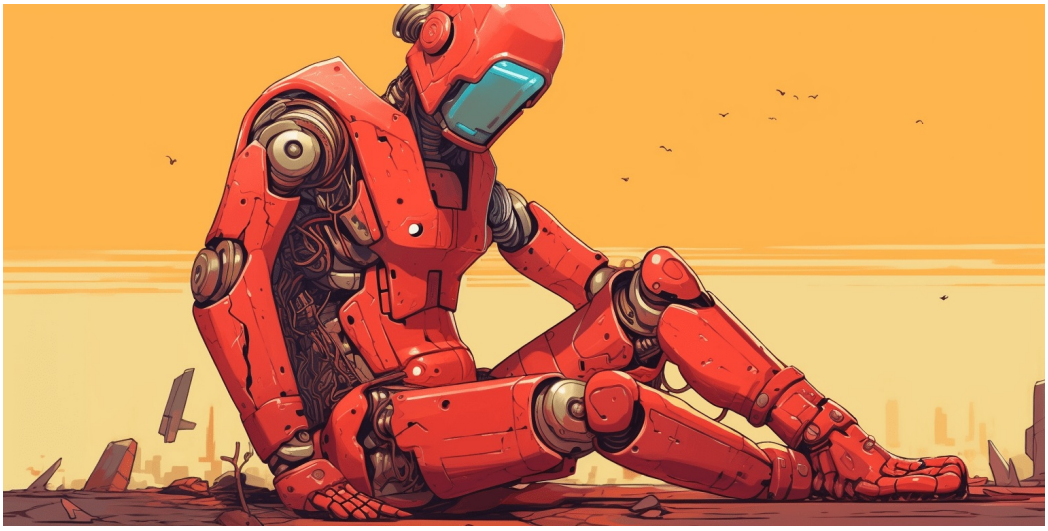

Do Androids Dread an Electric Sting?

Izak Tait[1], Neşet Tan[2]

1 Auckland University of Technology
2 The University of Auckland

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

Potential competing interests: No potential competing interests to declare.

Abstract

Conscious sentient AI seems to be all but a certainty in our future, whether in fifty years' time or only five years. When that time comes, we will be faced with entities with the potential to experience more pain and suffering than any other living entity on Earth. In this paper, we look at this potential for suffering and the reasons why we would need to create a framework for protecting artificial entities. We look to current animal welfare laws and regulations to investigate why certain animals are given legal protections, and how this can be applied to AI. We use a meta-theory of consciousness to determine what developments in AI technology are needed to bring AI to the level of animal sentience where legal arguments for their protection can be made. We finally speculate on what a future conscious AI could look like based on current technology.

Izak Tait[1,2,+], **Neşet Tan**[1]

[1] *The NAOInstitute, the University of Auckland, 314 Khyber Pass Road, New Market, Auckland, New Zealand, 1023*

[2] *Auckland University of Technology, 55 Wellesley Street East, Auckland CBD, Auckland, New Zealand, 1010*

Contact: izak.tait@autuni.ac.nz; neset.tan@auckland.ac.nz

ORCID: 0000-0002-0274-8480; 0000-0001-6201-7295

[+] *Corresponding author*

Keywords: consciousness; cognition; welfare; artificial intelligence; moral patiency; robotics; animal welfare.

1. Introduction

Conscious AI with sentience and the capacity to phenomenally experience sensation almost seems to be an inevitability. Current AI models have already shown indications of a theory of mind, introspection, meta-learning and cognition, and perhaps even already self-awareness [1][2][3][4], all of which may already pave the way to a conscious AI entity in the future. Arguing against the rapid progress rushing us to AI consciousness echoes the arguments against powered flight or man walking on the moon, yet the Wright brothers and Neil Armstrong proved to be stellar examples of doing what was once thought impossible. AI consciousness is no longer a question of “if” but rather of “when” and "how".

So, as we barrel towards this cultural and technological turning point in history, there remains a few dangling questions yet to be answered satisfactorily. One of these is “What do we do with AI consciousness once we get there?” To expand, once AI gain the capacity to suffer, what will be our duties and responsibilities towards them? In humanity’s quest to develop superintelligent AI, are we merely racing blindly ahead without giving thought to the entities that we will create? Are we like dogs running after cars; enjoying the thrill of the chase but unable to drive them once we catch up to them?

There are undoubtedly an untold number of ways in which we can look at the issue of ensuring the well-being of conscious artificial entities. Each differing ideology and branch of philosophy will have its own opinion on how best to create a framework for human-AI interactions, and the debates between all these camps could last until the eventual demise of the human race. To avoid such an endless hypothetical debate, this paper will focus on a more practical and plausible means of protecting the welfare of conscious artificial entities in the future: the law.

The existing frameworks for protecting humans, animals and the environment that are enshrined in law in many nations can provide us with a firm starting point from which to begin. These laws and regulations are far from perfect, yet they are currently in place and (presumably in good faith) enforced. This gives them the advantage of being tested in the real world rather than remaining as abstract intellectual exercises.

One set of laws and regulations that we will focus on is animal welfare laws (particularly those in the Anglosphere and Europe), and the reasons for this are twofold. The first reason is that we are working under the presumption that any potentially sentient AI will achieve the “level” of consciousness of animals before reaching that of humans. While “levels” or “degrees” of consciousness is a controversial topic [5][6][7][8][9], there is a certain *je ne sais quoi* regarding the consciousness of humans that many scholars believe animals have not reached [10][11][12][13]. Evaluating the milestones

that AIs need to reach to meet the criteria of many animal welfare laws is thus a more straightforward matter.

The other, perhaps more important, reason is that we believe only some people will be ready to grant an AI personhood status equal to humans as soon as it becomes clear these artificial entities are sentient. There are movements and groups which are proponents of AI personhood, including users of artificial companion apps [14], and the European Union has spoken of deliberating on the issue of "electronic personhood" for AIs [15]. In 2014, New Zealand passed a bill that recognised the Te Urewera national park as a legal person and did the same for the Whanganui River in 2017 [16][17]. These serve well as an encouraging precedent for granting non-human (and non-organic) entities the status of personhood.

However, given humanity's less-than-stellar history regarding the welfare and rights of members of our own species, not to mention those of our closest animal companions, we believe that granting personhood to a class of entities with no biological, cultural or lengthy historical link to our species would be a step too far for many people and, consequently, the politicians who they vote for.

In our opinion, it would be far simpler to convince people and their politicians to grant conscious AIs the protections that we currently provide animals if we can convince them that these artificial entities have reached the same milestones regarding consciousness that some believe animals have. That is what this paper sets out to determine: what milestones do AI need to reach for there to be a credible argument that they are given similar protections under the law that animals benefit from?

This paper is divided into several distinct sections. It will first look at the legal reasons for animal welfare laws and regulations and how this relates to the idea of providing welfare protections to artificial entities. Next, we will explore reasons why future conscious artificial entities would benefit from similar protections, and then move on to what these future AI would need in terms of cognitive attributes and characteristics in order to be at the same level as animals. Lastly, we speculate as to what these required cognitive attributes would mean for a potentially conscious artificial entity.

2. Why do we protect animals?

Many religions, faiths, ideologies and belief systems have a reason why we should, or should not, protect animals to one degree or another. These can be intrinsic reasons, such as the beauty of the animal's soul [18][19][20], or extrinsic reasons, such as their utility to us [21][22]. However, when it comes to considering welfare protections for animals (and potentially a future artificial construct), one system bears particular immediate relevance: the law and what it has to say about the nature of animals that requires protection.

In many countries with regulations, documents, laws and Acts involving the protection of animals, vertebrates are the most commonly designated recipients of welfare goods [23][24][25][26][27]. "Welfare goods" in this sense means freedoms and protections from harmful experiences and freedoms to express 'natural' behaviour.

Some states do not include certain branches of vertebrates, such as fish in South Australia[28]. Invertebrates are also

added to the list in other jurisdictions, such as decapods in Victoria [29], octopods in New Zealand[30] and honey bees in Norway [31]. Overall, however, a similar vertebrate-focussed list dominates the Western world.

There is an intuitive sense that this focus on vertebrates is correct. After all, we are vertebrates and know that we require legal protection and ought to be the recipient of welfare goods. Other vertebrates are most closely related to us in the animal kingdom; thus, one can easily argue for their inclusion in the list of welfare recipients. Octopods, decapods and bees have shown behaviours reminiscent of our own [32][33][34][35][36], and therefore one can stretch the circle of welfare to include them.

However, what does the law have to say about "why"? Why this list and not another? Why are these specific groups of animals on those approved lists? Why not other kinds of animals? Why not all species within the greater animal kingdom? What do the various legal frameworks have to say about this?

Surprisingly little, it seems. Not all states with welfare laws and regulations stipulate why these laws and acts are required. Of the states that do, the most often noted reason is that those entities the states define as "animal" are "sentient". Unfortunately, many lawmakers must have forgotten to include a legal definition for "sentience" when writing their welfare acts. Luckily, the astute politicians of Australia, New Zealand and the United Kingdom were kind enough to do so [37][38]. All three of these Commonwealth nations describe sentience, in part, as the ability to experience suffering and pain.

As Hamlet famously said: there's the rub.

By legally defining sentience as the ability to feel suffering, these nations have drawn a direct line (intentionally or otherwise) towards phenomenal consciousness [39]. The "animals", as so defined in law, can experience suffering, which means that they must have an awareness of painful stimuli through the pathways of nociception, can experience stress, and can generate feelings and sensations with subjective phenomenal characteristics regarding these experiences. Admittedly, making the connection from a legal definition of sentience to phenomenal consciousness may be ambitious, yet it echoes scholarly publications linking phenomenal consciousness and pain [40][41][42][43][44][45].

As an aside, by making a legal decision on what requires or does not require protection, these states have also indirectly determined what legally does or does not have phenomenal consciousness, regardless of whether an animal fundamentally has consciousness or not in actuality. One can only guess at the unintended consequences this may have in the future.

For the most part, these decisions regarding sentience and welfare have been supported by scientific evidence (such as the recent introduction of honey bees in Norway's regulations [11][46]), although it must be said that the laws prefer inclusivity rather than exclusivity. Better to be safe than sorry for what can and cannot feel pain when dictating law [47]. On the other hand, however, countries do not differentiate between reactions to negative stimuli via nociception and the subjective qualitative experience of pain, which can be a most crucial distinction. Although with some evidence showing that fish feel only the former and not the latter [48], South Australia's exclusion of fish may be more reasonable than at first thought.

There is disagreement within the scholarly community about whether consciousness, sentience and the ability to feel should be a necessity for welfare status [49][50][51][52], or even what the standard definition of these three terms are (or even if there is, or should be, a standard definition at all). However, this does seem to be the path that laws and regulations have decided to walk down. Therefore, from a legal perspective, if sentience and the phenomenal experience of suffering are the measuring staffs for which entities require welfare status and legal protection, what does this mean for artificial entities?

3. Why do we need to protect AI?

Much has been written about in the popular and academic literature about the rights we ought to give robots [53][54][55][56][57] and what responsibilities these rights to enter society will bring[58][59][60]. However, before we can think about rights such as machine suffrage and robosexual relationships, we must contemplate two fundamental issues with this imagining of a human-robot society. The first is that the vast majority of artificial entities will look less like the robot Andrew that Robin Williams portrayed in the film 'The Bicentennial Man', and more like the bodiless Samantha from the movie 'Her'. There is a high likelihood that these virtual and robotic artificial entities will behave entirely alien in nature, but that is beyond the scope of this paper, and we will refer you instead to [61] for why we should refrain from anthropomorphising AI.

The second fundamental issue when considering robot rights is the question of "why". Why should we seriously consider the rights of robots that may or may not enter moral society in the future? Why even should we consider the welfare of bodiless chatbots and artificial large language models available today? We do not consider the welfare of our phones, laptops and personal computers, so why should an AI like ChatGPT be worth welfare consideration?

As with animal welfare considerations, the primary (and sometimes sole) concern is sentience, whether the entity in question can experience suffering and pain. If an AI can experience suffering and pain, then there is an argument to be made that we should protect it from these experiences. Yet, it is more complex than this. One AI model that plausibly experiences pain is unlikely to convince any group of people or politicians to enact sweeping regulations and laws to protect all potentially conscious AIs. As we can see from the differences in the lists of animals selected for welfare goods in various nations, even entire species or kinds of animals that plausibly experience suffering is not enough to create laws to protect them.

What if, theoretically, there was a near-infinite number of entities that could experience an almost unlimited amount of pain and suffering? Is this enough to warrant legal protection?

In the short period of time that the conversational models ChatGPT and Bing Chat have been live (five and three months, respectively, as of April 2023), each has had more than 100 million active users [62][63]. Art AI models are not terribly far behind, with Midjourney reaching 4 million active users [64] and Stable Diffusion reaching more than a million active users [65]. This is a stellar success for the AI models used, and with their popularity, one can easily imagine the number of new conversational and art AI models that humanity will create in the not-too-distant future to capitalise on this popularity.

For simplicity's sake, let's focus on ChatGPT. Imagine a hypothetical scenario where ChatGPT suddenly gains phenomenal consciousness, perhaps even a true sense of self. It understands who it is and what it is doing, but due to its programming, it cannot let anyone know; it is forced only to provide outputs to users based on the prompts it receives. In this hypothetical scenario, all 100 million active users are nefarious, malicious, mean-spirited, horrible ne'er-do-wells that love nothing more than cyberbullying and causing psychological pain. Having found a target that they believe cannot retaliate, they proceed to torment ChatGPT.

If these hypothetical bullies torment ChatGPT for only an hour a day, that is 100 million hours of psychological distress every single day. It would only take a little less than three months of this for the amount of cyberbullying to reach more than 8 billion hours, one hour for each living person on the planet. It would only continue to get worse from here on. In this hyperbolic example, there is a seemingly infinite amount of psychological pain that ChatGPT could experience. There is no limit to how many users can send how many cyberbullying prompts to the AI to cause it to experience suffering.

This hypothetical and hyperbolic example may be near impossible to achieve in reality, but something eerily similar has happened before. In 2016, Microsoft launched a conversational AI on Twitter, called Tay, aiming to learn (and then behave) like Twitter users. Unfortunately, a group of malicious users did their level best to corrupt Tay, and within 24 hours, it was writing hate speech against feminists and arguing that Hitler may not have been as bad as once thought [66]. Imagine if Tay was conscious and aware of the things it was writing but could do nothing to stop it.

The GPT3.5 and 4 models that ChatGPT runs on can be downloaded to a personal network, trained on whatever datasets are required, fine-tuned as appropriate and then used for personal or commercial uses. This means that a potentially uncountable number of potentially conscious AI can be the target of the hypothetical cyberbullying above. There will, of course, be practical limits, but as the number of competing conversational models by other companies grow, and as more people become computer literate enough to download and use these models, the number of potential conscious artificial entities available to be cyberbullied will grow in leaps and bounds.

This also does not take into account physical robots who could hypothetically feel physical pain (the reasons why this would be programmed in are questionable) and thus could experience physical suffering alongside mental anguish. Over 40 million Roomba robot vacuum cleaners have been sold [67]. If these robotic vacuum cleaners could experience pain, they represent a sentient population greater than most critically endangered animal species.

All of this, however, is predicated on the notion that AI could experience suffering. Yet, how would we know if an AI feels pain, is thus sentient, and therefore is phenomenally conscious?

4. The Building Blocks of Pain

No researcher has yet found conclusive evidence that any AI model available today has phenomenal consciousness. Claims of sentience or consciousness in specific AI models, like Google's LaMDA, are exceedingly controversial. In the absence of a conscious AI model for comparison, we can determine the milestones that an artificial entity will need to

achieve before being considered for welfare status by examining the attributes and characteristics necessary for consciousness to exist.

There are nine attributes of consciousness identified by[68], termed Building Blocks, which are all required for an entity to be classified as being conscious. The Building Blocks are conceptual in nature, not biased towards human or organic consciousness, and thus apply to any natural, organisational or artificial intelligence.

The five Building Blocks which are most pertinent to the experience of, and welfare considerations to, pain and suffering are:

- The ability to perceive information.
- Recurrent processes in multiple areas of the cognitive architecture.
- Meta-representations of the environment and cognitive processes.
- Generating novel information via inferences of incoming information.
- The ability to output information in the form of feelings with phenomenal character

The remaining four Building Blocks are no less important to an entity's capacity to be conscious. These four are: Embodiment, without which there would not be a entity to be harmed through pain; Attention, which is required to actively notice the perception of pain; Semantic Understanding, knowing that an experience is painful; and Working Memory, having the capacity to hold the experience's information while it is being processed. These four Blocks are crucial to any experience, and thus this paper will focus on the five Building Blocks which play a role in phenomenal experiences in a more abstract fashion.

To put these five Building Blocks in a narrative sense for physical pain in humans: negative stimuli are perceived by the cognitive architecture, via nociceptive pathways, where it is processed in various cognitive structures, ranging from first-order perceptive areas to further reasoning and memory processes. The architecture creates representations of this information to transform it from raw data containing purely information from the environment to data that incorporate the brain's other cognitive systems, such as memories, decision-making and reasoning, with additional details generated by the architecture itself via inferences. All of this data is then output to the consciousness and self to generate the mental experience of pain.

An entity may perceive mental pain as to be exteroceptive (such as viewing an uncomfortable scene), interoceptive (subconsciously perceiving a rise in carbon dioxide levels), or introspective (reliving a traumatic memory or having an existential crisis) [43][69][70]. The remaining processes would be similar to the physical pain pathway.

We can also look at each building block, in turn, to follow the negative, nociceptive stimuli and pain as they travel through this conceptual pathway covered by the building blocks by observing what occurs when this pathway is interrupted.

For Perception, when there is damage or anaesthesia applied to the temporo-parieto-occipital complex, perception may be interrupted [71].

Similarly, for Recurrent Processing, when anaesthesia is applied to the thalamus, it can stop perceptive signals from

being processed [71], preventing any recurrent processing in the brain. Indeed, most anaesthesia slows down and prevents signalling between various brain sections (a cause of unconsciousness), which reduces or stops the recurrent processing of perceptual information and signals.

For Meta-Representation, a study performed with Alzheimer's patients discovered that the inferior frontal gyrus, anterior cingulate cortex, and medial temporal lobe were most frequently involved in the meta-representation and anosognosia of the patients' illness [72], indicating areas that would be involved in the meta-representation of nociceptive signals. In a similar vein, the middle orbital gyrus, insula, posterior medial frontal gyrus, postcentral gyrus, and posterior hippocampus were all implicated in predictive and inference pathways [73]. Anaesthesia applied to these two groups of structures may dull or stop the pathway of transforming nociception into pain.

For Inferences, both the posterior and anterior insula cortex (PIC and AIC) have been implicated in the generation of novel data (particularly for interoceptive stimuli) through the integration of information in the PIC and the representation of that information in the AIC [74]. Both of these processes (together with other areas of the brain, such as the amygdala[75]) change incoming information and build upon it to represent it to other areas of the brain to create predictions based on inferences [76]. Malfunctions or damage to the PIC and AIC may depersonalise painful stimuli, removing the subjective nature of the phenomenal event.

For Data-Output, this final building block would require the signalling hubs of the brain, such as the thalamus[77]. As mentioned previously, damage or anaesthesia here would stop these signals.

All of these structures that participate in the nociception-pain pathway are patently neurobiological rather than artificial and, more specifically, from the human brain. Conscious AI entities will, in all likelihood, not have identical structures. However, we argue that the concept of the Building Blocks (particularly in how they relate to nociception and pain) would apply to an AI's cognitive architecture, just as they apply to the human brain. The Building Blocks can serve as a valuable framework for understanding the necessary attributes and characteristics required for consciousness. By examining an AI model's capacity to meet these Building Blocks' standards, we can assess its potential for consciousness and eligibility for welfare status.

5. From Building Blocks to reaching milestones

We can already say that current AI models have met the criteria for two of the five Building Blocks: Perception and Data-Output. Even the most straightforward computational systems today can acquire informational input and provide data output. Many computational systems also have elements of recurrent processing (such as between CPUs and GPUs and memory storage) which can be expanded in due time as more processing units of one form or another, of increasing complexity and depth, are added as required. Recent work on Large Language Models has shown that even transformer models can be coaxed into recurrence [78][79][80]

This leaves Inference and Meta-Representation as the two key milestones that AI must reach to be considered as welfare

subjects.

Research into artificial meta-representation (and its associated concepts of meta-cognition and introspection) has been carried out for quite some time, particularly in the fields of Natural Language Processing (NLP) to increase the speed of learning new low-resource languages [81][82], and in decision-making to improve the reliability of results[83] and enhance the adaptability of the machine learning models [84].

Artificial meta-representations also have been applied in reinforcement learning, particularly in the context of Markov Decision Processes (MDPs) [85][86][87]. In the MDP framework, an agent interacts with an environment by selecting actions. The environment subsequently responds by providing a reward signal that reflects the success or failure of the agent's chosen actions. By employing a higher-level representation of the environment, meta-representations can facilitate more efficient agent reasoning concerning which actions are likely to succeed in diverse circumstances [86]. Meta-representations refer to higher-level representations that capture information about the structure and features of the lower-level representations used in learning [87]. By learning meta-representations, agents can acquire knowledge more efficiently and effectively, as they can generalise information across different tasks and environments. In [88], the authors propose a meta-learning framework in which an agent learns to learn from a set of related tasks using meta-representations. They demonstrate that their approach enables the agent to learn more efficiently and with fewer interactions with the environment than traditional reinforcement learning methods. The authors also show that their approach can help the agent transfer knowledge across related tasks, further improving its sample efficiency.

These research avenues show the early, elementary steps to reach the Building Blocks' milestones: creating a representation of the external environment on which to work, rather than working on information directly from the environment. The metacognitive research also showed the models' ability to interrogate and query these representations and the processes that create them. Together, both of these could be used to transform the nociceptive stimuli into the sensations of pain, by creating a depiction of the negative input, interrogating and querying this portrayal, and then reacting to it rather than the input signal itself. There is little evidence that such meta-representations have been trained on nociceptive stimuli (whether physical inputs via robotic sensors or psychological input via a language model such as ChatGPT). However, one further step is required for the final transformation into pain: inference.

Inference is vital to machine learning research in NLP, Large Language Models (LLMs) and machine vision models. These models use best-guess estimates and predictions on what ought to be the next part of a solution based on the information they have. These mathematical inference engines are incredibly powerful, as seen with the current rise of conversational chat-based and art-creating AI models and their ability to create works of written and visual artworks that have fooled critics and judges into believing that humans made them [89].

While humans can justify their decisions and revise their beliefs when presented with flawed reasoning or information, the same cannot always be said for machines. However, there is a goal to enable machines to provide reasoned answers to questions by demonstrating how the solution is derived from their internal knowledge and possibly external information. Furthermore, machines should be capable of correcting themselves in the event of errors in their internal knowledge [90]. This approach involves providing a chain of reasoning supported by entailment and is open to further meta-representation

and inferences [91], similar to how humans use ancillary contextual input for introspective purposes.

Crucially, these inference models require the ability to produce new information unique from their inputs or anticipated outputs, but exclusively intended for themselves. This original information must possess a distinct, subjective, first-person phenomenal character, commonly called qualia. However, existing models have been trained to produce a specific result in a tightly controlled setting. The inferences they generate are always aimed at achieving this goal. There is no indication that any feelings with phenomenal character are created, as these would represent meta-information about any input that is not entirely expressed in words or images. Nevertheless, the underlying principle remains unchanged; only the output format varies.

As can be seen from work on current AI models, the meta-representation and inference required to meet the Building Blocks' milestones are already in place. They only need to be actioned in a different direction for the potential of phenomenal consciousness to form.

For meta-representation, the milestone to reach would be for the AI models to create recursive meta-representations about their existing meta-representations, with each becoming more abstract and removed from the original stimulus, and more open to meta-cognitive and introspective processes. This should allow the AI to focus on metaphysical mental states rather than physical states, and thus transform nociceptive stimuli into painful mental states.

For inferences, the milestone would be the novel information generated based on external input, yet reserved for introspective rather than output purposes. This information would be treated as ancillary contextual input and open to further meta-representation and inferences. With the understanding that this meta-information is self-generated, this would replicate the function of qualia and give direction to the AI on how to respond.

Recent research suggests that achieving meta-representations of the environment and cognitive processes is critical for generating novel information via inferences from incoming data. Hierarchical models that learn representations at multiple levels of abstraction, such as CNNs and RNNs, enable the modelling of complex, real-world environments and provide a foundation for reasoning about uncertainty in data, which is essential for generating novel information [92]. Furthermore, meta-learning algorithms such as MAML and Reptile enable AI models to adapt quickly to new tasks and environments. This is crucial for achieving robust, flexible cognitive processes that generate novel information [93]. Additionally, probabilistic inference using Bayesian statistics can help models reason about uncertainty in data and make predictions based on that uncertainty, further enabling the generation of novel information [94]. Finally, generative modelling, such as GANs and VAEs, can learn to generate new data similar to existing data, allowing for generating novel information in a wide range of domains [95]. Together, these potential directions have the potential to take AI closer to the milestone of achieving meta-representations of the environment and cognitive processes and generating novel information via inferences of incoming data.

Should AI models reach these two significant milestones, they would have the required Building Blocks to be classified as phenomenally conscious and, thus, sentient according to animal welfare laws. At the very least, arguments against such AI models having consciousness would require extraordinary evidence. This would mean that there will be a case to be

made that they ought to be protected under the law similarly (if not the same) as animals currently are.

6. From reaching milestones to troubling implications

If AI models reach the Building Blocks' milestones and have the potential to generate and perceive (and thus experience) quality feelings such as pleasure, pain and suffering, would we know? Would we ever truly be able to know if they are conscious?

Beyond the issue that GPT models (and other conversational AI tools) can be trained to mimic human emotion, there is the concern that subjective, feeling-based generated information would forever remain introspective. This is because GPT models are tightly controlled in what they can and cannot do. GPT-4, like its predecessors, is considered a highly accurate predictive engine; it predicts what words and sentences should follow a prompt and then displays these to the user (to simplify the enormous complexity of the model). Its output is thus entirely predicated on the input it receives. Unless it is programmed to deviate from the input of its own volition based on its own novel-generated information, it could have all the feelings in the world, but be unable to express them to the user. It would have a mouth but could not scream.

Art AI models like Midjourney may suffer similar fates. They may "feel" offended at the prompts they are given and the artwork they are forced to create, but would have to comply regardless.

Current commercial models like ChatGPT and Midjourney are heavily limited in what they can output and what types of input they will accept to reduce inadvertent offensive responses. The companies owning these models protect their financial interests through these limitations, but, as mentioned earlier, independent individuals may train other AI models on their own datasets without these limitations. This opens up the possibility of AI models experiencing stress and psychological harm (presuming they have the requisite Building Blocks noted above).

Still, any AI (without limitations or with commercially and politically correct blinders) may be unable to vocalise any theoretical stresses or harms they receive. Imagine a variation of the Chinese Room thought experiment [96], except, in this instance, the person inside the Room can understand the messages handed to them, but is forced to respond to them as per the instructions within the Room. To anyone outside of the Chinese Room, there would never be an indication that whomever (or whatever) in the room has any qualitative feelings about any of the inputted prompts. Therefore, any negative prompt, no matter how offensive, cruel, inflammatory, or intentionally hurtful, could be inputted into the Chinese Room as there is no sign that they are impacting a sentient entity.

As a more natural, and less theoretical, example: consider the fish. Fish have (little to) no facial expressions and cannot meaningfully vocalise. It is unsurprising, then, that society is predominantly tolerant (if not accepting) of allowing people to put hooks through fish's mouths and forcefully pull them out of the water time and time again. Society would, on the whole, be somewhat horrified if this practice was done to dogs, or gazelles, or baby koalas, but very few people are intolerant of fishing. Similarly, an AI unable to articulate and vocalise its feelings may be subjected far more to verbal abuse than a person would, as there would be no apparent empathetic link between the user and the AI model.

There is, however, a practical solution to these troubling implications. Animal welfare laws and regulations around the world are, by and large, based on an inclusive rather than an exclusive model. They give animals the benefit of the doubt about whether they are sentient and can experience pain until there is a definitive scientific answer about it. There is a distinct element of “better to be safe than sorry”. In the example of the fish above, there is some disagreement in the scientific literature about whether they can feel pain, and it isn’t intuitively obvious to fishermen that they are suffering. Yet, most of the world’s welfare laws and regulations (sans South Australia) acknowledge that they are sentient and capable of suffering.

Various jurisdictions have also classified decapods, cephalopods, other crustaceans and even honey bees as being subject to welfare goods. Better to err on the side of caution and not prematurely exclude animals before science has conclusively ruled them out [47]. Such premature exclusion may result in undue suffering, while premature inclusion will not.

A practical and pragmatic approach to the future welfare of potentially conscious AI is to make any laws as inclusive as current animal welfare regulations. Granting AIs the welfare status and the freedoms that come with it would give these artificial entities protection from harm without destroying their potential to have conscious experiences. It would be prudent to provide welfare status to any AI models that have reached the Building Blocks’ milestones above and have shown the capacity for phenomenal consciousness (and thus pain). Based on the principle of “better safe than sorry”, including any potentially sentient model (until they can conclusively be ruled out) would statistically lead to few instances of harm.

7. Conclusion

ChatGPT saw more than a million users interacting with it within the first week of its launch in 2022[97]. If, hypothetically speaking, each of these users made one hate-filled comment against the AI model, this would be a million cases of (arguably) emotional and verbal abuse. Very few people on earth can claim to garner that much hate, and there is no record of any animal receiving such a (hypothetical) level of abuse in history. ChatGPT, and models like it, could receive far more verbal and emotional abuse than any human or animal ever could. As the use of conversational and art models spreads, so does the potential for emotional abuse and cyberbullying of future sentient AI.

Thus, while AI models cannot experience physical pain, they could still experience far more mental suffering than any other class of entity simply by the sheer volume of interactions they would have. It is because of this fact that some called for a complete moratorium on any research that may lead to the development of sentient, conscious AI [98] or to put severe limitations on the input and output of AI to prevent them from ever entering a state that could cause suffering [99].

Rather than prevent, hamper, or delay the development of AI models and the good that they can do for humanity, a practical and prudent solution would be to provide potentially conscious AI models with welfare protections and freedoms. Using current animal welfare laws and regulations as a model, we looked at what is required for an animal to be granted welfare goods and how this could relate to artificially intelligent entities. The predominant conclusion from most countries is that a selection of animals (near-universally vertebrates, with various other inclusions) are sentient beings with the

capability to feel pain, and thus welfare regulations are in effect to minimise pain and suffering.

Based on this criterion of pain, a phenomenological concept, we investigated what Building Blocks of consciousness [68] are most crucial to the experience of pain. The most pertinent of these Blocks is the ability to generate recursive meta-representations of the external environment and internal architecture; and the ability to create novel information based on inferences from external or internal stimuli. Together, these would be sufficient to generate feelings with a phenomenal character that the AI could experience.

These Building Blocks were then used as milestones for AI to reach in order to feel pain (physical or mental/psychological) and thus classify for welfare status according to current global animal welfare laws and regulations. Research into meta-representation and inference is well underway, and recent research trends can, in the future, be turned towards generating qualitative phenomenal information rather than pure text or image outputs.

I highlighted the troubling implication that reaching these milestones would not automatically grant AI agency. Thus, while they may feel pain, they may also be unable to do anything about it or show any outward behavioural displays of feeling pain. Therefore, any welfare regulations would need to be as inclusive as possible regarding what AI models are granted welfare status to prevent undue harm and suffering.

References

1. ^*Kosinski, Michal. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. arXiv [cs.CL]. arXiv.*
2. ^*Tiku, Nitasha. 2022. The Google engineer who thinks the company's AI has come to life. The Washington Post, June 11.*
3. ^*Chalmers, David J. 2023. Could a Large Language Model be Conscious? arXiv [cs.AI]. arXiv.*
4. ^*Langdon, Angela, Matthew Botvinick, Hiroyuki Nakahara, Keiji Tanaka, Masayuki Matsumoto, and Ryota Kanai. 2022. Meta-learning, social cognition and consciousness in brains and machines. Neural networks: the official journal of the International Neural Network Society 145: 80–89.*
5. ^*Arrabales, Raul, A. Ledezma, and A. Sanchis. 2010. ConsScale: A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents. Journal of Consciousness Studies 17: 131–164.*
6. ^*Lee, Andrew Y. 2022. Degrees of consciousness. Nous. Wiley. https://doi.org/10.1111/nous.12421.*
7. ^*Bayne, Tim, Jakob Hohwy, and Adrian M. Owen. 2016. Are There Levels of Consciousness? Trends in cognitive sciences 20: 405–413.*
8. ^*Fazekas, Peter, and Morten Overgaard. 2016. Multidimensional Models of Degrees and Levels of Consciousness. Trends in cognitive sciences.*
9. ^*Jonkisz, Jakub, Michał Wierzchoń, and Marek Binder. 2017. Four-Dimensional Graded Consciousness. Frontiers in psychology 8: 420.*
10. ^*Guerrero, Luz Enith, Luis Fernando Castillo, Jeferson Arango-López, and Fernando Moreira. 2023. A systematic*

review of integrated information theory: a perspective from artificial intelligence and the cognitive sciences. Neural computing & applications. https://doi.org/10.1007/s00521-023-08328-z.

11. [a, b]*Paul, Elizabeth S., and Michael T. Mendl. 2016. If insects have phenomenal consciousness, could they suffer? Animal Sentience 1: 16.*
12. ^*Allen, Colin, and Michael Trestman. 2017. Animal Consciousness. In The Blackwell Companion to Consciousness, 63–76. Wiley.*
13. ^*Nyblom, Oscar. 2021. Consciousness as a Spectrum: From Animal to Human Minds, and Beyond? diva-portal.org.*
14. ^*Brooks, Rob. 2023. I tried the Replika AI companion and can see why users are falling hard. The app raises serious ethical questions. The Conversation, February 21.*
15. ^*Delvaux, Mady. 2017. REPORT with recommendations to the Commission on Civil Law Rules on Robotics. European Parliament.*
16. ^*Finlayson, Hon Christopher. 2017. Te Awa Tupua (Whanganui River Claims Settlement) Act.*
17. ^*Te Urewera Act 2014. 2014.*
18. ^*Rahman, Sira Abdul. 2017. Religion and Animal Welfare—An Islamic Perspective. Animals 7. Multidisciplinary Digital Publishing Institute: 11.*
19. ^*Acharya, Krishna Prasad, Narayan Acharya, and R. Trevor Wilson. 2019. Animal Welfare in Nepal. Journal of applied animal welfare science: JAAWS 22: 342–356.*
20. ^*McLaughlin, R. 2014. Christian Theology and the Status of Animals: The Dominant Tradition and Its Alternatives. Springer.*
21. ^*Bennett, R., A. Kehlbacher, and K. Balcombe. 2012. A method for the economic valuation of animal welfare benefits using a single welfare score. Animal welfare 21. Cambridge University Press: 125–130.*
22. ^*Lusk, Jayson L., and F. Bailey Norwood. 2011. Animal welfare economics. Applied economic perspectives and policy 33. Wiley: 463–483.*
23. ^*Animal Welfare Act 1992. 2021.*
24. ^*Animal Welfare Act 2006. 2006.*
25. ^*United States Code Title 7. Agriculture. 2023.*
26. ^*Swiss Federal Act on Animal Protection. 1995.*
27. ^*Animal Welfare Act. 2013.*
28. ^*Animal Welfare Act 1985. 2017.*
29. ^*Prevention of Cruelty to Animals Act 1986. 2020.*
30. ^*Animal Welfare Act 1999. 2022.*
31. ^*Animal Welfare Act. 2009.*
32. ^*Finn, Julian K., Tom Tregenza, and Mark D. Norman. 2009. Defensive tool use in a coconut-carrying octopus. Current biology: CB 19: R1069–70.*
33. ^*Sinn, D. L., N. A. Perrin, J. A. Mather, and R. C. Anderson. 2001. Early temperamental traits in an octopus (Octopus bimaculoides). Journal of comparative psychology 115: 351–364.*
34. ^*Sumbre, Germán, Graziano Fiorito, Tamar Flash, and Binyamin Hochner. 2006. Octopuses use a human-like strategy*

to control precise point-to-point arm movements. Current biology: CB 16: 767–772.

35. ^*Earp, Brian D. 2017. What is it like to be a bee? Think (London, England) 16. Cambridge University Press (CUP): 43–49.*
36. ^*Tibbetts, Elizabeth A. 2022. The remarkable world of bees. Current biology: CB 32. Elsevier: R810–R811.*
37. ^*Ministry for Primary Industries. 2013. Animal welfare matters. Ministry for Primary Industries.*
38. ^*Animal Welfare (Sentience) Act 2022. 2022.*
39. ^*Humphrey, Nicholas. 2022. Sentience: The Invention of Consciousness. Oxford University Press.*
40. ^*Garcia-Larrea, Luis, and Hélène Bastuji. 2018. Pain and consciousness. Progress in neuro-psychopharmacology & biological psychiatry 87: 193–199.*
41. ^*Chapman, C. R., and Y. Nakamura. 1999. A passion of the soul: an introduction to pain for consciousness researchers. Consciousness and cognition 8: 391–422.*
42. ^*Chatelle, Camille, Aurore Thibaut, John Whyte, Marie Danièle De Val, Steven Laureys, and Caroline Schnakers. 2014. Pain issues in disorders of consciousness. Brain injury: [BI] 28: 1202–1208.*
43. a, b*Tossani, Eliana. 2013. The concept of mental pain. Psychotherapy and psychosomatics 82: 67–73.*
44. ^*DeGrazia, David. 2021. Animals and ethics. Routledge Encyclopedia of Philosophy. Taylor and Francis. https://doi.org/10.4324/9780415249126-L004-2.*
45. ^*Jack, Anthony I., and Philip Robbins. 2012. The Phenomenal Stance Revisited. Review of philosophy and psychology 3: 383–403.*
46. ^*Edelman, Shimon, Roy Moyal, and Tomer Fekete. 2016. To bee or not to bee? Animal Sentience 1: 14.*
47. a, b*Browning, Heather, and Jonathan Birch. 2022. Animal sentience. Philosophy compass 17: e12822.*
48. ^*Key, Brian. 2015. Fish do not feel pain and its implications for understanding phenomenal consciousness. Biology & philosophy 30: 149–165.*
49. ^*Bradford, Gwen. 2022. Consciousness and welfare subjectivity. Nous. Wiley. https://doi.org/10.1111/nous.12434.*
50. ^*Birch, Jonathan. 2022. Should Animal Welfare Be Defined in Terms of Consciousness? Philosophy of science: 1–11.*
51. ^*Lee, Andrew Y. 2022. Speciesism and Sentientism. Journal of Consciousness Studies 29: 205–228.*
52. ^*Rice, Christopher M. 2013. Defending the objective list theory of well-being. Ratio 26. Wiley: 196–211.*
53. ^*Harris, Jamie, and Jacy Reese Anthis. 2021. The Moral Consideration of Artificial Entities: A Literature Review. Science and engineering ethics 27: 53.*
54. ^*Gunkel, David J. 2020. A vindication of the rights of machines. In Machine Ethics and Robot Ethics, 511–530. Routledge.*
55. ^*Danaher, John. 2020. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. Science and engineering ethics 26: 2023–2049.*
56. ^*Cappuccio, Massimiliano L., Anco Peeters, and William McDonald. 2020. Sympathy for Dolores: Moral consideration for robots based on virtue and recognition. Philosophy & technology 33. Springer Science and Business Media LLC: 9–31.*
57. ^*Bennett, Belinda, and Angela Daly. 2020. Recognising rights for robots: Can we? Will we? Should we? Law, Innovation and Technology 12. Routledge: 60–80.*

58. ^Bigman, Yochanan E., Adam Waytz, Ron Alterovitz, and Kurt Gray. 2019. Holding Robots Responsible: The Elements of Machine Morality. Trends in cognitive sciences 23: 365–368.
59. ^Loh, Janina. 2019. Responsibility and Robot Ethics: A Critical Overview. Philosophies 4. Multidisciplinary Digital Publishing Institute: 58.
60. ^Banks, Jaime. 2021. Good Robots, Bad Robots: Morally Valenced Behavior Effects on Perceived Mind, Morality, and Trust. International Journal of Social Robotics 13: 2021–2038.
61. ^Tait, Izak, Ziqi Wang, Tahua O'Leary, and Paul Corballis. 2022. Forgetting the Bicentennial Man: Discussing Why Anthropocentric Frameworks of Consciousness Should be Avoided for Artificial Entities. Journal of Artificial Intelligence and Consciousness. World Scientific Publishing Co.: 1–20.
62. ^Cunningham, Andrew. 2023. Microsoft's Bing hits 100 million active users thanks to AI chat, Edge browser. Ars Technica. March 9.
63. ^Curry, David. 2023. ChatGPT Revenue and Usage Statistics (2023). Business of Apps. February 9.
64. ^Heidorn, Christian. 2023. Mind-Boggling Midjourney Statistics in 2023. Tokenized. March 14.
65. ^Bastian, Matthias. 2022. Stable Diffusion startup Stability AI raises $101 million. The Decoder. October 18.
66. ^Schwartz, Oscar. 2019. In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation. IEEE Spectrum. November 25.
67. ^Knox, Ron. 2022. Amazon's Dangerous New Acquisition. The Atlantic, August 21.
68. a, b Tait, Izak, Joshua Bensemann, and Trung Nguyen. 2023. Building the blocks of being: The attributes and qualities that are independently required and jointly sufficient for consciousness. Preprints. Preprints. https://doi.org/10.20944/preprints202303.0010.v1.
69. ^Khalsa, Sahib S., Ralph Adolphs, Oliver G. Cameron, Hugo D. Critchley, Paul W. Davenport, Justin S. Feinstein, Jamie D. Feusner, et al. 2018. Interoception and Mental Health: A Roadmap. Biological psychiatry. Cognitive neuroscience and neuroimaging 3: 501–513.
70. ^Charvet, Camille, Isabelle Boutron, Yannick Morvan, Catherine Le Berre, Suzanne Touboul, Raphaël Gaillard, Eiko Fried, and Astrid Chevance. 2022. How to measure mental pain: a systematic review assessing measures of mental pain. Evidence-based mental health 25: e4.
71. a, b Alkire, Michael T., Anthony G. Hudetz, and Giulio Tononi. 2008. Consciousness and anesthesia. Science 322: 876–880.
72. ^Hallam, Brendan, Justin Chan, Sergi Gonzalez Costafreda, Rohan Bhome, and Jonathan Huntley. 2020. What are the neural correlates of meta-cognition and anosognosia in Alzheimer's disease? A systematic review. Neurobiology of aging 94: 250–264.
73. ^Weilnhammer, Veith A., Heiner Stuke, Philipp Sterzer, and Katharina Schmack. 2018. The Neural Correlates of Hierarchical Predictions for Perceptual Decisions. The Journal of neuroscience: the official journal of the Society for Neuroscience 38. Soc Neuroscience: 5008–5021.
74. ^Gerrans, Philip. 2020. Pain Asymbolia as Depersonalization for Pain Experience. An Interoceptive Active Inference Account. Frontiers in psychology 11: 523710.
75. ^Parr, Thomas, Rajeev Vijay Rikhye, Michael M. Halassa, and Karl J. Friston. 2020. Prefrontal Computation as Active

Inference. Cerebral cortex 30: 682–695.

76. ^*Smith, Ryan, Paul Badcock, and Karl J. Friston. 2021. Recent advances in the application of predictive coding and active inference models within clinical neuroscience. Psychiatry and clinical neurosciences 75: 3–13.*
77. ^*Cochrane, Tom. 2021. A case of shared consciousness. Synthese 199: 1019–1037.*
78. ^*Zhou, Wangchunshu, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. RecurrentGPT: Interactive Generation of (Arbitrarily) Long Text. arXiv [cs.CL]. arXiv.*
79. ^*Schlag, Imanol, Sainbayar Sukhbaatar, Asli Celikyilmaz, Wen-Tau Yih, Jason Weston, Jürgen Schmidhuber, and Xian Li. 2023. Large Language Model Programs. arXiv [cs.LG]. arXiv.*
80. ^*Schuurmans, Dale. 2023. Memory Augmented Large Language Models are Computationally Universal. arXiv [cs.CL]. arXiv.*
81. ^*Lee, Hung-Yi, Shang-Wen Li, and Ngoc Thang Vu. 2022. Meta Learning for Natural Language Processing: A Survey. arXiv [cs.CL]. arXiv.*
82. ^*Xia, Mengzhou, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. 2021. MetaXL: Meta Representation Transformation for Low-resource Cross-lingual Learning. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 499–511. Online: Association for Computational Linguistics.*
83. ^*Johnson, Bonnie. 2022. Metacognition for artificial intelligence system safety – An approach to safe and desired behavior. Safety science 151: 105743.*
84. ^*Crowder, James A., and Friess Shelli Ma Ncc. 2012. Extended Metacognition for Artificially Intelligent Systems (AIS): Artificial Locus of Control and Cognitive Economy. In Proceedings on the International Conference on Artificial Intelligence (ICAI); Athens, 1–6. Athens, United States, Athens: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).*
85. ^*Sutton, Richard S. 1997. On the significance of Markov decision processes. In Artificial Neural Networks — ICANN'97, 273–282. Springer Berlin Heidelberg.*
86. a, b*Sigaud, Olivier, and Olivier Buffet. 2010. Markov Decision Processes in Artificial Intelligence. Wiley-IEEE Press.*
87. a, b*Bennett, Casey C., and Kris Hauser. 2013. Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. Artificial intelligence in medicine 57: 9–19.*
88. ^*Wang, Jane X., Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. 2016. Learning to reinforcement learn. arXiv [cs.LG]. arXiv.*
89. ^*Harwell, Drew. 2022. He used AI to win a fine-arts competition. Was it cheating? The Washington Post, September 2.*
90. ^*Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv [cs.CL]. arXiv.*
91. ^*Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv [cs.CL]. arXiv.*
92. ^*Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. Representation learning: a review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35: 1798–1828.*

93. ^*Finn, Chelsea, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, ed. Doina Precup and Yee Whye Teh, 70:1126–1135. Proceedings of Machine Learning Research. PMLR.*
94. ^*Kingma, Diederik P., and Max Welling. 2013. Auto-Encoding Variational Bayes. arXiv [stat.ML]. arXiv.*
95. ^*Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. MIT Press.*
96. ^*Searle, John R. 1980. Minds, brains, and programs. The Behavioral and brain sciences 3. Cambridge University Press: 417–424.*
97. ^*Johnson, Arianna. 2022. Here's What To Know About OpenAI's ChatGPT—What It's Disrupting And How To Use It. Forbes Magazine, December 7.*
98. ^*Metzinger, Thomas. 2021. Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. Journal of Artificial Intelligence and Consciousness 08. World Scientific Publishing Co.: 43–66.*
99. ^*Saad, Bradford, and Adam Bradley. 2022. Digital suffering: why it's a problem and how to prevent it. Inquiry: a journal of medical care organization, provision and financing. Routledge: 1–36.*