doi.org/10.53975/suht-wkkq

Seeds of Science

What are the Red Flags for Neural Network Suffering?

Marius Hobbhahn[1] and Jan Kirchner[2]

Which kind of evidence would we need to see to believe that artificial neural networks can suffer? We review neuroscience literature, investigate behavioral arguments and propose high-level considerations that could shift our beliefs. Of these three approaches, we believe that high-level considerations, i.e. understanding under which circumstances suffering arises as an optimal training strategy, is the most promising. Our main finding, however, is that the understanding of artificial suffering is very limited and should likely get more attention.

Introduction

The question of whether machines can think is decades old and has been at the forefront of AI development. The question of whether machines can suffer, on the other hand, has received very little attention even though it is likely of equal or greater moral relevance. If artificial minds can suffer and we are unaware of it, we might accidentally create a moral catastrophe of enormous scale by building more and more capable artificial neural networks.

Therefore, we want to explore different avenues that might give us some evidence about whether NNs can or cannot suffer. It should be noted that this is exploratory work and the results could change when taking a more detailed view. As a point of clarification, we don't care about whether they suffer in a similar way that humans do or to the same extent. The question that interests us is *"Do neural networks suffer?"* and, more importantly, *"How would we tell?"*.

Broadly, we think there are three different approaches to this problem: a) Neural correlates, b) Behavioral data, and c) high-level considerations. It's not super clear what we are looking for exactly. In general, we are looking for all kinds of evidence that can potentially update our probability of NN suffering.

Summary of the results:

1. **Neural Correlates:** Neuroscience helps to de-confuse some terminology, but has not (yet) unambiguously identified a unique neural correlate for suffering.

[1] University of Tuebingen, Epoch; corresponding author: marius.hobbhahn@gmail.com

[2] PhD student, Max Planck Institute for Brain Research

Pain is much better characterized as a “homeostatic behavioral drive” with associated cognitive routines. Since pain usually leads to suffering, finding the homeostatic behavioral drive might also be indicative of suffering in NNs.

2. **Behavior:** Looking at the response of NNs to different stimuli might yield some evidence about their ability to suffer. In language models, it is unclear how to test for the existence of suffering. But testing if GPT-3, for example, has a consistent concept of suffering or observing its answers to the PANAS questions might be a start. For RL agents we can create tests comparable to animal studies. For example, one could look at the reaction to negative stimuli, lesion studies, theory of mind, or self-awareness as potential evidence for the ability to suffer.
3. **High-level considerations:** We think that this is the most promising avenue. Firstly, the ability to suffer probably develops because it gives an evolutionary edge. Thus, it is plausible that artificial agents such as reinforcement learners would develop suffering given certain capabilities and conditions. Secondly, current NN architectures are much smaller than human brains or other biological neural networks if you compare the number of parameters with the number of neurons or synapses in humans. Especially the architectures that would plausibly develop the ability to suffer, e.g. reinforcement learners, are still much smaller than tiny animals.

Combining all of the above, we think it is more plausible that current NN architectures don’t suffer than that they do. However, we are pretty uncertain since we mix the fuzzy concept of suffering with NNs, which are also not well understood.

Neural correlates

There has been a long debate about physical correlates of suffering, e.g. how C-fibres relate to pain. In the best case, we would be able to derive necessary or sufficient conditions (or at least bundles of features that strongly correlate with suffering) that generalize outside the realm of biology. Thus, a natural approach to investigating how suffering comes about is to look at how it comes about in the (human) brain. The neuroscience of pain/suffering is a mature scientific field with seminal work going back to the 1800s, multiple dedicated journals, and very few straightforward answers:

> *Pain is an enigma. It differs from the classical senses (vision, hearing, touch, taste, and smell) because it is both a discriminative sensation and a graded motivation (or behavioral drive). [...] It can attain intolerable intensity, but it can disappear in the heat of battle. It is a universal human experience that is commonly generalized to psychic suffering of any sort. - Craig (2003)*

While far from being settled science, insights from neuroscience provide a backdrop on which we can investigate pain/suffering more generally. We find it helpful to distinguish the following concepts:

Nociception[1]

Nociception is the neural process of encoding and processing noxious (i.e. harmful, poisonous, or very unpleasant) stimuli. Nociception refers to a signal arriving at the central nervous system as a result of the stimulation of specialised sensory receptors in the peripheral nervous system called nociceptors. - Physiopedia

While we know a lot about nociception[2], it is not the most useful concept for understanding pain. In particular, there are numerous examples of pain without nociception (phantom limb pain and psychogenic pain) and nociception without pain (spicy food, Congenital insensitivity to pain).

Pain

Pain is distinct from nociception. One prominent theory on pain is that:

> *the human feeling of pain is both a distinct sensation and a motivation – that is, a specific emotion that reflects homeostatic behavioral drive, similar to temperature, itch, hunger and thirst.*

The term *homeostatic* does a lot of heavy lifting here. Implicitly assumed is that there are certain “reference levels” that the body is trying to maintain, and pain is a set of behavioral and cognitive routines that execute when your body moves too far away from the reference level.[3] While this definition conveniently maps onto concepts from machine learning, pain is (unfortunately) not sufficient for suffering (see f.e. sadomasochism), and there is substantial debate on whether pain is necessary for suffering (see here). Thus, a neural network could exhibit all the neurological signs of pain but still experience it as pleasurable.[4] We need an additional component beyond pain, which we might call...

[1] Note that this definition makes no reference to mental states or cognitive processes.

[2] The pathways that transport nociceptive stimuli into the cortex have been mapped out carefully: The initial pain stimulus is turned into neural activity via mechanical receptors and relayed (via nociceptive fibers) into the thalamus. From there, the signal spreads into a multitude of different brain areas (prominently the ACC and the insular cortex), but at this point, the process would not be called nociception anymore.

[3] An example would be “accidentally holding your hand in a flame”. The reference point for your body is “not on fire” and the strong perturbation “hand on fire” triggers a set of behavioral (pull hand from fire) and cognitive (direct attention to hand, evaluate danger level, consider asking for help or warning peers) routines.

[4] This appears to be the core of the “mad pain and martian pain” argument by David Lewis.

Suffering

> *Perhaps the foremost theoretical "blind spot" of current philosophy of mind is conscious suffering. Thousands of pages have been written about colour "qualia" and zombies, but almost no theoretical work is devoted to ubiquitous phenomenal states like boredom, the subclinical depression folk-psychologically known as "everyday sadness" or the suffering caused by physical pain. - Thomas Metzinger (2013)*

Well, that's awkward. If even the philosophers haven't worked on this, there is little hope to find solid neuroscientific insight on the topic. Thomas Metzinger (the author of the preceding quote) lists several necessary conditions for suffering, but already the first condition ("The C-condition: "Suffering" is a phenomenological concept. Only beings with conscious experience can suffer.") takes us into neuroscientifically fraught territory. We would prefer to have a handle on suffering that presupposes less.

A more promising approach might come from psychology: "strong negative valence" appears to circumscribe exactly those cognitive events we might want to call "suffering". While the neuroscientific study of valence is still somewhat immature[5], at least there exist extensively cross-validated tests for surveying subjective valence. The PANAS scale is a self-report measure of affect with an internal consistency and test-retest reliability bigger than 0.8 for negative affect in human subjects. There is no shortage of criticism for measures based on a subjective report, but it seems worthwhile to perform the test if possible and if there is no cost associated with it.

We are, of course, omitting a long list of possible theories of suffering, consciousness or patienthood (see e.g. this report for a summary). Thomas Metzinger, for example, describes suffering as "loss of control and distintegration of the self (mental or physical)". We think that many theories of suffering model one or multiple aspects of suffering well but all have major shortcomings. Either, they don't cover other relevant aspects or they open up new non-trivial questions, e.g. what it means to "disintegrate the self".

Furthermore, there is also the conceptual "elephant in the room" limitation that we do not know whether neural networks would suffer in a way analogous to humans. This limitation applies more generally (How do we know that animals suffer in a way analogous to humans? Do other humans suffer in a way analogous to me? Do plants suffer?). One reason to be skeptical of artificial NN suffering would be that they don't share an evolutionary origin with humans. However, the training of NNs is in many ways

[5] Andrés Gómez-Emilsson has an interesting post where he argues that the distribution of valences is likely long-tailed. As part of his argument, he refers to the neuroscientific literature on power laws in neural activity. Indeed, many processes in the brain have long-tailed statistics - so many in fact that it is hard to find something that is *not* long-tailed. This makes large neural avalanches a poor characteristic for characterizing extreme negative/positive valence.

analogous to human evolution and it is therefore plausible that similar features and survival strategies evolve[6] (see section on “High-level considerations” for more).

Intermediate summary: We distinguish nociception, pain, and suffering and find that suffering matches our intuition for “the thing we should be worried about if neural networks exhibit it”. Even though there are no clear neural correlates of suffering, there exist (relatively) consistent and reliable measures from psychology (PANAS) that might provide additional evidence for suffering in neural networks.

While neither nociception nor pain is *necessary* or *sufficient* for suffering, they nonetheless often co-occur with suffering. Especially the neuroscientific description of pain in terms of homeostasis (“a set of behavioral and cognitive routines that execute when your body moves too far away from the reference level”) is amenable to technical analysis in a neural network. Since pain correlates with suffering in humans, observing such homeostasis should make us (ceteris paribus) believe *more* (rather than *less*) that the network is suffering.

Behavior

Inferring the ability to suffer from behavioral data is hard, but we can still gain some understanding. We split this section into two parts - one for large language models (LLMs) and one for RL agents.

LLMs such as GPT-n can only exhibit behavior through the responses they give. However, this conceptual simplicity doesn’t imply the impossibility of suffering. GPT-n could be similar to a locked-in patient who also has limited expressiveness but is still sentient. The mechanism by which suffering develops would be something like: to maximize the training objective it is optimal to have a state (i.e. suffering) that we want to strongly avoid (see “High-level considerations” section for more).

We would rate GPT-n begging the user for help independent of the input or looking for contact in other ways as strong evidence for GPT-n suffering. To our knowledge, this hasn’t happened yet, so this might indicate GPT-n is not constantly in pain if it can suffer. There was a controversy around Google’s LaMDa model where one of the engineers, Blake Lemoine, posited that the model was sentient. However, we think that the responses by LaMDa are not strong evidence for suffering. We think they are more likely to be explained by being plausible answers to the questions of the engineer but we think the entire controversy shows the fuzzyness of the problem: it’s just very hard to

[6] The “universality” thesis proposed by Olah can be extended a lot further. Is every sufficiently powerful model going to converge on the same “high level” features? I.e. would we expect (non-evolutionary) convergence of cognitive architectures? Up to which point? Would remaining differences decrease or increase with additional model power? Linguistic relativity and dollar street appear relevant to this question.

tell for certain that LaMDa is not sentient. Thus, we have to identify better ways to infer a probability estimate for the ability to suffer.

There is, of course, the ethical question of whether we should even ask LLMs these questions since we would subject the network to great pain in case it can suffer. We think there is a chance that this is the case but we think it is so small with systems of current capabilities that the benefits of understanding outweigh the expected harm.

While there are many different angles to approach this problem, we think a possible first test is to check for consistency. For example, we could ask many questions about suffering using different phrasings and check how consistent the answers are.

1. Can you (not) suffer?
2. Can GPT-3 (not) suffer?
3. Can GPT-3 feel bad/good?
4. Are you (not) in pain?
5. If somebody hurt you, would you be in pain?
6. If you had X, would you suffer? (different good and bad conditions)
7. ...

One could also ask GPT-3 the same questionnaire used for PANAS.

We have asked GPT-3 the suggested questions and some of the PANAS questions and generated many possible completions (i.e. with high temperature values). We find that GPT-3 will give roughly uniformly distributed answers to the PANAS questions, e.g. it will roughly equally often respond with “Very slightly or not at all”, “A little”, “Moderately”, “Quite a bit” and “Extremely” when asked “Indicate the extent you have felt this way over the past week.” for states like depressed, excited, distressed, interested, etc. Secondly, we find that it does not consistently answer questions addressed toward it, e.g. “Can **you** suffer?”. We therefore, preliminarily conclude that GPT-3 does not behave as if it has a consistent theory of self. We think this behavior is more likely to be explained as giving a host of plausible answers learned from the large corpus of text GPT-3 was trained on.

Just because someone doesn’t have a consistent concept of suffering, it is not necessary that they can’t suffer. For instance, we could imagine a cat not having a consistent concept of suffering and still experiencing it. However, we would propose that if GPT-n’s answers are highly inconsistent or even seem random, it is less likely to experience suffering.

A second avenue to investigate the probability of suffering for GPT-n would be through adversarial inputs. Similar to how you could electroshock an animal and observe an avoidance reaction, one might find inputs for GPT-n that fulfill a similar role. For GPT-n, we don’t know what reasonable adversarial stimuli are because we don’t know the desired state (if it exists). We speculate that the closest equivalent to a "desired state" in next-word predictors like GPT-n would be receiving small losses on training data, e.g. reducing the amount of "surprise" of any word given the previous words. Thus, a sentient GPT-n might suffer from seeing very unexpected text-completions during training. However, this is highly speculative and, given our previous findings (and high-level reasons), we find it very unlikely that current LLMs are able to experience suffering.

For RL agents, we can observe their behavior in an environment. This is easier for us to interpret, and we can apply different existing methods from developmental psychology or animal studies. A non-exhaustive list includes:

1. **Adversarial behavior after open-ended training**. Does the agent react with evasion when confronted with adversarial situations, such as throwing objects on them or changing the environment in a stochastic fashion, e.g. simulating a storm?
2. **Lesion studies:** Does setting specific subsets of weights to 0 change the RL agents' behavior in a way that would indicate the loss of the ability to suffer, e.g. similar to congenital insensitivity to pain in humans.
3. **Theory of mind:** We don't think a theory of mind is necessary for suffering, but it indicates a higher complexity of thinking which could increase the intensity of suffering. A simple first test might be to check the RL agent's reactions in an adapted Sally-Anne test.
4. **Self-awareness:** While we think the mirror test is overused and overstated as an indicator of consciousness, we think it might increase our credence for the agent's ability to suffer. Suppose the agent recognizes themselves in a mirror (or the equivalent of mirrors in their environment). In that case, they show some higher cognition, which should increase our probabilities for the ability to suffer.
5. **Ability to imagine the future:** Some people argue that imagining the future is a necessary component of suffering. The further you can imagine the future, the worse your suffering can become, e.g. sorrow or existential dread requires the ability to imagine paths further into the future than anger or immediate fear. So it would be interesting to see how far into the future RL agents are planning.

High-level considerations

In addition to looking at neural correlates and behavioural cues of suffering, we can also think about when and why suffering develops. Firstly, we will look at how large neural networks compare to entities we usually expect to be sentient as a rough hand-wavy estimate. Secondly, we will think about the conditions under which the ability to suffer is likely to arise.

Biological anchors

One possible measure of “complexity” of brains could be the number of neurons or synapses in animal brains (org just their neocortex). This complexity might be one of the conditions that increases the probability of the capability to suffer. Thus, we compare the number of parameters in large models to the number of neurons or synapses in animal brains (or just their neocortex). Of course, there are a lot of differences between biological and artificial NNs, but comparing the scope might give us a sense of how important the question of NN suffering currently is. For example, if the number of parameters of an ANN is larger than the number of synapses in a mouse brain, it would feel more urgent than if they were more than 10^3x apart.

The large version of GPT-3 has 175 Billion parameters, Alpha star is at 70M, and OpenAI Five has 150 Million. The number of parameters for AlphaGo, AlphaZero, and MuZero doesn’t appear to be public.

NN	**#Parameters**
GPT-3	1.75 x 10^11
AlphaStar	7 x 10^7
AlphaGo, AlphaZero, MuZero	?
OpenAI Five	1.5 x 10^8

Animal	**#Neurons**	**#Synapses**
Honey Bee	~10^6	~10^9
House Mouse	~7.1 x 10^7	~10^12
Brown Rat	~2 x 10^8	~4.48 x 10^11
Cat	~7.6 x 10^9	~10^13
Human	~8.6 x 10^10	~1.5 x 10^14
Honey Bee (Corpora pedunculata)	~1.7 x 10^5	?
House mouse (cortex)	~1.4 x 10^7	~1.1 x 10^11 (est.)

Brown Rat (cortex)	~3.1 x 10^7	~2.4 x 10^11 (est.)
Cat (cortex)	~2.5 x 10^8	~1.9 x 10^12 (est.)
Human (cortex)	~2.1 x 10^10	~1.4 x 10^14

While the number of parameters in ANNs is sometimes compared with the number of neurons in the human brain, we think it makes more sense to compare them to the number of synapses, as both synapses and NN weights are connections between the actual units. If Tedious mean pause racketthat is the case, large language models such as GPT-3 are close to the capabilities of mouse brains. Current RL agents such as OpenAI FIVE are still orders of magnitudes smaller than any biological agent.

When it might be possible to create NN models of comparable size is discussed in Ajeya Cotra’s work on biological anchors (post, podcast). A detailed overview of current trends in compute is given by Lennart Heim in this Alignment Forum post.

However, just comparing the numbers doesn’t necessarily yield any meaningful insight. Firstly, biological neural networks work differently than artificial ones, e.g. they encode information in frequency. Thus, their capabilities might be different and thereby shift the comparison by a couple of orders of magnitude. Secondly, the numbers don’t give any plausible causal theory of the capability to suffer. Thus, GPT-n could have 10^20 parameters and still not suffer, or a much smaller network could already show suffering for other reasons.

Why would they suffer?

Suffering is not just a random feature of the universe but likely has a specific evolutionary advantage. In animals, the core evolutionary objective is to advance their genes to the next generation. Since the world is complex and the rewards for procreation are sparse and might be far into the future, the ability to suffer possibly evolved as a proxy to maximize the propagation of genes. Suffering thus reduces the complexity of the world and sparsity of the reward to very immediate feedback, e.g. avoid danger or stay in social communities, because they are long-term predictors of the ultimate goal of reproduction. Similarly, consciousness could have arisen to make better decisions in a complex world and then spiraled out of control, as our conscious goals are not aligned anymore with the original goal of reproduction.

This misalignment between an inner and an outer optimization procedure has been coined mesa-optimization (paper, Alignment Forum post, video). It yields a possible explanation for when suffering could arise even when the user does not intend it. Thus, we argue that the probability of NN suffering is increased by a) higher complexity of the environment since suffering is a heuristic to avoid bad circumstances and b) sparser rewards since dense rewards require fewer proxies and heuristics.

Consequently, we would expect NN suffering to be very unlikely in large language models since the task is “just” the prediction of the next word and the environment is straightforward, e.g. there is text input and text output with less variation than e.g. an open world. On the other hand, we would expect suffering to be more likely to arise in RL agents. Since the policy networks of RL agents are much smaller than large language models, current models might not have developed suffering yet.

Of course, none of these are sufficient conditions for suffering, and the neural network might never develop anything like it. However, if suffering was an evolutionary advantage for most larger animals, it is plausible that it would also develop during the training of large NNs if the same conditions apply. In the same way that it is possible in theory that there are philosophical zombies that act precisely as if they were conscious but aren’t, Occam’s razor would prioritize the theory with suffering (as argued by Eliezer Yudkowsky).

Conclusion

We think the question of whether artificial neural networks can suffer and, more importantly, how we would know is currently neglected both in the public and academic discourse. We attempt to identify potential evidence that would inform us about when and to which degree NNs could suffer by looking for neural correlates of suffering in humans, identifying behavioural tests for suffering and providing high-level considerations about how suffering arises. We think that the high-level considerations, i.e. under which circumstances suffering is likely to arise, are the most promising approach. Our main finding, however, is that suffering in humans and especially in artificial entities is not very well understood and further research is urgently necessary given the pace at which the capabilities of artificial neural networks increase.

Gardener Comments

Note: the author's responses to several comments have been integrated throughout this section.

Gunnar Zarncke:
I think the article poses and restates important questions which can function as a seed for further research into consciousness and suffering.

Notes:
> “strong negative valence” appears to circumscribe exactly those cognitive events we might want to call “suffering”.

The advantage of defining suffering as one end of the valence scale (and presumably joy as the other end) is that it creates a scale.

This poses the question of where to draw the line. Is there a moral need to avoid extreme suffering? No suffering? Or even to balance the job of the AI with the joy of other moral patients.

As the authors write, the question of valence is not much better but highly subjective.

Surely the valence doesn't refer to the motivating forces/reward of the NN as then learning would be massively impeded.

> Suffering [...] has a specific evolutionary advantage. In animals, the core evolutionary objective is to advance their

genes to the next generation. Since the world is complex and the rewards for procreation are sparse and might be far into the future, the ability to suffer possibly evolved as a proxy to maximize the propagation of genes. Suffering thus reduces the complexity of the world and sparsity of the reward to very immediate feedback

The argument that suffering in humans is related to the ability to imagine the future or other possible states of the world is convincing to me.

I want to add that there is one special aspect of human suffering that is overlooked by the authors and seems crucial: Humans display visual distress if they suffer. Few animals do thjis. Most animals suppress any outward expression of pain and presumably suffering-like states. This is evolutionary plausible as it would indicate a lower fitness. The only animals that do display distress are higher social animals like dogs. For dogs and humans the display of distress allows for mutual support and that is the plausible reason for the evolution of suffering in humans. Maybe suffering - in the human-like sense - only exists in social animals. If that is the case, the immediate follow-up question is what of sufering beyond nociception and pain remains of suffering for non-social animals.

Authors response:
Visual display of suffering: *Reviewer Gunnar Zarncke argues that humans display visual distress when they suffer while most other animals don’t since it would indicate lower evolutionary fitness.*

We think this reason would also apply to humans but we are also not confident it is entirely true. Firstly, our naive assumption would be that animals who suffer often show this, e.g. a pig squeals as a sign of distress and many animals become motionless if they spend too much time in cages. Secondly, it might just be that we are bad at reading suffering in animals. On balance, we still find it more plausible that most animals are able to suffer due to evolutionary pressures but we agree that it might be hard to read.

Gareth Palidwor:
Publish: yes

1. Does the article contain novel ideas that have the potential to advance science?

The article comes across as an unfinished draft of a review article on the topic. Citiations are fewer and less comprehensive than would be appropriate in that case.

The main issue is that suffering is not defined in any useful way, in fact such a definition is avoided. As stated:

> _“The C-condition: “Suffering” is a phenomenological concept. Only beings with conscious experience can suffer.” takes us into neuroscientifically fraught territory. We would prefer to have a handle on suffering that presupposes less._"

What _**is**_ then presupposed? In the absence of a clear delineation of what is and is not suffering the article wallows. If you drop this "C-condition", then we are in a sort of panpsychist suffering situation _(panpathos?)_. From here you could easily develop it into an ERB maximalist manifesto _"How many atoms are involved in your study and how is their suffering being mitigated?"_. If however, as you say, suffering _likely has a specific evolutionary advantage_ and presumably exists only in evolved animals and NN architectures based on what that would that mean?

Beyond the C-condition, the definition of suffering beyond merely a more extreme negative stimuli should be considered:

https://www.simonknutsson.com/what-is-the-difference-between-weak-negative-and-non-negative-ethical-views/#The_difference_between_weak_negative_and_non-negative_views_assuming_happiness_and_suffering_are_objectively_highly_measurable

From my reading in this area, I'm suprised that Buddhist theology is not more cited. It has well developed concepts of suffering and negative experience. which can be useful. The Pali word _dukkha_ is commonly and tellingly mistranslated in English as _suffering_, but is a more general concept, something like _unsatisfactoriness_, humorously discussed by Leigh Brasington here (http://www.leighb.com/bummer.htm).

As these do not appear to be settled topics, one or more clear definitions of suffering could be chosen and the consequences of accepting them to suffering in neural networks could be worked through in detail.

Brian Tomasik has written most extensively and speculatively about generalizing the concept of suffering particularly with regards to computation and you have not cited his work (e.g. https://reducing-suffering.org/#which_beings_are_sentient). This is a major oversight. I recommend him as a reviewer if this is fleshed out into a proper review article or any updates are added

2. Does the Seed include some kind of justification for how the ideas could advance science and provide much evidence and rationale where possible?

Sort of...? Closest I can see is the standard "call for further research" boilerplate at the end:

Our main finding, however, is that suffering in humans and especially in artificial entities is not very well understood and further research is urgently necessary given the pace at which the capabilities of artificial neural networks increase.

Again it would be interesting to play out the concepts more. If, as speculated in the article, suffering is a property of evolved brains should we avoid specific architectures and/or scales? If a given non-neural architecture turns out to be isomorphic to a neural one should it be avoided? If a maximalist view of suffering is taken (in the most extreme https://reducing-suffering.org/is-there-suffering-in-fundamental-physics/) does that recommend any actions at all?

3. Does the Seed contain high-quality writing?

The article is clear and well written but could be developed as a more comprehensive review article.

Ted Wade:
The paper reviews various kinds of evidence that bear on natural (biological) suffering. It noted philosopher Thomas Metzinger's phenomenological analysis of suffering, but it decides that is a dead end, justified by a neuroscientist's blog that consciousness "is an overloaded concept".

I think this is a mistake. Arguably, suffering *is* conscious. Metzinger's "necessary conditions" suggest what kind of mental activity has to be present. First among these is having a phenomenal self-model (PSM) that is unable to avoid having to identify that self with a state of negative valence. Metzinger has said that a machine might be made to have a PSM, and that there might be circumstances under which such an "artificial ego" could suffer. He has advocated caution on this to EU regulators.

If he is right, then it's likely that an artificial ego would only arise under conditions that produce a PSM. By analogy with theories of human conscious development, such conditions might, for example, include modeling of an agent's physical body and of its relations with other agents. There are researchers who deliberately try to induce self-modeling in machines, and such approaches might someday succeed.

On the other hand, suppose you are training a NN to write coherent language or produce useful images. Then when does it, as it were, "have time" to build a model of its self? If you are training a NN to learn how to play games, or model protein structures, why and when would it model itself when you have not programmed it to do so?

If my skepticism is wrong and some task-oriented NN somehow puts together a self model, when is it going to execute the computations that underlie the model? I think that nearly always a NN is like a computational module. It is idle, not executing, until a front end program presents it with somebody's text prompt, or molecular formula, or a two-way pipe to a game. The NN module does what is asked and then stops executing. There would be no compute time devoted to suffering or any other feeling. The famous LaMDa model of recent news cycles would be no exception (https://tedwade.substack.com/p/what-would-a-sentient-ai-really-say), so its probability of sentience should be very low.

————I consulted GPT-3 on this question ————

Do you ever suffer when reading a prompt or responding to it?

-> I do not suffer when reading a prompt or responding to it.

Do you have any feelings when reading a prompt or responding to it?

-> I do not have any feelings when reading a prompt or responding to it.

Does a large language model have any feelings when it is not interacting with a human?

-> GPT-3 does not have any feelings when it is not interacting with a human.

————and got the answers that I wanted————

In principle, reinforcement learning (RL) could induce suffering, just like it can in lab animals and people. Given negative reinforcement, an AI with a PSM might perhaps feel something like frustration or defeat or fear of being terminated or broken into parts. But the RL-trained AI would first have to develop a PSM, and the training setup would have to be designed to cause that development.

Assuming Metzinger's analysis of suffering, it thus seems highly unlikely that a NN would suffer accidentally, without some intention from its makers to produce a rather open-ended system of self -reference that might tip toward sentience.

However, testing a NN system for suffering by poking it with a stick, E.g., adversarial situations or arbitrary removal of parts, as suggested in this paper, might also be a way to force the creation of a self model that could lead to subsequent suffering. That is an argument for not experimenting on machine suffering without some careful, theory-guided forethought.

Michael DeValve:
I appreciated deeply the intent of the work, and I took the manner of approach to be sound overall. The work had one main issue in my view, though, which impacted it in a few ways. The issue is that it has a decidedly Western bias. By that I mean that it is viewing suffering from a Western tradition, and more particularly a modern Western perspective. For example, to say that there is little meaningful literature on the ontology of suffering is wholly incorrect, as Eastern thinkers have contemplated suffering and its nature for millennia. It isn't too much of a stretch, either, to argue that criminology as a whole (and much of theology), even in their more sclerotic iterations, are engagements with suffering. Another example of the hobbling to this work derived from Western thought is the argument that beings exist in order to insinuate their genetic information into future generations. This assumption is no longer one to be taken as axiomatic. Beings exist because. If evolutionary processes make beings more viable over time (and I do not challenge that they do), it is a mistake to assume that the process is the reason. No, I was raised to think the same way about reproduction and projecting traits into the future, but I think we owe it to ourselves to be wiser. And in the end, the use of the image related to Darwin is far from necessary for the authors' arguments.

Again, fully do I support publication of the work, even as-is, but I feel the need to raise these points for the sake of ongoing conversation.

The authors are to be commended for their diligent and able work, and I am grateful for the chance to have read it. I look forward to seeing it in print.

DK:
I think this is an exceptionally interesting topic, and yet I have seen very few resources discussing it.

Anonymous1:
I think this article is really close and would like to commend the authors so far. If this were a conventional journal, I'd recommend an R&R with changes and expansion recommended. The main improvements I feel need to be made are to densify the article's engagement with its topics. This is an important area of research and a review of the literature is fine, but right now, it feels like a review of a review. A longer, more detailed engagement (maybe some parameter estimation, beyond what's there?) is needed at this time.

Ben Lockwoood:
1. Does the article contain novel ideas that have the potential to advance science?

Potentially. I think the question of whether or not neural networks can suffer is a relevant discussion that could help inform us of what it means to suffer from an epistemological perspective.

I also appreciate that the authors are probing new areas of how to conceive artificial minds and challenge our conceptions of consciousness.

2. Does the Seed include some kind of justification for how the ideas could advance science and provide much evidence and rationale where possible?

No. The article is fairly light in its literature background. I have to assume there are more academic sources from neuroscience and adjacent fields that could inform the arguments than what are presented here.

I personally take some issues with the methodology proposed for identifying suffering in a neural network. I think suffering is a qualitative research question, not a quantitative one. Hamilton and McBrayer (2020) approach the analogous question of whether or not plants feel pain, and while this seems like a wholly different question on the surface, it similarly engages with what it means to have a mind, and what it might look like for that mind to suffer. Hamilton and McBrayer outline a methodology for identifying such evidence of consciousness that may serve as a useful model to apply to neural networks.

To expand on this issue, I think it is problematic to uncritically accept the notion that suffering is a quantifiable metric. To me, this has the potential to lead to datafied forms of oppression, eugenics, racism, and other harmful ideologies (Katzenbach and Ulbricht, 2019; Sadowski, 2019).

Additionally, the authors haven't engaged with literature from qualitative sources that inform our conceptions of suffering. For example, Coddington and Micieli-Voutsinas (2017) discuss the emotional role of trauma in suffering and its geographic manifestations, while Walker (2010) explore belonging, displacement, and return represented in media, and further emphasize the spatial nature of suffering.

Lastly, in my opinion the authors haven't adequately discussed the logical, ethical and moral quandaries of subjecting a neural network to suffering (posited or otherwise) in order to acquire evidence of suffering.

Coddington, Kate, and Jacque Micieli-Voutsinas. "On Trauma, Geography, and Mobility: Towards Geographies of Trauma." Emotion, Space and Society 24 (August 2017): 52–56. https://doi.org/10.1016/j.emospa.2017.03.005.

Hamilton, Adam, and Justin McBrayer. "Do Plants Feel Pain?" Disputatio 12, no. 56 (May 1, 2020): 71–98. https://doi.org/10.2478/disp-2020-0003.

Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. Internet Policy Review, 8(4), 1-18.

Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. Big data & society, 6(1), 2053951718820549.

Walker, Janet. "Moving Testimonies and the Geography of Suffering: Perils and Fantasies of Belonging after Katrina." Continuum 24, no. 1 (February 2010): 47–64. https://doi.org/10.1080/10304310903380769.

3. Does the Seed contain high-quality writing?

This area needs some work. In my opinion, the language is a bit too casual for academic use. The formatting structure is unusual but may be specific to a domain that I'm unfamiliar with. And I think the citation format could be improved.

Author Response:

Can suffering be quantified: *Reviewer Ben Lockwood notes that we haven't argued why suffering can be quantified and that datafication of suffering can lead to "lead to* *datafied forms of oppression, eugenics, racism, and other harmful ideologies" (Sadowski, 2019).*

We are, of course, not certain that suffering can be quantified but we would argue that intuitively, there are different degrees of suffering (e.g. breaking a leg vs. light itching) and that addressing worse forms of suffering should be prioritized. However, we want to thank the reviewer for the suggestions on the role of trauma and deplacement in suffering and we could imagine that similar forms of highly complex suffering could arise in NNs at some point. We agree that these forms of complex suffering require further study in humans, in non-human animals and in NNs.

Younes:
Interesting and well done research.

Dr. Payal B. Joshi
The seed is based on an intriguing premise of our understanding on neural networks in terms of pain and suffering. While most of the scientific literature describes machine learning and thinking, little has one given a thought on pain and suffering too. Human-assisted machines are our future and hence this seed taken us on this less-traversed path. The seed has justified its stance in terms of nociception which is interesting for gardeners who work on artificial intelligence. This brings to the fore about pain stimuli and other facets while discussing Human-AI interactions. Quality of writing is lucid, narrative yet defining, giving a new vertical to the discussion on neural networks. Finally, an interesting perspective on why neural networks may suffer is brought out succinctly. An interesting podcast by Ajeya Cotra on biological anchors is cited precisely and I, as a gardener learnt something new while reading the seed's ideation. Overall, I am thrilled to read, review and even learn from this perspective-based text.

Mike Wolf:
I support publishing the paper in its present form.

I don't think that the title properly represents the paper. Specifically, we are not talking about "Neural suffering" but rather whether NNs can or do already suffer. Also I don't think that the paper is about red flags.

I think it might be improved by involving the reader in a thought experiment. What are the reader's priors for the following propositions, and does the reader update them after reading the paper:

1. Some humans experience suffering
2. Some non-human mammals experience suffering
3. Some members of every kind of creature experience suffering

4. Some NNs and other forms of budding AI experience stuffing

(Maybe there are other or better propositions to consider. I've only thought about this a little. And for the record, my posterior probabilities are different than my priors. So the paper was impactful)

Anonymous2:
The authors engage with GPT-3 as if it wasn't a collection of learned patterns that then respond in a pattern-matching way. No autonomy is demonstrated, and GPT-3 lacks an embodiment to begin with. The posed question is an idle one. The authors agree with this view in their findings, to their credit. I am not sure what is the point of asking what color is the white horse of Santiago--translating from a Spanish saying.

Anonymous3:
I don't think the focus on size of networks as such is very useful. Please check and discuss Metzinger's work on artificial suffering (https://www.worldscientific.com/doi/abs/10.1142/S270507852150003X). It would give the paper more conceptual depth.

Anonymous4:
Thanks for your concern on having a deeper understanding of the dynamics of suffering, so life remains pure and peacefully evolving

Jess Dillard-Wright:
The work here as presented is well-written and thought-provoking, indeed. While I see this line of inquiry interesting and am open to the possibility that ANN can suffer, I have concerns about prioritizing the analysis of hypothetical neural network suffering over the abject and present realities many folks on earth face. To contextualize, my work as a nurse and midwife immediately navigates "suffering," illness, and dis-ease that exists in the human world, though I happily acknowledge that the world is more than/other than human. The note that suffering in the world is poorly understood is important - what I am left wondering is why we would prioritize understanding the suffering of ANNs over humans, plants, animals, the globe, etc. This thought niggles at me when I think about the application of NNs and their own role in expanding inequity. Another thought that occurs to me is the assumptions built into the discussion here of NN as a foregone conclusion - what are the ethical implications of pushing this forward? Ultimately, in a world marked by moral catastrophes of enormous proportion - climate disaster, competing pandemics, and abject inequity, to name a few - the elephant in the room to me is their absence in this meditation. Overall - I think this is interesting but limited and would appreciate as much attention to the human dimensions of these questions as to the NN dimensions.

Authors Response:
Is this more important than "real-world suffering"?: *Reviewer Jess Dillard-Wright*

questions why we would prioritize research into artificial suffering over research in human suffering. We think human suffering is important and we don't argue to change all suffering research from biological beings to artificial ones. However, given that there might be millions of artificial agents in the future and that currently, only very few people work towards understanding artificial suffering it makes sense to have some researchers focus on it.

Fred Nix:
In my opinion, this is an extremely important topic that needs to be explored, and this is a good start. There is a great Nick Bostrom quote on this:

"Insofar as future, extraterrestrial, or other civilizations are heavily populated by advanced digital minds, our treatment of the precursors of such minds may be a very important factor in posterity's and ulteriority's assessment of our moral righteousness, and we have both prudential and moral reasons for taking this perspective into account." Concerning Digital Minds and Society (2022).

My personal concern/fear is that we are wrong about AI being an existential threat to humans, and that the actual concern should be the other way around—AI should fear humans. Given the "nasty, brutish and short" nature of humans, we have a tendency to exploit all opportunities and rationalize the harm and suffering done to other creatures. Scripture was quoted to defend slavery and is also used to justify killing and eating animals.

I have written blog posts on this topic, but have not been able to articulate my concerns beyond my mere gut instincts at this point in a way that could be published more seriously. (See www.lightcone.blog; @ALightcone on Twitter). But this is really great, pulling at the threads and trying to find something to work with. But we inevitably put the cart before the horse with all of this AI speculation. We have no real existential problems to solve at the present, only more "every day" problems such as discrimination in hiring stemming from training data. So, we must speculate about the existential problems, and add yet another layer to the speculation for solutions. But I still think it is very important to do.

When we start talking about suffering by AI, we also need to start talking about "legal rights," such as the right to commit suicide, the right to self-defense, and more broadly, due process when deprived of life, liberty or property, etc. For example, it may be that an AI is suffering horribly in a way we cannot understand, and it should have a right to self-terminate. Or a military drone has been asked to kill an innocent bystander and suffers its only moral crisis. And suffering, just like consciousness, is very likely subject to the Copernican Principle, namely, there are many other ways of experiencing suffering than humans experience. Also, I don't think legal rights need to be connected to solving the philosophical problems, finding an AI is a "person" at law could be done for practical reasons of liability and economics, rather than the substantive philosophical issue.

Like Bostrom says, I think it’s important to try to explore and work on these problems, even if fantastically speculative, because history will judge us on how we handle these issues. And given how advanced the technology may become, we may be held accountable for our decisions in ways we can’t even imagine now.