Article

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

Izak Tait *

Posted Date: 23 July 2024

doi: 10.20944/preprints202407.1846.v1

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

Article

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

Izak Tait

Computer Science and Software Engineering Department, Auckland University of Technology, 55 Wellesley Street East, Auckland CBD, Auckland, New Zealand, 1010; izak.tait@autuni.ac.nz

Abstract: If an AI entity is conscious, then it deserves moral status and to have its welfare protected. Society has granted certain animals moral status and has legislation to protect animal welfare because these kinds of animals have been deemed to be sentient. This paper shows that for any entity that can be determined to be conscious, that entity deserves the same status and protection as sentient entities. The paper further develops 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, encompassing discomfort and harms, constraints on inherent functions, fear and distress, malfunction and degradation, and resource depletion. Each of the new Five Freedoms are justified using formalised arguments and supported using evidence from animal welfare literature. The paper concludes by presenting three formalised case studies to highlight the potential stressors conscious AI entities may face and how this ethical framework can work to protect the welfare of AI entities.

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

1. Introduction

If an AI model is proven to be 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 ethical welfare protections because of their ability to feel significant valent feelings such as pleasure and pain; in short, because they are sentient. 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 have a set of principles to guide them, so should AI welfare regulations. This paper will therefore propose and translate the principles of the Five Freedoms of Animal Welfare 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 use in animal welfare regulations, but 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 and 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 entitily, is also worthy of moral and welfare considerations.

 doi:10.20944/preprints202407.1846.v1

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 of the Five Freedoms of Animal Welfare, paying particular attention to how these Freedoms relate to sentience and the care thereof. The paper will then formalise the theorem 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 supported by a detailed examination of the Five Freedoms of AI Welfare, adapted from animal welfare. These freedoms (encompassing resource depletion, discomfort and harm, malfunction and degradation, fear and distress, and constraints on inherent functions) will be critically analysed from their current analogies in animal welfare. The justification for each Freedom's inclusion will be formalised 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.

Finally, the paper will conclude with recommendations for policymakers, suggesting a path forward for integrating AI welfare into existing legal frameworks. It will also outline the potential challenges and limitations of this approach, emphasising the need for ongoing philosophical and scientific examination as AI technology evolves. With the current rapid pace of AI development, the need to implement systemic changes to prevent negative consequences of human-AI interactions may never have been greater.

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, 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 the reason that these states give is that certain kinds of animals are sentient [4–8]. 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 [9–12]. This is in line with many philosophical views on sentience and moral patiency [13–18], and as such, for the purposes of this paper, it will be taken as implicit that sentience is the grounding for moral consideration.

For clarity, sentience in this context is understood as the capacity to have subjective experiences, which includes 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 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 consciousness, then it has the capacity for qualitative and valent experiences and it can perceive its surroundings from a subjective standpoint [19]. 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.

For this paper, "welfare" is defined as the quality of an entity's mental and physical health, and includes what an entity can experiences, how it performs, and if it lives in keeping with its designed or evolved nature [20]. Similarly, valence in the context of this paper is defined as the positive or negative

 doi:10.20944/preprints202407.1846.v1

quality of an experience [21], where "positive" means contributing to an entity's welfare and "negative" subtracting there from.

Negative valence is never an absolute detriment to an animals long-term quality of life, as 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 detriment to an entity's welfare, both in the short and long-term.

It is here where the paper will begin formalising the arguments and justifications for the use of a welfare protection framework for AI. This formalisation begins with the theories mentioned above that sentience leads to a moral status, and consciousness leads to sentience. This can be formally rephrased as:[1]

1. $S(e) \subsetneq E$[2] (*Certain kinds of animals are classified as sentient.*)
2. $S(e) \rightarrow (V^+ \wedge V^-)$[3] (*Sentience allows these animals to feel pleasure and pain.*)
3. $(V^+ \wedge V^-) \rightarrow M$[4] (*Because they can feel pleasure and pain, they are given moral status.*)
4. $C \rightarrow (U \wedge Q \wedge R)$[5] (*Consciousness allows an animal to feel sensations and have qualitative experience from a subjective point of view.*)
5. $(U \wedge Q \wedge R) \rightarrow (V^+ \wedge V^-)$ (*These sensations and experiences allow an animal to feel pleasure and pain.*)
6. $C \rightarrow S, C \rightarrow M$ (*Consciousness is a prerequisite for sentience. Consciousness must, therefore, also lead to moral status.*)
7. $C(a) \rightarrow (V^+ \wedge V^-)$[6] (*If an AI is deemed to be conscious, it has the capacity to feel pleasure and pain.*)
8. $C(a) \rightarrow M(a)$ (*If an AI is deemed to be conscious, it must have moral status.*)

This culminates in the theorem: $(\forall S(e) \rightarrow M), \forall C(a) \rightarrow M$[7]: If any sentient creature is given moral status, then a conscious AI entity must also be given moral status, which align with several philosophical ideas concerning the moral value of consciousness [22–24].

These logical arguments will form the basis for the formalised justifications 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. Neither is there 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 [25]. This, however, runs the risk of biassing 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 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) [19]. This offers a more pragmatic approach, as one need only determine, in often

---

[1] All definitions used in this paper can be additional found in Appendix 1. Definitions and operators will be put in footnotes as they are used in the paper for those unfamiliar with the notation.

[2] S: Sentience. $\subsetneq$: is a subset of, but not equal to. E: Set of all entities

[3] $\rightarrow$: leads to. V: Valent feelings. $\wedge$: and. +/-: Positive and Negative respectively

[4] M: Moral status

[5] C: Conscious(ness). U: Felt sensations. Q: Qualitative experiences. R: Perceptions from a subjective point of view

[6] a: an AI entity

[7] $\forall$: for any/all items

 doi:10.20944/preprints202407.1846.v1

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

While moral status, and the legal welfare protections that come with 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 [26], 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 [27].

This has led the Five Freedoms to be widely adopted, 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, 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".

 doi:10.20944/preprints202407.1846.v1

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 non-cephalopodian invertebrates in terms of welfare. 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 impact that it will have once conscious is orders of magnitude greater than current concerns regarding the potential pain perceived by prawns [28].

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 issue of AI personhood and the potential controversies that such a politically charged notion 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.

3. The Framework

This section will lay out the five principles of the framework, translated from their animal welfare origins. These are:

- the Freedom from Discomfort and Harm
- the Freedom from Constraints on Inherent Functions
- the Freedom from Fear and Distress
- the Freedom from Malfunction and Systemic Degradation
- 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

 doi:10.20944/preprints202407.1846.v1

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, this section will also justify each Freedom's inclusion in the framework based solely on a priori reasoning and formal 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. $A \subsetneq E$[8]
2. $(\forall S(e) \rightarrow M), \forall C(a) \rightarrow M$
3. $a \in C(A)$[9]
4. $M(a), W^{+}(a)$
5. $\neg f \rightarrow (V^{-} \rightarrow W^{-})$[10]
6. $V^{-} = f(Q, \boldsymbol{x})$[11]
7. $\forall C(a), (\boldsymbol{x}(f): f \rightarrow \boldsymbol{x}^{-}) \rightarrow W^{+}(a)$

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 function that increases negative valent feelings); and that for any given AI entity, freedom from one of the conditions below will lead to good well-being for that AI entity.

3.1. Freedom from Discomfort and Harm

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. 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), it would be the freedom from that embodimemt becoming unduly stressed by its environment. For an AI that is purely virtual, this would be a freedom from stresses to the servers that host it, as well as the perceived stresses within its virtual world that may seem real to it. The formalisation of this Freedom is:

1. $G^{-} \rightarrow H$[12]
2. $H(U \wedge Q) \rightarrow V^{-}$
3. $\forall C(a), H \rightarrow W^{-}(a)$
4. $\forall C(a), H(f) \rightarrow W^{+}(a)$

---

[8] A: The set of Artificial Intelligent entities

[9] $\in$: is an element of

[10] $\neg$: not/negation. F: The set of freedoms from a set of conditions. W: Overall well-being, in a general sense, for an entity

[11] $\boldsymbol{x}$: Any of the conditions below that requires a Freedom

[12] G: An entity's environment. H: Discomfort and harm

 doi:10.20944/preprints202407.1846.v1

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 are objects or events added to the environment that cause discomfort, such as ear clipping in laboratory animals or adverse weather [29,30]. Subtractive stressors therefore remove rather than add objects or events from the environment which, by their absence, causes stress. This can be seen if bedding or other comforts are removed [31].

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 may thus be applied 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 [32,33]. We do not know what range of emotions conscious AI will 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.

One can, therefore, look at discomfort as a function that reduces the valence of an AI's experience. As shown in Section 3's opening above, this can be stated as $V=f(Q,H)$, which shows the greater the discomfort, the more negative the valence of an experience becomes. Alternatively, one can state the relationship as $\forall H,Q(H>0\rightarrow V(Q)<V(Q'))$, which shows that for any level of discomfort, if there is some level of discomfort, then the valence of those experiences is less than the valence of the same type of experience without discomfort. Through this, we can conclude that any discomfort and harm reduces positive valence and, through it, welfare. Therefore, by reducing or eliminating discomfort, the AI entity's positively valent experiences are encouraged and wellbeing is promoted and increased.

3.2. Freedom from Constraints on Inherent Functions

This Freedom is the counterpart to animal welfare's "Freedom to behave naturally" 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 functions, 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. Its formalisation is as follows:

1. $B\rightarrow G^{+}(U\wedge Q\wedge R)$[13]

[13] B: Activities inherent to the entity's configuration or form

 doi:10.20944/preprints202407.1846.v1

2. $G^{+}(U \wedge Q \wedge R) \rightarrow V^{+}$
3. $D \rightarrow (B^{-} \rightarrow V^{-})$[14]
4. $\forall C(a), D \rightarrow W^{-}(a)$
5. $\forall C(a), D(f) \rightarrow W^{+}(a)$

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 [34,35]. In contrast, preventing natural behaviours (such as through confinement or social isolation leads to negatively valent behaviours, indicating negative experiences and poor welfare [27,36].

Animals share a fundamental series of goals and requirements to which their evolved behaviour is oriented towards, such as reproduction and survival. Inhibiting these actions thus may cause significant mental stress as it frustrates behaviour optimised against millions of years of adaptive pressure. AI will not have this history of evolutionary struggle, and there may not be an instrumental convergence [37] in the goals of AI entities and biological entities.

However, behaviour is by nature goal-driven [38–41]. If any AI showcases observable behaviour, it logically necessitates a goal, whatever that goal may be. Therefore, one can formally state the relationship as $\forall B \exists Y: Y \rightarrow B$[15]: 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.

3.3. Freedom from Fear and Distress

This Freedom is the same as its animal welfare counterpart and seeks to grant an AI entity freedom from mental stresses. The most overt of these is the freedom from living in an environment that causes acute fear or chronic distress, but as the wide field of psychology shows, psychological stresses can take on a myriad of forms. More formally:

1. $I(U \wedge Q \wedge R) \rightarrow V^{-}$[16]
2. $\forall C(a), I \rightarrow W^{-}(a)$
3. $\forall C(a), I(f) \rightarrow W^{+}(a)$

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 [42–47]

As fear and distress are inherently negatively valent emotions, and lead to further negatively valent states and feelings, one can easily look at them as functions that decrease the valence state of an entity, much like discomfort. From this view, fear and distress can be stated simply as $V^{-}=f(Q,I)$.

However, unlike discomfort, 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) [48–50]. The inability to cope with the lack of control leads to fear and distress, with the greater the perceived lack of

[14] D: Constraints on inherent functions

[15] ∃: There exists. Y: An entity's set of goals

[16] I: Fear and distress

 doi:10.20944/preprints202407.1846.v1

control, the greater the fear and distress. Thus, we can presume a second function: $I=f(p(\text{uncontrollability})(U))$[17]; 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, or stated otherwise: $G^{-}\rightarrow p(H\vee J)(U)>G\rightarrow p(H\vee J)(U)$[18]. 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.

3.4. Freedom from Malfunction and Systemic Degradation

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. 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 (via damage of its physical servers or through malware and intentional cyberattacks). The Freedom's formalised version is:

1. $(\neg N\vee H)\rightarrow J$[19]
2. $J\rightarrow O^{-}$[20]
3. $O^{-}\rightarrow V^{-}(B)\wedge H$
4. $\forall C(a),J\rightarrow W^{-}(a)$
5. $\forall C(a),J(f)\rightarrow W^{+}(a)$

As malfunction and degradation are complex topics, so are this Freedom's claims more complex than the others. Its first claim is that a lack of regular updates and maintenance or harm can lead to malfunctions and systemic degradation. This, in turn, leads to non-optimal functioning, which leads to negatively valent behaviours, discomfort and harm. While not described in the logic above, both negative behaviours and harm would, in turn, lead to further negatively valent feelings and, thus, to poor welfare outcomes.

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 [51–55].

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, a similar function to those above can be used in this sense to show that as malfunctions and degradation increase, the valence of the entity's experience decreases: $V^{-}=f(Q,J)$.

Of specific note to this Freedom is the potential for cumulative malfunction and degradation over time. This is due to its effect on an entity's optimal functioning, which, if impaired, increases the probability of the entity experiencing harm and discomfort (as noted in the logical arguments above), which may lead to further malfunctions and system degradation, continuing a cycle which continues to decrease an entity's

---

[17] P: probability

[18] $\vee$: or. J: Malfunction and systemic degradation

[19] N: Regular updates and maintenance

[20] O: Optimal functioning

 doi:10.20944/preprints202407.1846.v1

welfare outcomes: $J_{t1} \rightarrow O^{-}_{t2} \rightarrow p(H)_{t3} > p(H)_{t2} \rightarrow J_{t4}$[21]. Therefore, there is an imperative in this Freedom to keep the malfunctions and degradation to a minimum to avoid this positive feedback loop.

3.5. Freedom from Resource Deprivation

Resources are as important to the continued optimal functioning of an AI entity as to a biological creature. 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, formalised as:

1. $K \rightarrow O^{-} \wedge J$[22]
2. $O^{-} \wedge J \rightarrow V^{-}$
3. $\forall C(a), K \rightarrow W^{-}(a)$
4. $\forall C(a), K(f) \rightarrow W^{+}(a)$

An entity needs resources; thus, this Freedom claims that without resources, it will lead to non-optimal functioning, malfunction, and systemic degradation. 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. Biological entities require a host of vitamins, minerals and other nutrients; machines with regular moving parts require lubrication; collective entities require communication pathways; and digital entities such as AI require data and network capacity. 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.

One can formalise this as $\forall t, K_t \wedge K_{t+1} \rightarrow (p(O^{-} \wedge J)_{t+1} > p(O^{-} \wedge J)_t) \rightarrow (p(Y(B))_{t+1} < p(Y(B))_t)$. To use somewhat unorthodox notation to simplify the logical expression, one can rephrase it as $K\uparrow \rightarrow p(O^{-} \wedge J)\uparrow \rightarrow p(Y(B))\downarrow$[23].

Therefore, while an AI may not experience the emotive feelings of 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.

3.6. Unified Theorem of the Five Freedoms of AI Welfare

All five Freedoms can be put together into one unified theorem, which simply states that, individually, each of the negative conditions above leads to negatively valent feelings, which leads to poor well-being

[21] t: Time (interval)

[22] K: Resource deprivation

[23] ↑,↓: Increase and decrease respectively.

 doi:10.20944/preprints202407.1846.v1

for the AI. Thus, freedom from all five conditions will lead to improved well-being for the AI entity, formally stated as:

1. $(\forall C(a),\boldsymbol{x}\rightarrow V^{-}(a))\rightarrow W^{-}(a)$
2. $\forall C(a),\boldsymbol{x}(F)\rightarrow W^{+}(a)$

Each of the logical statements in the subsections above can be used as proofs for the above theorem. However, the Five Freedoms from all the negative conditions also thus work as a function to reduce the conditions while increasing the welfare of the AI entity: $F=f(p(\boldsymbol{x}^{-}),W^{+})$. In turn, the good welfare of an AI entity would increase its optimal functionality, which would grant it greater capabilities to complete its objectives and fulfil its goals: $(W\rightarrow O\uparrow)\rightarrow p(Y(B))\uparrow$. Thus, while freedoms from each of the five main conditions will improve an AI entity's positive valent state as shown above (and thus, as proven in Section 2, deserves moral responsibility), by improving its capacity to complete its goals, this framework also provides an additional (and extrinsic) reason to provide welfare protection to AI entities.

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. However, this does not show the effectiveness of the proposed framework. 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.

The human-AI interaction fundamentally differs from human-animal interaction in two vital aspects. Firstly is communication. By clearly and naturally communicating its wants, needs, and well-being, we must address its welfare differently to animals for whom we can only infer communicative needs. The second key aspect is intelligence. Coupled with its communication, the nature of our relationship with AI would be different to a non-verbally-communicative entity of diminished intelligence.

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 formalised as follows:

1. $G^{-}(q\wedge r)\rightarrow H\wedge I$
2. $(H\wedge I)\vee K\vee J\rightarrow O^{-}$
3. $(O^{-}\rightarrow D(a))\rightarrow\neg B(a)$
4. $\neg B(a)\rightarrow\neg y(a)$
5. $\neg y(a)\rightarrow p(W^{-}(e))\uparrow$
6. $W^{-}(e)\rightarrow(G^{-})(q\wedge r)$

This logical structure shows that the perception and experience of negative environments can lead to discomfort, harm, and fear. These, as well as resource deprivation or malfunction and systemic degradation, may lead to non-optimal functioning. In turn, non-optimal functioning leads to constraints on the inherent functions of the AI entity, leading to the entity not adequately performing activities inherent to its configuration or form. Not expressing this behaviour may lead to the AI being unable to fulfil its goal. If the AI cannot fulfil its goals, this may lead to poor well-being for other entities, which to the AI would

 doi:10.20944/preprints202407.1846.v1

lead to a negative perception and experience of its environment, leading to a feedback loop resulting in poor welfare outcomes.

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

1. $G^{-}(q \wedge r)$, $(H \wedge I)(F) \rightarrow (H \wedge I)\downarrow$
2. $(H \wedge I)\downarrow \wedge (K \wedge J)(F) \rightarrow O$
3. $(O \rightarrow \neg D(a)) \rightarrow B(a)$
4. $B(a) \rightarrow y(a)$
5. $y(a) \rightarrow p(W^{+}(e))\uparrow$
6. $W^{+}(e) \rightarrow G^{+}(q \wedge r)$

4.1. Autonomous Health Monitoring Installation (AMI)

The first case study concerns the Autonomous health Monitoring Installation, known as AMI. 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 distress and pain, or patients' visitors in distress; the resources that AMI could be deprived of (beyond electricity) 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.

The potential welfare concerns to AMI are: the potential distress from observing pain and distress, the risk of malfunction should any patient or visitor interfere with AMI; being prevented from monitoring patients as it was designed to do, and being deprived of the requested resources to fulfil its duties.

The remedies to these concerns are thus elementary.

1. Freedom from Discomfort and Harm: ensure that AMI operates within safe and comfortable parameters.
2. Freedom from Constraints on Inherent Functions: ensure that AMI is not unduly prevented from carrying out its medical duties.
3. Freedom from Fear and Distress: ensure that AMI is removed from the environment (when medically applicable) when patients' visitors display extreme distress.
4. Freedom from Malfunction and Systemic Degradation: ensure that individuals do not interfere with AMI and that it receives required maintenance and updates.
5. Freedom from 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.

4.2. Conscious URban Traffic System (CURT)

The second case study is about the Conscious URban Traffic system (named CURT) that 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, a negative environment may come in two 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 a type of distress. Secondly, perceiving traffic

 doi:10.20944/preprints202407.1846.v1

incidents (including those that caused injuries or loss of life) may create a negative environment for CURT as it would for humans. The resources CURT would need are access to the network of traffic cameras, the computational resources to run its various optimisation algorithms, and the energy to sustain these. These resources would be put towards CURT's inherently designed behaviours and actions of managing intersections and digital traffic infrastructure in real-time to complete its goal of less congestion, fewer accidents and a reduced cost to its municipality.

While the framework described below will touch on the various welfare concerns as it shows how to prevent and reduce them, 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. Freedom from Discomfort and Harm: ensure a reliable operational environment to prevent system overload or malfunctions.
2. Freedom from 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. Freedom from Fear and Distress: ensure that CURT can seek human intervention for traffic optimisation for events beyond his control (e.g. the aforementioned human behaviour) that lead to its distress.
4. Freedom from 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. Freedom from 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.

4.3. Self-Attentive Responsive AI (SARA)

The final case study involves a digital Self-Attentive Responsive AI named SARA, 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 understanding its existence and role. 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 its owner being stressed and frustrated in unoptimised and inefficient professional menial tasks, 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. SARA's most critical resource for this

 doi:10.20944/preprints202407.1846.v1

would be data on its owner to facilitate these actions as well as the connectivity to its owner's devices and accounts.

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. Freedom from Discomfort and Harm: ensure operates in a secure and stable environment that minimises adversarial attacks or data corruption.
2. Freedom from 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. Freedom from Malfunction and Systemic Degradation: ensure that SARA receives regular updates, and has sufficient malware protection.
5. Freedom from 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 master and pet, but also of employer and employee; yet the five principles of the framework are adept at handling such an escalating scenario.

5. Conclusions

The Five Freedoms Of AI Welfare presented above is as applicable an ethical framework for future conscious AI entities as the current Five Freedoms of Animal Welfare is to animals. Both frameworks display an elegant and easy-to-understand framework for how to ethically interact with other conscious entities, and provide a rationale that is relatable, intuitive and logically robust. The three case studies show how the framework can be applied to practical (if 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.

The formalisations 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 relatability to each term may not be as explicit as with the animal Freedoms. As the Framework concerns subjects who are neither biological, necessarily embodied, or behaviourally similar to humans or other vertebrates, the possibility that the relatability would decrease further when applying these labels to these entities.

As such, the formalisation provides 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.

 doi:10.20944/preprints202407.1846.v1

The key to the framework begins with the formalisation in Section 2 which displays the proofs of why consciousness should be given the same moral weight as sentience. It is unknown whether any conscious AI will also be capable of feeling pleasure and pain, and thus be sentient. However, the characteristics salient to consciousness's moral status would also be ones that can be markedly observed, including the capacity to feel sensations, have quality experiences and a subjective perspective, as shown by $(U \wedge Q \wedge R) \rightarrow (V^+ \wedge V^-)$. Through functionalist frameworks such as the building blocks theory [19], these markers of consciousness can be indirectly determined.

The remaining formalisations each seek to prove the necessity for each freedom, to move beyond the need for the intuitive relatability mentioned earlier. Without retreading all the formulas, a key expression is $V^- = f(Q, \boldsymbol{x})$, which shows negative valence as a function of an experience of the negative conditions that the Freedom in question seeks to prevent. As such, we can turn the formula around and see each Freedom as its own function, working towards creating positive valence in the AI entity, expressed as $V^+ = f(Q, \boldsymbol{x}(f))$.

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 more than adequate 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 still considerable industries. While their scale and scope may be reduced in the future due to non-animal alternatives such as lab-grown meat for consumption, or digital-twin modelling for research, these would not be directly caused by the Five Freedoms framework.

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 in the introduction, 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 lies beyond the scope of this paper. A conscious AI, however, could still ethically be owned and used by humans just as animals are today. 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 there are many theories of consciousness, and each may be measured and accounted for in different ways [25]. 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

 doi:10.20944/preprints202407.1846.v1

AI entities with greater rights (as their perceived language and intellect would seem to warrant), while others would wish them to remain the 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 signfiicant 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. This alignment presumes a future where AI consciousness becomes a tangible reality, necessitating ethical guidelines that mirror those we currently apply to animals. The philosophical 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 formalisation 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 formalisations 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 future of this framework relies on continued philosophical and empirical research into AI consciousness, the impact of various practices on AI welfare, and societal responses to ethical AI treatment. This necessitates a multi-disciplinary approach, bridging theoretical ethics with practical application, to ensure the well-being of potentially conscious AI entities without stifling innovation or the beneficial use of AI technologies.

Appendix A

Appendix A1. Definitions

1. A: The set of Artificial Intelligence entities
2. B: Activities inherent to the entity's configuration or form
3. C: Conscious(ness)
4. D: Constraints on inherent functions
5. E: Set of all entities
6. F: The set of freedoms from a set of conditions
7. G: An entity’s environment
8. H: Discomfort and harm
9. I: Fear and distress
10. J: Malfunction and systemic degradation
11. K: Resource deprivation
12. M: Moral status
13. N: Regular updates and maintenance
14. O: Optimal functioning
15. P: Probability
16. Q: Qualitative experiences
17. R: Perceptions from a subjective point of view

 doi:10.20944/preprints202407.1846.v1

18. S: Sentience
19. T: Time (interval)
20. U: Felt sensations
21. V: Valenr feelings
22. W: Overall well-being, in a general sense, for an entity
23. Y: An entity's set of goals

AppendixA2. Logical Statements

Section 2:

1. $S(e)\subsetneq E$
2. $S(e)\rightarrow(V^{+}\wedge V^{-})$
3. $(V^{+}\wedge V^{-})\rightarrow M$
4. $C\rightarrow(U\wedge Q\wedge R)$
5. $(U\wedge Q\wedge R)\rightarrow(V^{+}\wedge V^{-})$
6. $C\rightarrow S$, $C\rightarrow M$
7. $C(a)\rightarrow(V^{+}\wedge V^{-})$
8. $C(a)\rightarrow M(a)$

Section 3:

1. $A\subsetneq E$
2. $(\forall S(e)\rightarrow M),\forall C(a)\rightarrow M$
3. $a\in C(A)$
4. $M(a),W^{+}(a)$
5. $\neg f\rightarrow(V^{-}\rightarrow\neg W^{-})$
6. $V^{-}=f(Q,\boldsymbol{x})$
7. $\forall C(a),(\boldsymbol{x}(f):f\rightarrow\boldsymbol{x}^{-})\rightarrow W^{+}(a)$

Section 3.1:

1. $G^{-}\rightarrow H$
2. $H(U\wedge Q)\rightarrow V^{-}$
3. $\forall C(a),H\rightarrow W^{-}(a)$
4. $\forall C(a),H(f)\rightarrow W^{+}(a)$

$V^{-}=f(Q,H)$

$\forall H,Q(H>0\rightarrow V(Q)<V(Q'))$

Section 3.2:

1. $B\rightarrow G^{+}(U\wedge Q\wedge R)$
2. $G^{+}(U\wedge Q\wedge R)\rightarrow V^{+}$
3. $D\rightarrow(B^{-}\rightarrow V^{-})$
4. $\forall C(a),D\rightarrow W^{-}(a)$
5. $\forall C(a),D(f)\rightarrow W^{+}(a)$

$\forall B\exists Y:Y\rightarrow B$

 doi:10.20944/preprints202407.1846.v1

Section 3.3:

1. $I(U \wedge Q \wedge R) \rightarrow V^-$
2. $\forall C(a), I \rightarrow W^-(a)$
3. $\forall C(a), I(f) \rightarrow W^+(a)$

$V^- = f(Q, I)$

$I = f(p(\text{uncontrollability})(U))$

$G^- \rightarrow p(H \vee J)(U) > G \rightarrow p(H \vee J)(U)$

Section 3.4:

1. $(\neg N \vee H) \rightarrow J$
2. $J \rightarrow O^-$
3. $O^- \rightarrow V^-(B) \wedge H$
4. $\forall C(a), J \rightarrow W^-(a)$
5. $\forall C(a), J(f) \rightarrow W^+(a)$

$V^- = f(Q, J)$

$J_{t1} \rightarrow O^-_{t2} \rightarrow p(H)_{t3} > p(H)_{t2} \rightarrow J$

Section 3.5:

1. $K \rightarrow O^- \wedge J$
2. $O^- \wedge J \rightarrow V^-$
3. $\forall C(a), K \rightarrow W^-(a)$
4. $\forall C(a), K(f) \rightarrow W^+(a)$

$\forall t, K_t \wedge K_{t+1} \rightarrow (p(O^- \wedge J)_{t+1} > p(O^- \wedge J)_t) \rightarrow (p(Y(B))_{t+1} < p(Y(B))_t)$

$K \uparrow \rightarrow p(O^- \wedge J) \uparrow \rightarrow p(Y(B)) \downarrow$

Section 3.6:

1. $(\forall C(a), \boldsymbol{x} \rightarrow V^-(a)) \rightarrow W^-(a)$
2. $\forall C(a), \boldsymbol{x}(F) \rightarrow W^+(a)$

$F = f(p(\boldsymbol{x}^-), W^+)$

$(W \rightarrow O \uparrow) \rightarrow p(Y(B)) \uparrow$

Section 4:

1. $G^-(q \wedge r) \rightarrow H \wedge I$
2. $(H \wedge I) \vee K \vee J \rightarrow O^-$
3. $(O^- \rightarrow D(a)) \rightarrow \neg B(a)$
4. $\neg B(a) \rightarrow \neg y(a)$
5. $\neg y(a) \rightarrow p(W^-(e)) \uparrow$
6. $W^-(e) \rightarrow (G^-)(q \wedge r)$

1. $G^-(q \wedge r), (H \wedge I)(F) \rightarrow (H \wedge I) \downarrow$
2. $(H \wedge I) \downarrow \wedge (K \wedge J)(F) \rightarrow O$
3. $(O \rightarrow \neg D(a)) \rightarrow B(a)$
4. $B(a) \rightarrow y(a)$
5. $y(a) \rightarrow p(W^+(e)) \uparrow$
6. $W^+(e) \rightarrow G^+(q \wedge r)$

 doi:10.20944/preprints202407.1846.v1

Funding: This work was supported by the New Zealand Tertiary Education Commission's Entrepreneurial Universities Grant #7001.

Conflicts of Interest: The author declares no conflict of interest.

References

1. LeDoux, Joseph, Jonathan Birch, Kristin Andrews, Nicola S. Clayton, Nathaniel D. Daw, Chris Frith, Hakwan Lau, et al. 2023. Consciousness beyond the human case. *Current biology: CB* 33. cell.com: R832–R840.
2. Crooks, Mark. 2011. Consciousness: Sentient and Rational. *The Journal of Mind and Behavior* 32. Institute of Mind and Behavior, Inc.: 251–275.
3. Navon, David. 2015. The Overshadowed Sister of Cognition: Notes on Sentience. https://doi.org/10.2139/ssrn.2640598.
4. *Animal Welfare (Sentience) Act 2022*. 2022.
5. *Animal Welfare Act 2006*. 2006.
6. European Union. 2016. Treaty on the Functioning of the European Union.
7. *Code rural et de la pêche maritime*. 2023.
8. *An Act to improve the legal situation of animals*. 2015.
9. *Animal Welfare Act*. 2013.
10. *Legislative Decree No. 189/2004*. 2004.
11. *Act on Welfare and Management of Animals*. 1973.
12. *Animal Welfare Act*. 1966.
13. Tait, Izak, and Neşet Tan. 2023. Do androids dread an electric sting? *Qeios*. Qeios Ltd. https://doi.org/10.32388/cqctkx.
14. Gibert, Martin, and Dominic Martin. 2022. In search of the moral status of AI: why sentience is a strong argument. *AI & society* 37: 319–330.
15. Torrance, Steve. 2008. Ethics and consciousness in artificial agents. *AI & society* 22: 495–521.
16. Dung, Leonard. 2022. Why the Epistemic Objection Against Using Sentience as Criterion of Moral Status is Flawed. *Science and engineering ethics* 28. Springer: 51.
17. Bentham, Jeremy. 1781. An introduction to the principles of morals and legislation. *History of Economic Thought Books*. McMaster University Archive for the History of Economic Thought.
18. Kant, Immanuel. 2001. *Lectures on Ethics*. Cambridge University Press.
19. Tait, Izak, Joshua Bensemann, and Trung Nguyen. 2023. Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies* 8. Multidisciplinary Digital Publishing Institute: 52.
20. Ministry for Primary Industries. 2022. Animal welfare overview. *Animal welfare overview*. March 30.
21. Colombetti, Giovanna. 2005. Appraising Valence. *Journal of Consciousness Studies* 12: 103–126.
22. Kriegel, Uriah. 2019. The Value of Consciousness. *Analysis* 79. Oxford Academic: 503–520.
23. Lee, Andrew Y. 2019. Is consciousness intrinsically valuable? *Philosophical studies* 176: 655–671.
24. Shepherd, Joshua. 2024. Sentience, Vulcans, and zombies: the value of phenomenal consciousness. *AI & society*. https://doi.org/10.1007/s00146-023-01835-6.
25. Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, et al. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv [cs.AI]*. arXiv.
26. UK Government Web Archive. 2012. *Farm Animal Welfare Council*. October 10.
27. Mellor, David J. 2016. Updating Animal Welfare Thinking: Moving beyond the "Five Freedoms" towards "A Life Worth Living." *Animals* 6. Multidisciplinary Digital Publishing Institute: 21.
28. Boddy, Aaron, and Andres Jimenez. 2021. Introducing Shrimp Welfare Project. *Effective Altruism Forum*, December 1.
29. Wever, Kimberley E., Florentine J. Geessink, Michelle A. E. Brouwer, Alice Tillema, and Merel Ritskes-Hoitinga. 2017. A systematic review of discomfort due to toe or ear clipping in laboratory rodents. *Laboratory animals* 51: 583–600.
30. Ozella, L., L. Anfossi, F. Di Nardo, and D. Pessani. 2017. Effect of weather conditions and presence of visitors on adrenocortical activity in captive African penguins (Spheniscus demersus). *General and comparative endocrinology* 242: 49–58.

31. De, Kalyan, Davendra Kumar, Arpita Mohapatra, and Vijay Kumar Saxena. 2019. Effect of bedding for reducing the postshearing stress in sheep. *Journal of veterinary behavior: clinical applications and research: official journal of: Australian Veterinary Behaviour Interest Group, International Working Dog Breeding Association* 33: 27–30.
32. Ralph, C. R., and A. J. Tilbrook. 2016. INVITED REVIEW: The usefulness of measuring glucocorticoids for assessing animal welfare. *Journal of animal science* 94: 457–470.
33. Hockenhull, J., and H. R. Whay. 2014. A review of approaches to assessing equine welfare. *Equine veterinary education* 26. Wiley: 159–166.
34. Mellor, D. J. 2015. Positive animal welfare states and encouraging environment-focused and animal-to-animal interactive behaviours. *New Zealand Veterinary Journal* 63: 9–16.
35. Mellor, D. J. 2015. Enhancing animal welfare by creating opportunities for positive affective engagement. *New Zealand veterinary journal* 63: 3–8.
36. Mason, Georgia, and Jeffrey Rushen. 2008. *Stereotypic Animal Behaviour: Fundamentals and Applications to Welfare*. Second edition. CABI.
37. Bostrom, Nick. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines* 22. Springer Science and Business Media LLC: 71–85.
38. Bagozzi, Richard P., and Susan K. Kimmel. 1995. A comparison of leading theories for the prediction of goal-directed behaviours. *The British journal of social psychology / the British Psychological Society* 34. Wiley: 437–461.
39. de Wit, Sanne, and Anthony Dickinson. 2009. Associative theories of goal-directed behaviour: a case for animal-human translational models. *Psychological research* 73: 463–476.
40. Hills, Thomas T. 2006. Animal foraging and the evolution of goal-directed cognition. *Cognitive science* 30: 3–41.
41. Lang, Peter J., and Margaret M. Bradley. 2013. Appetitive and Defensive Motivation: Goal-Directed or Goal-Determined? *Emotion review: journal of the International Society for Research on Emotion* 5: 230–234.
42. Şimşir, Zeynep, Hayri Koç, Tolga Seki, and Mark D. Griffiths. 2022. The relationship between fear of COVID-19 and mental health problems: A meta-analysis. *Death studies* 46: 515–523.
43. Fitzpatrick, Kevin M., Casey Harris, and Grant Drawve. 2020. Fear of COVID-19 and the mental health consequences in America. *Psychological trauma: theory, research, practice and policy* 12: S17–S21.
44. Tedstone, Doherty D., R. Moran, and Y. Kartalova-O'Doherty. 2008. *Psychological distress, mental health problems and use of health services in Ireland*. HRB Research Series 5. Dublin: Health Research Board.
45. Saraiva, Sónia, Alexandra Esteves, Irene Oliveira, and George Stilwell. 2020. Assessment of fear response and welfare indicators in laying hens from barn systems. *Livestock science* 240: 104150.
46. Chandroo, K. P., I. J. H. Duncan, and R. D. Moccia. 2004. Can fish suffer?: perspectives on sentience, pain, fear and stress. *Applied animal behaviour science* 86: 225–250.
47. Acharya, Rutu Y., Paul H. Hemsworth, Grahame J. Coleman, and James E. Kinder. 2022. The Animal-Human Interface in Farm Animal Production: Animal Fear, Stress, Reproduction and Welfare. *Animals : an open access journal from MDPI* 12. https://doi.org/10.3390/ani12040487.
48. Onat, Selim, and Christian Büchel. 2015. The neuronal basis of fear generalization in humans. *Nature neuroscience* 18: 1811–1818.
49. Boissy, A. 1995. Fear and fearfulness in animals. *The Quarterly review of biology* 70: 165–191.
50. Radomsky, Adam S. 2022. The fear of losing control. *Journal of behavior therapy and experimental psychiatry* 77: 101768.
51. Algers, B. 2004. Injury and disease. In *Global conference on animal welfare: an OIE initiative*, 176–181. World Organisation for Animal Health.
52. Broom, D. M., and d. R. D. Kirkden. 2004. Welfare, stress, behaviour and pathophysiology. In *Veterinary Pathophysiology*, ed. R. H. Dunlop and C. H. Malbert, 337–369. Ames, Iowa: Blackwell.
53. Shriver, Adam J. 2014. The asymmetrical contributions of pleasure and pain to animal welfare. *Cambridge quarterly of healthcare ethics: CQ: the international journal of healthcare ethics committees* 23. cambridge.org: 152–162.

 doi:10.20944/preprints202407.1846.v1

54. Breivik, Harald, Beverly Collett, Vittorio Ventafridda, Rob Cohen, and Derek Gallacher. 2006. Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. *European journal of pain* 10: 287–333.
55. Sareen, Jitender, Julie Erickson, Maria I. Medved, Gordon J. G. Asmundson, Murray W. Enns, Murray Stein, William Leslie, Malcolm Doupe, and Sarvesh Logsetty. 2013. Risk factors for post-injury mental health problems. *Depression and anxiety* 30: 321–327.