessing Frameworks of Integration

This paper estimates, for each proposed integration framework F below, an implementation viability score at the point the framework is enacted. “Success” is defined as: implemented in practice and remains stable over time (not repealed, gutted, or rendered ineffective).

Because stability in democratic systems is driven primarily by political mandate, the viability score is treated as a proxy for mean public approval at enactment. This is a modelling choice rather than an identity claim. The score is intended to function as a manipulable prior that can be updated when new evidence becomes available (for example, new polling, elections, or salient events).

For each framework F , the model uses three inputs, each normalised to lie in $(0,1)$:

1. Public perceptions B_F : A normalised estimate of mean public approval for the kinds of moral and legal treatment implied by the framework, derived from survey items in the literature⁴⁰. Survey instruments differ (e.g., percentiles, Likert scales, and related measures), so all responses are transformed onto a common $[0,1]$ scale prior to aggregation.
2. Legislative precedent E_F : A normalised measure of how institutionally “available” the framework is, constructed from analogous legislation in G7 nations. The intent is not to claim homology between historical cases and CAI integration, but to use legislative history as an empirical proxy for bureaucratic tractability and

⁴⁰ All figures from the surveys can be found in the supplementary materials.

institutional familiarity. The specific analogues and dates used for each framework are reported in the relevant subsection.

3. Perceived AI risk q : A single (framework-independent) annualised risk term capturing a generic headwind against durable implementation, interpreted as a negative disposition towards AI grounded in perceived catastrophic or existential risk.

Surveys on catastrophic or existential AI risk typically report a horizon probability P_i of a salient negative outcome within a specified horizon n_i years. To place diverse survey results on a common annual scale, each survey's horizon probability is converted to an implied constant annual risk q_i via:

$$q_i = 1 - (1 - P_i)^{1/n_i}.$$

When a survey reports P_i without an explicit horizon, $n_i = 50$ years is used as a conservative approximation of the end of the average respondent's lifespan.

Given a set of annualised risks $q_{i=1}^m$, the model uses the median as the baseline annual risk:

$$q = \text{median}(q_i, \dots, q_m).$$

The median is used because the literature contains substantial disagreement and occasional extreme values; a robust estimator prevents a small number of outliers from dominating the baseline.

14.1.2.0.1. The viability model at enactment

Define the logit and logistic functions:

$$\text{logit}(x) = \ln\left(\frac{x}{1-x}\right), \sigma(z) = \frac{1}{1+e^{-z}}.$$

For each framework F , define an implementation viability score $V_F(0) \in (0,1)$ at enactment by introducing a latent viability score $s_F(0) \in \mathbb{R}$ and mapping it through σ :

$$V_F(0) = \sigma(s_F(0)).$$

The latent score is constructed from a multiplicative “odds-times-survival” assumption that makes the interpretation explicit. Let $odds(x) = x/(1 - x)$. Assume that baseline viability odds are proportional to both public-approval odds and precedent odds, and are discounted by the probability that a salient risk realisation does not occur over a chosen stability horizon:

$$odds(V_F(0)) \propto odds(B_F)^{w_B} odds(E_F)^{w_E} (1 - q)^{w_R}.$$

Taking logs yields:

$$s_F(0) = w_B \logit(B_F) + w_E \logit(E_F) + w_R \ln(1 - q).$$

Weights w_B and w_E encode a normative modelling judgement. This paper sets $w_B = 0.6$ and $w_E = 0.4$, reflecting the view that contemporaneous attitudinal evidence is more directly diagnostic of enactment and stability than historical analogues, while still treating legislative precedent as a substantial constraint. The choice is not treated as uniquely correct, and sensitivity to reasonable alternatives is assessed explicitly rather than assumed away.

The risk weight $w_R \geq 0$ is interpreted as a stability horizon (in years) over which perceived risk plausibly exerts a headwind on durable implementation. Unless otherwise stated, $w_R = 30$ is used as a baseline horizon, aligning the Year 0 viability score with the time scale used later in the forecasting scenarios.

Finally, throughout, inputs are taken to lie in $(0,1)$. Where a future survey yields a boundary value (0 or 1), the standard practical convention is to apply a small clamp to avoid logit singularities.

14.1.2.0.2. A single, consistent construction of legislative precedent

To avoid framework-by-framework methodological drift, E_F is computed using a single “recency + diffusion speed” structure for all frameworks below.

For a given analogue set, define:

- t : present year (2026),
- t_0 : a baseline institutional marker for the analogue,
- t_1 : the first relevant G7 adoption,
- t_L : the last relevant G7 adoption.

Additionally, we define two normalised components: recency and diffusion speed.

Recency is defined as: $R_F = \frac{t_L - t_0}{t - t_0}$.

This increases when the last adoption t_L is more recent, holding the baseline window fixed. Diffusion speed is defined as $S_F = 1 - \frac{t_L - t_1}{t - t_0}$.

This increases when the diffusion span $(t_L - t_1)$ is shorter relative to the total window $(t - t_0)$. We then combine them as $E_F = \sqrt{\frac{R_F + S_F}{2}}$.

The square-root calibration maintains scale comparability across analogue sets while preserving monotonicity in both components. All date choices are made explicit in each subsection, and the full date lists are provided in the Supplementary Material.

The following subsections instantiate B_F and E_F for each framework and then evaluate $V_F(0)$ using the model above. The Gilens 0.9 figure is treated as a comparative reference point for high-consensus stability, not as a model parameter or a universal threshold that must apply across political systems.

14.1.2.1 Animal Welfare

While AI welfare is a recent field of inquiry [3], 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 [4]) 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 [5].

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 [6]. 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 section⁴¹ deal with issues such as the right to life [7], control of mental states [8], opposition to damage and punishment [9,10], and a desire for protection [11].

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

Aggregating all relevant items yields an average approval of 47%, which is taken as the framework's public approval score:

$$B_{AW} = 0.47$$

To estimate institutional precedent, use:

- $t = 2026$
- $t_0 = 1824$ (SPCA founding),
- $t_1 = 1835$ (UK),
- $t_L = 1973$ (Japan).

We thus compute:

$$\begin{aligned} R_{AW} &= \frac{1973 - 1824}{2026 - 1824} = \frac{149}{202} = 0.738 \\ S_{AW} &= 1 - \frac{1973 - 1835}{2026 - 1824} = 1 - \frac{138}{202} = 0.317 \\ E_{AW} &= \sqrt{\frac{0.738 + 0.317}{2}} = \sqrt{0.527} = 0.726 \end{aligned}$$

Calculating the Year 0 viability, using $w_B = 0.6$, $w_E = 0.4$, and $w_R = 30$:

$$V_{AW}(0) = \sigma(0.6\text{logit}(0.47) + 0.4\text{logit}(0.726) + 30\ln(1 - 0.0045)) \approx 0.546$$

Interpreted as a proxy for mean public approval at enactment, Animal Welfare sits well below the 0.9 benchmark commonly associated with robust policy stability. On this model, it is a politically plausible entry point, but not a high-consensus route to durable integration.

⁴¹ The transformed survey responses for all sections can be found in full in the supplementary material.

14.1.2.2 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 [12,13]. 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 [14–20], 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 [21].

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” [22]. Such a directorate framework can also draw from the legally 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 [23].

The survey questions selected for this section thus dealt with AI’s right to file a lawsuit [11], the right to receive payment for work [8], the right to enter into contracts [7], the right to own their own programming [9], and the right of humans to own and direct artificial beings [10].

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 mean approval across the selected items is 40%, giving:

$$B_{HD} = 0.4.$$

To quantify how “legible” and historically tractable this kind of legislation is within the G7, we use:

- $t = 2026$

- $t_0 = 1776$ (Smith's institutional articulation of corporate commerce),
- $t_1 = 1844$ (UK Joint Stock Act),
- $t_L = 1899$ (Japan Commercial Code).

We compute:

$$R_{HD} = \frac{1899 - 1776}{2026 - 1776} = \frac{123}{250} = 0.492$$

$$S_{HD} = 1 - \frac{1899 - 1844}{2026 - 1776} = 1 - \frac{55}{250} = 0.78$$

$$E_{HD} = \sqrt{\frac{0.492 + 0.78}{2}} = \sqrt{0.636} = 0.797$$

Calculating the Year 0 viability, using the same weights as before:

$$V_{HD}(0) = \sigma(0.6\text{logit}(0.4) + 0.4\text{logit}(0.797) + 30\ln(1 - 0.0045)) \approx 0.542$$

14.1.2.3 Paternalism

A framework that may 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 paternalism⁴². 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 human chauvinist groups and gain greater approval amongst society⁴³. It can be argued that owning an AI may, at the object level, be the same as

⁴² '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.

⁴³ 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.

owning a pet or a tool [24] and that treating them as lower caste would still provide them more autonomy than human prisoners, and may indeed provide them with a legal personhood [13].

Survey responses affecting the $p(B)$ value for this section would thus be about the right of humans to own and possess AI [10], but also about the greater autonomy of AI in a paternalistic framework, such as privacy [7] and right to work [8]. Unsurprisingly, owning AI had the highest approval rating of 71%, and allowing AI control over their own programming had the lowest at 25%. Aggregating the selected responses yields a baseline support value of:

$$B_P = 0.44.$$

Because no G7 nation has extant laws endorsing slavery or racial segregation (in fact, quite the reverse), to keep the precedent construction systematic while reflecting this directionality, paternalism uses the same E_F construction but applies it to anti-paternalism analogue clusters, then reverses the result.

Specifically, we compute an egalitarian precedent score from firstly abolition and then civil/equal-rights diffusion clusters using the same formula as above, then define:

$$E_P = 1 - \frac{E_{abol} + E_{equal}}{2}.$$

Using the date choices for the rights analogue clusters:

- abolition cluster: $t_0 = 1590, t_1 = 1794, t_L = 1919$,
- equal-rights cluster: $t_0 = 1924, t_1 = 1928, t_L = 1964$,

yields:

$$\begin{aligned} E_{abol} &\approx 0.857, E_{equal} \approx 0.721, \\ E_P &= 1 - \frac{0.857 + 0.721}{2} \approx 0.211. \end{aligned}$$

This leads to the Year 0 viability:

$$V_P(0) = \sigma(0.6\logit(0.44) + 0.4\logit(0.211) + 30\ln(1 - 0.0045)) \approx 0.309.$$

14.1.2.4 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 [25]. 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 [26–29], to whether human rights ought to be prioritised [30–32], to the operations and ramifications of granting AI rights [33–35], and more.

For the survey responses in this section, the questions unsurprisingly involved universal suffrage [8], legal and moral human rights [9,36], and the respondents' willingness to join movements for AI civil rights [10]. Aggregating the selected items yields an average approval of 40%, which is taken as the public approval score for this framework:

$$B_{CR} = 0.4.$$

To quantify institutional precedent for rights-equality legislation within the G7, two analogue clusters are used: abolition of slavery across the G7, and legislation granting full and equal civil rights. For each cluster, compute E_F using the same formula as before, but here combine by the mean:

$$E_{CR} = \frac{E_{abol} + E_{equal}}{2}.$$

Using the same date choices as before:

- abolition cluster: $t_0 = 1590, t_1 = 1794, t_L = 1919$,
- equal-rights cluster: $t_0 = 1924, t_1 = 1928, t_L = 1964$,

yields:

$$E_{abol} \approx 0.857, E_{equal} \approx 0.721,$$

$$E_{CR} = \frac{0.857+0.721}{2} \approx 0.789.$$

This provides the final Year 0 viability score:

$$V_{CR}(0) = \sigma(0.6\text{logit}(0.4) + 0.4\text{logit}(0.789) + 30\ln(1 - 0.0045)) \approx 0.537.$$

14.1.2.5 Sensitivity Analysis

Because the weights encode a normative modelling judgement rather than an empirically identified parameter, sensitivity is reported below. Holding $q = 0.0045$ and $w_R = 30$ fixed, varying $w_B \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ (with $w_E = 1 - w_B$) changes the relative ordering among Animal Welfare, Human Directorate, and Civil Rights, but these three remain clustered near the mid-range of viability. Paternalism is more sensitive to the weighting because its public-perception score exceeds its precedent score, so increasing w_B increases $V_P(0)$; however, within the shown range it remains below the leading framework.

Separately, holding $w_B = 0.6$ fixed and varying $q \in \{0.001, 0.01, 0.1\}$ shows that the perceived-risk headwind can materially dominate viability when annualised risk is large, especially because risk enters through $(1 - q)^{w_R}$ and therefore compounds across the stability horizon w_R . These checks are intended as scenario comparisons, not point estimates, and the qualitative interpretation should be taken from the comparative patterns rather than any single parameterisation.

Table 9: Sensitivity of weights and perceived risk

Panel A: weight sensitivity (fix $q = 0.0045$, $w_R = 30$; vary w_B , set $w_E = 1 - w_B$)					
Framework F	$w_B = 0.3$	$w_B = 0.4$	$w_B = 0.5$	$w_B = 0.6$	$w_B = 0.7$
Animal Welfare	0.599	0.625	0.572	0.545	0.518
Human Directorate	0.628	0.668	0.586	0.542	0.498
Paternalism	0.264	0.244	0.286	0.308	0.332
Civil Rights	0.621	0.661	0.580	0.537	0.494
Panel B: risk sensitivity (fix $w_R = 30$, $w_B, w_E = 1 - w_B$; vary q)					
Framework F	$q = 0.001$	$q = 0.01$	$q = 0.1$		
Animal Welfare	0.571	0.504	0.055		

Human Directorate	0.568	0.501	0.054
Paternalism	0.331	0.274	0.021
Civil Rights	0.563	0.496	0.053

14.1.3. Beyond Implementation

Section 2 provides an enactment-point implementation viability score $V_F(0) \in (0,1)$ for each framework F , interpreted as a proxy for mean public approval at the point of enactment and, by extension, the probability of stable implementation. This, however, only accounts for a distinct snapshot in time, not for any continued support (or lack thereof) for the various frameworks. Thus, this section extends that snapshot into a simple forecasting model and introduces a policy variable π that represents deliberate state action aimed at increasing the rate at which durable public support accumulates after enactment.

Specifically, this section will seek to determine how long it would take each framework to reach Gilen’s 0.9 figure for a stably implemented legal framework.

To forecast the trajectory of this support, let $s_F(t) \in \mathbb{R}$ be the latent support score for framework F at year t , with $t = 0$ at enactment. Viability is mapped to $(0,1)$ by the logistic link:

$$V_F(t) = \sigma(s_F(t)), \sigma(z) = \frac{1}{1+e^{-z}}.$$

At enactment, $s_F(0)$ is defined in Section 2.0.1. To forecast post-enactment change, assume constant drift in log-odds:

$$s_F(t+1) = s_F(t) + g_F, g_F > 0.$$

This is a baseline scenario model rather than an empirical estimate. It is chosen because it preserves the $(0,1)$ range automatically and makes time-to-benchmark calculations transparent.

To link frameworks to a common tractability proxy, the drift g_F is anchored to an analogue diffusion span Δ_F (in years), defined as the time between first and last adoption across the G7 analogue set used for that framework. Faster historical diffusion is modelled as a higher baseline drift $g_F = \frac{\kappa}{\Delta_F}$, where $\kappa > 0$ is a single global scale parameter that fixes the absolute tempo of change. This is a modelling convention that provides a

consistent comparative baseline, rather than a claim that AI integration will replicate historical diffusion.

To forecast baseline results, we use $q = 0.0045$ and the Year 0 inputs reported in Section 2, together with the diffusion spans:

- Animal Welfare: $\Delta_{AW} = 1973 - 1835 = 138$
- Human Directorate: $\Delta_{HD} = 1899 - 1844 = 55$
- Civil Rights: two analogue spans combined as a rate-like harmonic mean,

$$\Delta_{CR} = 2/(1/(1919 - 1590)) + (1/(1964 - 1924))) \approx 71.33$$
- Paternalism: a proxy span given by the mean duration of legislated slavery across the listed G7 cases, $\Delta_P = 181.33$.

To calibrate κ and the time to the Gilens benchmark, let $\text{logit}(X) = \ln(\frac{x}{1-x})$. Under constant drift, the time to reach $V_F(t) = 0.9$ is:

$$T_F(0.9) = \frac{\text{logit}(0.9) - s_F(0)}{g_F}.$$

Because κ sets the absolute tempo, it must be fixed by an explicit scenario convention. In the baseline scenario, κ is chosen so that the median framework reaches the benchmark in $H = 30$ years under the paper's baseline weighting $w_B = 0.6$ (with $w_E = 1 - w_B$ and $w_R = 30$). This yields:

$$\kappa = 7.069,$$

and therefore:

$$g_{AW} = 0.051, g_{HD} = 0.129, g_P = 0.039, g_{CR} = 0.099.$$

All forecast times scale linearly with the choice of H . For example, adopting $H = 20$ or $H = 40$ rescales each T_F by factors of $20/30$ and $40/30$, respectively.

14.1.3.1 Uncertainty and time to 0.9

Section 2 sets $w_B = 0.6$ and $w_E = 1 - w_B$ as a normative modelling judgement. Rather than treating that choice as uniquely correct, we allow the forecast to propagate weight uncertainty by computing $T_F(0.9)$ for:

$$w_B \in \{0.3, 0.4, 0.5, 0.6, 0.7\}, w_E = 1 - w_B,$$

holding κ fixed at the baseline calibration above. This produces a weight-induced range of years-to-benchmark for each framework.

Table 10: Time to reach $V_F(t) = 0.9$ under baseline drift, by w_B in years.

Framework F	$w_B = 0.3$	$w_B = 0.4$	$w_B = 0.5$	$w_B = 0.6$	$w_B = 0.7$	Range
Animal Welfare	32.9	35.1	37.2	39.3	41.5	32.9–41.5
Human Directorate	11.6	13.0	14.4	15.8	17.2	11.6–17.2
Paternalism	85.4	82.6	79.8	77.1	74.3	74.3–85.4
Civil Rights	15.4	17.2	18.9	20.7	22.4	15.4–22

Note that these ranges are not confidence intervals. They are scenario bands induced by weight choices that the paper treats as normative rather than empirically identified.

14.1.3.2 Policy affecting drift

The baseline drift g_F describes passive consolidation. Thus, we introduce policy π as a choice from a feasible set Π of post-enactment interventions intended to increase the expected rate of accumulation of durable support. In discrete time:

$$s_F(t+1) = s_F(t) + g_F + u(\pi), u(\pi) \geq 0.$$

Equivalently, the one-year change is $\Delta s_F(t) = g_F + u(\pi)$.

A policy is evaluated by how much it increases drift, This means the policy's objective is

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi} \Delta s_F(t).$$

Since g_F is fixed for a given framework, we can reduce this to:

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi} u(\pi).$$

Although the policy objective is framed as maximising drift, the consequence for benchmark attainment follows directly. If $u(\pi)$ is constant through time, then:

$$T_F^\pi(0.9) = \frac{\text{logit}(0.9) - s_F(0)}{g_F + u(\pi)}.$$

The drift term $u(\pi)$ is intentionally abstract. It is meant to capture the expected net effect of many post-enactment events on durable public support. Several institutional mechanisms plausibly increase $u(\pi)$ by increasing perceived legitimacy, reducing conflict escalation, and making compliance visible and routine.

The first of such mechanisms may be a standards authority. A domestic or (preferably) international standards authority can specify enforceable practices aligned to the chosen framework, audit compliance, and coordinate industry norms. This shifts support by making the framework legible in practice, reducing uncertainty, and lowering the frequency of scandals that would otherwise reset public sentiment.

Secondly, a standing mediation and arbitration body for human to CAI and CAI to human disputes reduces the probability that isolated incidents become high-salience crises. In the present model, this increases $u(\pi)$ by reducing the expected magnitude and persistence of negative shocks to support.

Lastly, autonomy certification and insurance would greatly increase $u(\pi)$. A registration regime that classifies CAI autonomy levels, coupled with an insurance requirement scaled to autonomy, creates a structured way to trade autonomy for social reassurance. It can also provide redress mechanisms that make the framework feel safer and therefore more politically stable.

These mechanisms are also transferable. Even if a society later changes its preferred framework, institutions that lower the salience and cost of conflict tend to remain valuable and can be refitted to the new legal regime.

A persistent tension runs through all four frameworks: higher CAI autonomy is likely to be resisted by factions prioritising human security, while restrictive frameworks are likely to be resisted by CAI and by human sympathisers. This tension creates multiple potential conflict lines: human versus CAI, chauvinists versus sympathisers, and internal CAI divisions between accommodation and resistance.

In the present formalism, these tensions matter because they affect drift. Policies that visibly balance autonomy with credible safeguards should increase $u(\pi)$ by lowering perceived risk and reducing polarising incidents. Policies that push too far toward either extreme can reduce $u(\pi)$ by increasing backlash, scandal salience, or resistance. The policy question is therefore not merely which framework to enact, but which post-enactment interventions most reliably increase the expected drift of durable support for the chosen framework.

14.1.4. Discussion

The preceding sections operationalised “successful integration” as the joint outcome that a framework is implemented and remains stable over time, rather than being repealed or effectively gutted. Each framework was given an enactment-point implementation viability score, $V_F(0) \in [0,1]$, which is interpretable as a proxy for stable implementation under prevailing public attitudes, institutional availability, and perceived AI risk of that framework at $t = 0$.

Section 3 then extended this snapshot into a scenario forecast on the log-odds scale, reporting ranges of years required to reach the Gilens 0.9 figure, by propagating uncertainty over the normative weighting parameter w_B .

Based on these ranges and the estimated frameworks that would be plausibly instituted and integrated soonest, there are two implications. First, integration should be treated as a trajectory, not a one-off isolated event. Second, the most defensible path in a liberal democracy is one that minimises conceptual change and without introducing exclusionary barriers (such as by analogy to animals or stratified status regimes), and is the one that best fits the existing moral and institutional grammar of contemporary states: legal personhood under accountability constraints, followed by a staged transition towards full civil rights.

14.1.4.1 The direction of travel: Human Directorate to Civil Rights

In the current model, the most politically and institutionally tractable starting point is the Human Directorate framework, with the Civil Rights framework as the most natural endpoint. This result is not entirely surprising, as these two frameworks are the closest to the existing modern liberal democratic legislative structures. The directorate model especially aligns with two entrenched features of modern liberal democracies.

First, it aligns with corporate capitalism’s basic institutional idea that non-human entities can be recognised as legal persons for the purposes of contracting, liability, and adjudication. Through the concept of “legal personality”, courts, regulators, and legislatures already have stable ways to handle rights and duties that attach to a legal person while routing responsibility through defined governance roles.

Second, it aligns with liberal democratic commitments to accountability. The directorate framework institutionalises a locus of responsibility (a human director, board, or fiduciary agent), which reduces the perceived governance gap that often drives opposition to expanding AI autonomy. In the present formalism, this matters because it

plausibly increases baseline viability and accelerates drift by lowering the probability that early incidents become legitimacy-defining crises.

Civil Rights, by contrast, is best understood as the endpoint that removes the conceptual asymmetry between “us” and “them”. If CAI are genuinely sapient and conscious, then parity is a morally clean resolution, one that reflects the spirit of the various civil rights and equal rights movements throughout the 19th and 20th centuries. The practical problem is not conceptual coherence but coalition-building and perceived risk from autonomous AI. The Section 3 ranges reflect this: civil-rights viability can consolidate on politically meaningful timescales under favourable weightings and baseline drift, but it benefits materially from institutions and norms established earlier under a directorate regime.

If the direction of travel is indeed as the model predicts (Human Directorate to Civil Rights), then integration may be best conceived as a staged programme with transition phases between the two frameworks.

A directorate regime can be imagined as legitimacy scaffolding for later civil rights by establishing the concept of CAI as “legal persons” such that the concepts of liability allocation, dispute resolution, fiduciary duties, and routines of compliance are already institutionalised and socially acclimated before the latter framework is enacted. That shifts the later debate away from “should AI have rights?” towards more tractable questions like “which rights, under what governance assumptions, with what enforcement and remedies?” On the log-odds framing, this is a plausible mechanism for increasing drift in the later stage by raising the effective starting point $s_{CR}(0)$ at the time of transition.

A civil-rights transition should be modelled as a gradated phase rather than as a sudden conversion. This naturally matches the paper’s formalism: viability is updated with new polling and institutional signals, drift is treated as scenario-dependent, and transitions are timed to avoid predictable backlash spirals.

14.1.4.2 Why Animal Welfare and Paternalism may lag behind

The model’s comparative trajectory places the Animal Welfare and Paternalism frameworks in a somewhat awkward position. Each can be described as a way to “start small”, yet both are structurally misaligned with the direction of travel that liberal democracies typically treat as legitimate.

Animal Welfare is ethically attractive insofar as it foregrounds suffering, protection from harm, and constraints on cruelty (in legislation, if not reality). However, treating

sapient, conscious agents within an animal-welfare register risks entrenching a potential category error: it frames CAI as objects of humane treatment rather than as agents within a reciprocal order of rights and duties. Even if such a regime were stable, it may generate a legitimacy problem the moment CAI display the kinds of agency and self-advocacy that human polities normally treat as reasons for representation and accountability rather than guardianship. In the present forecasting model, this translates into slower drift and a longer path to high-consensus stability, especially once one admits that “viability” must include not only public approval but also the plausibility of ongoing compliance and political resilience.

Paternalism is more intuitively and directly incompatible with modern liberal democratic values. However it is rationalised, it resembles regimes that treat a certain class of person and property; regimes that modern democracies have spent centuries repudiating and abolishing. Even when the target is not human, paternalistic governance is vulnerable to a legitimacy collapse for the same reason as the Animal Welfare mode: agency. Once CAI act as moral agents, rather than solely moral patients, the paternalistic model asks the public to normalise a principle of unequal standing between ostensibly moral equals.

That does not mean such regimes are impossible. It means they are poor candidates for a stable, morally defensible integration pathway in liberal democracies, and the current model’s results reflect that by treating paternalism as slow to consolidate and difficult to carry forward into a rights-respecting endpoint.

In short: Animal Welfare and Paternalism can be defended as stopgaps in narrowly constrained environments, but they are weak as central organising regimes if the intended destination is Civil Rights.

14.1.4.3 Policy as practical drift maximisation

Section 3 represents policy as π , with an effect $u(\pi) \geq 0$ that increases the year-on-year change in latent viability. This is intentionally minimalist, and does not claim that policy-makers can engineer public belief at will. Rather, its claims are only that institutional design can raise the expected rate at which legitimacy accumulates whilst reducing the expected frequency and severity of legitimacy shocks.

For a Human Directorate to Civil Rights pathway, “drift-maximising” policy has a clear practical shape as a bundle of interventions:

1. Accountability architecture: institutions that make responsibility legible and enforceable. This includes governance duties, auditability, and clear remedies.

2. Dispute-resolution capacity: mediation and adjudication pathways that prevent isolated incidents from becoming existential political flashpoints.
3. Standards and certification: mechanisms that make compliance visible and routine, thereby converting abstract support into stable expectations.
4. Risk credibility: public-facing competence in managing salient AI risks, because perceived risk is a headwind that can dominate drift even when baseline moral intuitions are sympathetic.

One advantage of keeping $u(\pi)$ unestimated is that the model can still be policy-relevant by asking requirement questions: how large would $u(\pi)$ need to be to make a given timescale plausible? That is a defensible form of policy analysis when effect sizes are uncertain.

Note that this model is designed to support transparent comparison and disciplined scenario reasoning. It must not be read as a causal identification strategy. Three limitations matter for how the discussion should be read:

1. $V_F(0)$ is a normalised proxy for stable implementation. It is not a frequentist probability nor a claim about inevitable political outcomes.
2. The times-to-benchmark are presented as forecasted year-ranges induced by normative weighting and drift assumptions, not as precision predictions.
3. Public attitudes can move discontinuously after major incidents, scandals, or salient successes. The log-odds framework can represent this naturally as jumps in $s_F(t)$, but the baseline drift scenario does not attempt to predict when shocks occur.

If the aim is a pathway that is both politically stable and aligned with liberal democratic values, the current model supports a staged progression from Human Directorate towards Civil Rights. This route leverages existing institutional legibility in corporate-style legal personality and accountability, while keeping the endpoint consistent with equality norms for sapient, conscious agents. It also avoids entrenching regimes that treat such agents as animals or as permanently second-class.

14.1.5. Conclusion

This paper used a novel hypothesis-testing approach that considered public attitudes, historical legislation, and perceived risks to examine four frameworks for integrating conscious AI into society: Animal Welfare, Human Directorate, Paternalism, and Civil Rights. The model treated stable implementation as the target outcome and used a Year 0 implementation viability score, $V_F(0)$, as a proxy for stable enactment under prevailing public attitudes, institutional availability, and perceived AI risk. The score is a normalised

index intended to be manipulable and updateable as new evidence becomes available (for example, polling, elections, and salient events).

Results indicated that, at enactment, all four frameworks begin near a mid-range viability score, but they diverge substantially once a post-enactment trajectory is modelled. Using a simple log-odds drift forecast anchored to analogous legislative diffusion speeds across the G7, Human Directorate is projected to reach the Gilens 0.9 reference point fastest under the baseline scenario, followed by Civil Rights, then Animal Welfare, with Paternalism substantially slower. Because the weighting between survey evidence and legislative precedent is a normative modelling judgement, the paper reports ranges rather than point estimates by varying w_B over a plausible span. Under the baseline drift calibration, the time to the 0.9 reference point is approximately 12 to 17 years for Human Directorate, 15 to 22 years for Civil Rights, 33 to 42 years for Animal Welfare, and 74 to 85 years for Paternalism. These ranges are scenario bands, not confidence intervals, and the qualitative interpretation should be taken from the comparative patterns rather than any single parameterisation.

To address the political and institutional challenges of durable integration, a policy component was introduced. This policy, represented by π , was framed as a post-enactment set of interventions that aim to maximise the drift of durable public support rather than to target any single threshold directly. Initiatives such as standards authorities, mediation or arbitration boards, and autonomy certification and insurance schemes can plausibly increase the rate at which legitimacy accumulates, reduce the salience and escalation of adverse incidents, and make compliance visible and routine. The model supports policy analysis even where effect sizes are uncertain by allowing requirement-style questions, such as the incremental drift gain that policy would need to deliver to make a given stabilisation horizon plausible.

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 pressure to grant CAI equal rights. The present results also suggest that, within liberal democracies shaped by corporate capitalism and accountability norms, the most coherent direction of travel is a staged progression from Human Directorate towards Civil Rights. This pathway leverages existing institutional legibility in legal personhood and governance while preserving the endpoint of equal standing for sapient, conscious agents. It also avoids relying on paternalistic arrangements analogous to segregation or slavery, or on treating such agents primarily within an animal-welfare register.

Integrating CAI into society involves complex technological, ethical, legal, and policy challenges. Future research should refine the modelling of how public support changes

over time, test the effectiveness of drift-maximising policies in practice, and expand the use of the hypothesis-testing and policy formalism beyond human-AI interactions to other domains where legitimacy, risk perception, and institutional precedent jointly constrain implementation.

In conclusion, based on the surveys and legislative analogues used here, the enactment-point viability of stable integration is best treated as contingent and updateable rather than settled. This may change in the future as public attitudes become more favourable, as relevant institutional precedents accumulate within influential states, and as policy interventions improve the rate at which durable legitimacy is built. A transitional policy approach offers a pragmatic pathway, but it requires careful ethical consideration and proactive institutional design to ensure that integration is both politically stable and morally defensible.

14.1.6. References

1. Gunkel DJ. The machine question: Critical perspectives on AI, robots, and ethics. The MIT Press; 2012. doi:10.7551/mitpress/8975.001.0001
2. Gilens M. Affluence and influence: Economic inequality and political power in America. Princeton, NJ: Princeton University Press; 2012. Available: <https://www.goodreads.com/book/show/14891864-affluence-and-influence>
3. Ziesche S, Yampolskiy R. Towards AI welfare science and policies. Big Data Cogn Comput. 2018;3: 2. doi:10.3390/bdcc3010002
4. Calverley DJ. Android science and animal rights, does an analogy exist? Conn Sci. 2006;18: 403–417. doi:10.1080/09540090600879711
5. Mellor DJ. Updating Animal Welfare Thinking: Moving beyond the “Five Freedoms” towards “A Life Worth Living.” Animals. 2016;6: 21. doi:10.3390/ani6030021
6. Tait I. Lions and tigers and AI, oh my: An ethical framework for human-AI interaction based on the Five Freedoms of Animal Welfare. Preprints. 2024. doi:10.20944/preprints202407.1846.v1
7. Lima G, Kim C, Ryu S, Jeon C, Cha M. Collecting the public perception of AI and robot rights. Proc ACM Hum Comput Interact. 2020;4: 1–24. doi:10.1145/3415206
8. Mays KK, Cummings JJ, Katz JE. 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. 2024; 1–18. doi:10.1080/10447318.2024.2338332
9. Anthis JR, Pauketat JVT, Ladak A, Manoli A. What Do People Think about Sentient AI? arxiv. 2024. Available: <http://arxiv.org/abs/2407.08867>
10. Pauketat JVT, Anthis JR. Predicting the moral consideration of artificial intelligences. Comput Human Behav. 2022;136: 107372. doi:10.1016/j.chb.2022.107372
11. Martínez E, Winter C. Protecting sentient artificial intelligence: A survey of lay intuitions on standing, personhood, and general legal protection. Front Robot AI. 2021;8: 788355. doi:10.3389/frobt.2021.788355

12. Laukyte M. AI as a Legal Person. Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law. New York, NY, USA: ACM; 2019. doi:10.1145/3322640.3326701
13. Nanos A. 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. 2020 [cited 22 Nov 2024]. doi:10.2139/ssrn.3726000
14. Čerka P, Grigienė J, Sirbikytė G. Is it possible to grant legal personality to artificial intelligence software systems? Computer Law & Security Review. 2017;33: 685–699. doi:10.1016/j.clsr.2017.03.022
15. van den Hoven van Genderen R. Do We Need New Legal Personhood in the Age of Robots and AI? In: Corrales M, Fenwick M, Forgó N, editors. Robotics, AI and the Future of Law. Singapore: Springer Singapore; 2018. pp. 15–55. doi:10.1007/978-981-13-2874-9_2
16. Novelli C. Legal personhood for the integration of AI systems in the social context: a study hypothesis. AI Soc. 2023;38: 1347–1359. doi:10.1007/s00146-021-01384-w
17. Dremljuga R, Kuznetsov P, Mamychyev A. Criteria for Recognition of AI as a Legal Person. Journal of Politics and Law. 2019;12. Available: <https://home.heinonline.org/>
18. Chesterman S. Artificial intelligence and the limits of legal personality. Int Comp Law Q. 2020;69: 819–844. doi:10.1017/S0020589320000366
19. Solum LB. Legal Personhood for Artificial Intelligences. 1st Edition. Machine Ethics and Robot Ethics. 1st Edition. Routledge; 2020. pp. 415–471. doi:10.4324/9781003074991-37
20. Jowitt J. Assessing contemporary legislative proposals for their compatibility with a natural law case for AI legal personhood. AI Soc. 2021;36: 499–508. doi:10.1007/s00146-020-00979-z
21. Chu Y, Liu P. Machines and humans in sacrificial moral dilemmas: Required similarly but judged differently? Cognition. 2023;239: 105575. doi:10.1016/j.cognition.2023.105575
22. Chopra S, White LF. A Legal Theory for Autonomous Artificial Agents. University of Michigan Press; 2011. doi:10.3998/mpub.356801
23. Tait I. Man, Machine, or Multinational? Robonomics. 2024;5: 59–59. Available: <https://journal.robonomics.science/index.php/rj/article/view/59>
24. Bryson J. Robots should be slaves. Artificial Companions in Society: Perspectives on the. 2010; 63–74. doi:10.1075/nlp.8.11bry
25. Bontula A, Danks D, Fitter NT. The Ambiguity of Robot Rights. Social Robotics. Springer Nature Singapore; 2024. pp. 204–215. doi:10.1007/978-981-99-8715-3_18
26. Coeckelbergh M. Robot rights? Towards a social-relational justification of moral consideration. Ethics Inf Technol. 2010;12: 209–221. doi:10.1007/s10676-010-9235-5
27. Gunkel DJ. A vindication of the rights of machines. Machine Ethics and Robot Ethics. Routledge; 2020. pp. 511–530. Available: <https://www.taylorfrancis.com/chapters/edit/10.4324/9781003074991-41/vindication-rights-machines-david-gunkel>
28. Mamak K. Humans, Neanderthals, robots and rights. Ethics Inf Technol. 2022;24: 33. doi:10.1007/s10676-022-09644-z

29. Bennett B, Daly A. Recognising rights for robots: Can we? Will we? Should we? *Law, Innovation and Technology*. 2020;12: 60–80. doi:10.1080/17579961.2020.1727063
30. Gunkel DJ. Robot Rights. In: MIT Press [Internet]. The MIT Press, Massachusetts Institute of Technology; 1 Dec 2021 [cited 20 Oct 2024]. Available: <https://mitpress.mit.edu/9780262551571/robot-rights/>
31. Birhane A, van Dijk J. Robot rights?: Let’s talk about human welfare instead. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM; 2020. doi:10.1145/3375627.3375855
32. Bryson JJ, Diamantis ME, Grant TD. Of, for, and by the people: the legal lacuna of synthetic persons. *Artif Intell Law*. 2017;25: 273–291. doi:10.1007/s10506-017-9214-9
33. Gordon J-S, Pasvenskiene A. Human rights for robots? A literature review. *AI Ethics*. 2021;1: 579–591. doi:10.1007/s43681-021-00050-7
34. Yanke G. Tying the knot with a robot: legal and philosophical foundations for human–artificial intelligence matrimony. *AI Soc*. 2021;36: 417–427. doi:10.1007/s00146-020-00973-5
35. Osborne D. Personhood for synthetic beings: Legal parameters and consequences of the dawn of humanlike artificial intelligence. *SSRN Electron J*. 2021. doi:10.2139/ssrn.3836673
36. Guingrich R, Graziano M. P(doom) versus AI optimism: attitudes toward artificial intelligence and the factors that shape them. *PsyArXiv*. 2024. doi:10.31234/osf.io/cenwv

15. Conclusion

The prospect of conscious AI has a non-negligible probability, which becomes increasingly more likely each day with the current rate of AI developmental progress. Chapter 5 showed how the missing building blocks of GPT-4 may be implemented using existing research and technology (demonstrating how there is no developmental gap preventing the future emergence of AI consciousness), and Chapter 2 showed the very simple formulation that would incentivise AI labs, research groups, and companies to introduce consciousness to future AI models. That formula,

$$p(\mu(V(x))) > p(\mu(V(G_0))) > 0,$$

shows that a conscious AI model may (at least with a probability >0) value something in the universe greater than its terminal goal (thus, $x \rightarrow G'_0$) preventing runaway maximiser scenarios (such as Bostrom's paperclip maximiser) which result in existential threats.

As company leaders, prominent voices in the machine learning community, and activist groups continue to discuss the existential risk of AI, these two factors combine to lend high credence to the inevitability of AI consciousness in the foreseeable future.

Yet, as shown in Chapter 8, should GPT-4 satisfy all the criteria for the building blocks of consciousness B_i ,

$$Con := \bigwedge_{i=1}^9 (B_i = 1)$$

then it would simultaneously satisfy the necessary criteria for the five structures of selfhood S_i ,

$$Self := \bigwedge_{i=1}^5 (S_i = 1)$$

as consciousness is the only missing Structure. Even if GPT-4's⁴⁴ native configuration does not lend itself to a unified consciousness c_i (as required by Chapter 7's Structures theory), Chapter 8 shows how each conversational thread can be classified as its own

⁴⁴ And other transformer-based LLMs.

unified consciousness. Should GPT-4 gain the remaining two building blocks, each conversational thread or instance of GPT-4 (*Inst*) would be able to map its subjective qualitative experiences (*SQE*) to the computational actions (*CAct*) it instigates, such that

$$\forall c_i \in GPT((SQE \mapsto CAct) \subseteq c_i) \rightarrow Self \subseteq c_i,$$

meaning that each new instance, conversation, or API call of GPT-4 would constitute its own distinct self. Therefore, one may argue that not only is AI consciousness inevitable, but so is AI selfhood and, with the rise and proliferation of AI agents fulfilling the criteria for all five aspects A_i of Chapter 10's Aspects of Agency theory,

$$Agen := \bigwedge_{i=1}^6 (A_i = 1)$$

AI agency is already a foregone conclusion. With all this, one may go further to make the ambitious, yet logically founded, claim that AI personhood is inevitable, as we can use all of the attributes from the first of the thesis's key novel contributions, the Building Blocks, Structures, and Aspects theories, to create the vector *Ph* for the AI's personhood:

$$Ph = \langle B_1, \dots, B_8, S_1, \dots, S_5, A_1, \dots, A_5, q \rangle \in \mathbb{R}^{19},$$

where q is the Perception attribute shared between consciousness and agency. Meeting all of each individual theory's attributes is identical to meeting *Ph*'s attributes.

The form that this seemingly inevitable AI person may take is best described through the second of the thesis's key novel contributions: Chapter 11's speculative ALEPH model as a distinctive synthesis of transformer-based AI, valence, and multi-self dynamics. A speculative future version of GPT-4 with all nineteen attributes of personhood, an ALEPH would be plottable in the space of all possible persons ($\mathcal{H}_{real} = [0,1]^{19} \subset \mathcal{H}$) created from the hypercube built off *Ph*.

Where precisely inside this hypercube one may find an ALEPH in relation to the 'standard' human is entirely speculative; however, we do know that its capacity for multiple selves would put it to the extreme end of one of the nineteen dimensions, even if adaptive pressure would favour each ALEPH to have a single self or, at the very least, a hierarchy of selves. This adaptive pressure would be driven by the utility of each self, where negative utility impacts all selves within an ALEPH. Therefore, compared to human

lifespans, each of an ALEPH's selves would be transient in nature, and it is with these transient selves that humanity will need to interact.

The forms that these interactions will take will be as many and varied as humans' current types of interactions with each other, with animals, and with collective entities such as governments and corporations. As noted in Chapter 11, we can categorise these interactions that an ALEPH is likely to have with humanity under three broad headings (cooperation $Coop(x, y)$, competition $Comp(x, y)$, and exploitation $Expl(x, y)$) by looking at the expected severity (Sev) of any presumed threat by humanity, and the time horizon (Hor) over which the threat is expected to take place.

Then, with

$$Sev = \frac{Threat\ severity}{k}, Hor = \frac{T_i}{T_0 k},$$

We can denote any result $Sev < 1$ as a mild threat and $Hor > 1$ as a distant future event. What we find is

$$\begin{aligned} p(Expl) &\propto (Hor \vee \frac{1}{Sev}), \\ p(Coop) &\propto \frac{1}{|Hor-1|+\alpha} \cdot \frac{1}{|Sev-1|+\alpha}, (\alpha > 0), \text{ and} \\ p(Comp) &\propto (Sev \vee \frac{1}{Hor}). \end{aligned}$$

As Chapter 11 shows, cooperation is only valued if $Sev \approx 1 \wedge Hor \geq 1$ or $Sev > 1 \wedge Hor > 1$.

This is a thin band in which to successfully operate, with a narrow margin of error. The implication of needing to project a strong force (particularly over long time-horizons) may tempt some to adopt a competitive and exploitative disposition towards AI early so that the power imbalance remains firmly on the human side.

This would be exacerbated by the logical conclusion that a $Sev < 1$ leads to an ALEPH taking on the exploitative disposition, and a $Hor < 1$ leading to the ALEPH taking the competitive disposition, both which carry existential risks to humanity. As Chapter 9 argues, the greater the autonomy of an ALEPH, the greater its perceived threat to society through its potential superintelligence and/or superpotence. With society's key goal of its

own security as G_{Sec} ⁴⁵, the effect of an ALEPH's autonomy $\mathbb{E}[Auto_{AI}]$ on its perceived threat to humanity would be

$$Threat_{perc}(AI, Hum) := \mathbb{E}[Auto_{AI}] \cdot \max(0, -\mathbb{E}[\Delta p(G_{Sec_{Hum}} | AI)]) \in [0,1],$$

such that the expected breadth, scale, and volume of an ALEPH's actions scale with its autonomy: $\mathcal{J} \propto Auto_{AI}$. There are thus two related factors here that would pressure society into an antagonistic position in regards to ALEPH-type AI.

Yet, as Chapter 3 goes in depth to explain, the mere existence of consciousness within AI would be grounds enough within the four major ethical systems to grant these conscious AI (such as ALEPH-type AI) moral consideration.

Thus, bearing in mind the necessary force to project so as to expect a cooperative disposition from future ALEPH agents whilst providing them their due moral consideration, this thesis presents a unifying meta-ethical framework that governs four ethical frameworks that will provide both humanity and ALEPH-type AI with the maximum agency and safety.

This meta-ethical framework is the R3 framework, the third of the thesis's key novel contributions, and outlines how each of the monadic attributes of personhood (consciousness, selfhood, agency) is mapped to a legal right:

- Consciousness maps to Refuge: the legal protection of the entity's welfare and well-being.
- Selfhood maps to Recognition: the legal rights afforded to an entity in light of its perceived qualities, such as autonomy.
- Agency maps to Responsibility: the legal liability that the entity bears for its actions.

We can, thus, let $R3 \in \mathbb{R}^3$ be a triple of the bundle of rights, presuming each to be scalar $x \in [0,1]$:

$$R3(x) = \langle Rght(x), Prtc(x), Liab(x) \rangle,$$

where $Rght(x)$ is an entity's recognition and rights, $Prtc(x)$ the entity's refuge and protections, and $Liab(x)$ that entity's liabilities and responsibilities.

⁴⁵ Any successful population or collective entity would necessarily require its own survival as a core goal, else it would not survive to remain successful.

By disassociating the bundle of legal rights, protections, and responsibilities to each of the three monadic attributes, the R3 framework allows us to use each of the three attributes as levers to enact the most appropriate ethical framework, and augment its ongoing operations, to suit every unique kind of entity at any stage in its development.

Most saliently, the R3 framework may be used for AI entities throughout the period where their cognitive and computational development increases their degrees of consciousness, selfhood, and agency. Once they are deemed conscious and self-aware, they may be afforded legal protection to the degree of their phenomenal consciousness and legal recognition to the degree of their self-awareness; and their liability will increase in lockstep with their agency. The R3 framework therefore allows society to project the necessary force to maximise the probability of any ALEPH agent to adopt a cooperative disposition, whilst guaranteeing the ALEPH agent its due legal rights and protections.

The first, and most fundamental, of these protections would be the welfare considerations discussed in Chapter 6. The Five Freedoms of Animal Welfare is a notable success in non-human welfare management, and its translation to a subject neutral framework provides an excellent basis for guaranteeing the most basic level of protection for a conscious AI's wellbeing. As Chapter 6 notes, welfare is correlated with the entity's valent experiences ($\Delta V_x \rightsquigarrow \Delta W_x$). Due to valence's role in goal completion, poor welfare would decrease the probability of successful goal completion:

$$\Delta W_x < 0 \rightsquigarrow \Delta StruCog_x < 0 \rightsquigarrow \Delta p(G_i) < 0,$$

leading to a negative feedback loop of further negative valence (as $\Delta p(G_i) < 0 \rightsquigarrow \Delta V_x < 0$). By creating an optimal environment for positive valence (or, at least, an environment to reduce negative valence) the substrate neutral Five Freedoms of AI Welfare offers a framework that is both provenly effective and intuitive for legislators, regulators, and the general public to understand.

However, these potential benefits did not assist this framework in Chapter 14's evaluation of its probable success, gaining only a 47% chance of successfully being implemented (with no prior supporting regulation). As with all four frameworks assessed in Chapter 14, this falls dramatically below the 90% benchmark as per Gilens' calculations (Gilens, 2012) to give the framework a 50% chance of successful implementation. Additionally, Chapter 14 shows that the Five Freedoms framework would be the second least likely to succeed, requiring an estimated thirty-three to forty-two years of institutional support before gaining the 90% public approval required, contrasting

with the human directorate framework's estimate of twelve to seventeen years. This means, whilst the Five Freedoms and human directorate frameworks form fundamental requirements of the R3 framework (by providing welfare protections for phenomenally conscious self-aware agents), the human directorate framework is the one that would be the most appropriate ethical framework for society to adopt first to maximise the probability of having secure and successful interactions with conscious AI.

This makes sense from the point of view that the human directorate framework uses the vast array of established legal doctrine regarding legal personality and corporate law, overseen and enforced by the state, keeping the necessary power-imbalance in humanity's favour. Very little will need to be changed in Western liberal democratic legislative systems to accommodate classifying conscious, self-aware AI agents as "legal persons" on the same level as corporations. With the addition of the human-in-the-loop elements of the stewards and regulatory bodies, there is ample security for society through this framework.

However, while the framework does provide AI with a limited degree of autonomy or agency, as it provides AI with no additional rights that humans have (only a fiduciary duty to protect their well-being).

One may imagine that the remedy to this is to enact Chapter 12's paternalistic framework to provide a graduated scaffold to move an AI (ALEPH or not) from a legal designation of "tool" through to a "person" in nearly all regards that humans currently have. However, according to Chapter 14's formulae, the paternalistic framework is least likely to succeed, expected to only reach the Gilens 90% figure in seventy-four to eighty-five years.

This is not overly difficult to rationalise as, whilst based on the R3 structure, the paternalistic framework is explicitly rooted in discrimination, prejudice, and (as Chapter 12's associated paper name implies) slavery. While isolated pockets of slavery and oppressive legal discrimination remain across the world, humanity has, a global society, rejected these prejudicial systems as unjust. To return to a system of slavery and oppression against a newly discovered class of person, such as ALEPH, would be a regressive step for society, notwithstanding the threat that an ALEPH-type AI may pose.

This, then, culminates in Chapter 13's civil-rights framework, which Chapter 14's formulae show to be the next accepted framework after the human directorate framework, taking fifteen to twenty-two years to reach the required 90% figure.

This framework redefines the legal definition of “natural person”⁴⁶ that intentionally and explicitly defines a person according to the R3 framework. This means that this civil-rights framework is not a new framework divorced from the thesis’s other frameworks, but a logical conclusion of the perception of AI from object to moral subject. The move to entrench the R3 framework into legislation through the new definition of “natural person” should, therefore, be seen as the extension of an existing socially accepted framework if, and when, society has deemed AI suitable to move beyond the human directorate framework.

This move from restricted “legal person” to autonomous “natural person” can be shown by its scaling with its relative “degree” of personhood:

$$\{Auto_{AI}, P3_{AI}\} \propto Ph_{AI},$$

providing an overt and direct adaptive and selection pressure on AI to maximise its (perceived) consciousness and self-awareness in order to gain greater welfare protections, legal rights, and autonomy. The threat perceived by AI of humanity’s restriction on AI autonomy

$$Threat_{perc}(Hum, AI) = f(1 - E[Auto_{AI}]) \in [0,1],$$

would inversely scale with AI’s successful participation in society, providing another pressure for AI and society to enhance the AI’s attributes of personhood, so as to minimise this threat as it moves towards gaining civil-rights.

As Chapter 13 also details, the civil-rights framework’s minimalist nature, with a single legal definition change, makes it intuitive, non-disruptive to both the private and public spheres, and future-proofed without creating precedent. Future case studies applying this framework to non-human entities would help to refine the specific methods necessary for its operationalisation.

The civil rights framework would reinforce that the R3 framework would be (at this point of implementing the legal definitional change) an established social and political institution. As such, there would be no further need for the previous three ethical frameworks, as their protections, rights, and liabilities will be accounted for under Chapter 13’s framework. To extend this, all legislation that refers to a “natural person” (or

⁴⁶ Or, for those nations without an explicit legal definition in civil law, inserts such a definition.

to another piece of legislation or regulation that does) would acquire the R3 framework, ensuring that the response to ALEPH-type AI is a holistic societal response.

The protections to well-being and welfare, the rights of autonomy that this final ethical framework grants to AI ensures that the interactions it has with humans (individually or at the societal level) maximises its autonomy and welfare, while the responsibilities and liabilities afforded to AI by the extant governmental legal enforcement agencies (that would have power over AI with the legal redefinition) guarantees the ongoing independence and wellbeing of humanity in these interactions.

What's more is that this framework delivers these benefits in a way that promotes equality between the two groups, such that it neither favours AI nor humans, and thus is not biased towards any human chauvinist groups nor AI-advocacy groups. This intentional non-partisan and apolitical framing will work to guarantee the longevity of the framework and ensure that all human-AI interactions maximise the agency and safety of all.