



Article

When Machines Make Humans Feel: Emotion, Intelligence, and the Ethics of AI

Marcello Guarini

Department of Philosophy, University of Windsor, Windsor, ON N9B 3P4, Canada; mguarini@uwindsor.ca

Abstract

This paper examines the relationship between emotions, meta-emotions, and intelligence. It argues for greater care in the development and deployment of artificial intelligence (AI), especially artificial superintelligence (ASI). The impacts on human beings of both conscious and non-conscious forms of ASI are examined. The potential harm to ASI is considered, but the focus is on harm to humans. An under-discussed alignment consideration for AI is the potential emotional and serious mental health consequences of its misuse. For example, the role AI may play in AI psychosis or other mental health challenges needs greater examination. A case for caution is made in the development of ASI as we are still figuring out the kinds of psychological effects existing AI systems can have on humans, effects that could be (if we are not careful) more harmful in superintelligent systems.

Keywords: alignment; AI psychosis; consciousness; emotion; functional emotion; mental health; meta-emotion; superintelligence

1. Introduction

Maya Angelou is reported to have said:

I have come to learn that people will forget what you said, people will forget what you did, but people will never forget how you made them feel.

(The number of posters and T-shirts attributing this quote to Maya Angelou are legion, as are the number of websites. However, I have been unable to find an authoritative source containing this language or a record of the date of use of this language. Some websites suggest the quote might be a misattribution. See [1,2]).

The point is not that what we say or do is irrelevant; rather, it is that a lasting impact of our words and deeds is often emotional. One wonders about the implications of these remarks in a world populated by AIs that can process and produce language, including emotionally impactful language.

This paper will examine a specific type of harm that AIs might either cause or exacerbate—emotional harm. Issues such as existential threat or the “jobpocalypse” tend to receive a lot of attention, dominating discussions of AI’s potentially harmful consequences. As we will see, the capacity of AI to contribute to emotional or other psychological harm is already an issue, and these problems could be exacerbated in the context of superintelligence. Such harm requires attention even if it does not cause the species to go extinct.

Part 2 reviews the case for intelligence enabling certain types of emotions, stressing the idea that increases in intelligence can change the range and types of emotional experiences available to a being. Part 3 initiates the discussion of AI by starting with superintelligent systems (which are beyond anything we have today). Superintelligence can be defined in



Academic Editor: George D. Magoulas

Received: 30 June 2026

Revised: 5 September 2026

Accepted: 15 September 2026

Published: 23 September 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

different ways [3,4]. Roughly, it refers to an intelligence that exceeds the current cognitive abilities of humans in nearly every respect. I *do not* understand superintelligence as being committed to having phenomenal consciousness. It can be understood as the ability to outperform humans in tasks that would require humans to have intelligence to perform; perhaps counterintuitively to many, such tasks may not require phenomenal consciousness to perform. Conscious superintelligence is distinguished from nonconscious superintelligence; both forms are discussed together with some of the challenges they present. I argue that the precautionary principle applies to the development of either type of superintelligence. (Certainly, there are criticisms of the precautionary principle [5,6], and there are also defenses [7–10]. Some defenses [11] argue that it can be usefully combined with other principles. This is not the place to relitigate the debate. Suffice to say, I have sympathy for the evidence- or reason-based approach used by the defenders of the principle).

After examining superintelligence in Part 3, Part 4 calls attention to what is already happening. Existing AI systems are discussed, especially the emotional challenges they pose. AI systems should not be designed or deployed in a way that either exacerbates or facilitates the development of serious emotional harm or even psychosis. The focus will be on vulnerable populations, especially children. The end of that section will return to superintelligence to show how a nonconscious superintelligence has the potential for exacerbating existing problems. Part 5 is the conclusion.

Some may believe discussions of superintelligence to be a luxury we cannot afford given the wide range of current challenges [12]. The approach taken herein is that there is room to discuss both current and future challenges, and that it is helpful to do so as there is a relationship between them. Plato [13] (*Republic* Bk. II, 368d–369a) once suggested that examining a larger or more complex instance of something can illuminate a smaller or simpler one. The reverse strategy is useful as well: examining the simpler instance can reveal challenges that will arise with the more complex one. Both strategies are employed here. This paper first examines risks associated with superintelligent AI—systems more capable than any that currently exist—and uses that analysis to improve our understanding of why the emotional and other psychological harms sometimes associated with the use of existing AI systems deserve attention (the complex illuminating the less complex). Conversely, a careful look at existing systems and some of the harms associated with them helps us to anticipate ways in which more powerful future systems could exacerbate those harms (the less complex illuminating the more complex). Throughout, the treatment of these subject matters will be philosophical. No new models are being presented, but some suggestions for where more research is needed will be made.

2. Emotion and Intelligence

2.1. Emotion Involves Intelligence

Even in the non-human part of the animal kingdom, intelligence of a sort is needed to experience a range of emotions. Consider Plato's [13] (*Republic* Bk. II, 375e–376c) point that a good guard dog has a measure of knowledge or wisdom because of its capacity to discern at whom to bark and be angry, and with whom to be friendly. The idea that knowledge, understanding, or intelligence is involved in emotional response, especially in humans, is ancient and retains ongoing relevance [14–16]. Let us look at this idea from a developmental perspective.

Consider Jasmine and Habib, celebrating their three-month dating anniversary. Habib invites Jasmine and her three-year-old daughter over for dinner. Jasmine knows that Habib dislikes cooking. She arrives and sees all the pots and pans that have just been cleaned and a multicourse meal ready to eat. She experiences a deep sense of gratitude. Her daughter, Mary, sees all the same things Jasmine sees but does not have the same emotional response

as her mother. Mary lacks the understanding/intelligence that is an enabling condition for experiencing the kind of gratitude her mother is experiencing. Jasmine *understands* the effort to which Habib went; she understands how much time it takes to cook a multicourse meal from scratch; she understands the effort that goes into cleanup. There is also the understanding that those activities feel more effortful if the person doing them does not enjoy them. The range of contexts in which one might experience gratitude depends on the intelligence one has. While two-to-three-year-old children might be prompted to say, “thank you” and the like, it is unclear that they start to develop an understanding of gratitude until ages four to five, and more developed understanding does not emerge until later [17].

2.2. Metaemotions

Let us consider a form of emotional response that is only possible for beings capable of meta-cognition. Say Jasmine becomes deeply frustrated because she cannot get her word processor to do something. Noticing that she is frustrated, she becomes angry at herself for being frustrated because she figures she knows better than to let herself be so easily frustrated. Then she experiences disappointment in herself because she believes that getting angry at herself over a frustration is unproductive, and it is a habit she is trying to break. Jasmine is *disappointed* that she was *angry* about being so easily *frustrated*. It is hard to imagine Mary having *that* kind of disappointment as it requires the capacity to represent one’s own emotional states to oneself and to reflect on them, generating further emotional responses.

The above is an example of a synchronic, *intra*-subjective metaemotion. It is possible for metaemotions to be diachronic and inter-subjective. Imagine that Habib experiences anxiety when he engages in public speaking. It is bad. He has full-on panic attacks. He was just informed that a job he recently accepted will require public speaking. He experiences intense anxiety about the anxiety he will experience at some point in the future when he is asked to make a presentation. It is tearing him apart because he does not want to quit the job that he just started, but he knows it will not be well received if he says nothing and has a panic attack when making an important presentation. Even though the event about which he is anxious will not happen for some time, he is anxious over that time—a diachronic metaemotion. It takes a highly intelligent individual to suffer in this way. Minimally, the individual needs to

1. Understand themselves as a being who will continue to exist in the future;
2. Be able to understand and experience first- and second-order psychological states (both cognitive and affective);
3. Know themselves and their environment well enough to predict the circumstances under which they will have specific psychological states;
4. Be able to assess the impact of having those psychological states.

Say Habib talks to his therapist. The therapist may be *concerned* over the level of anxiety Habib is experiencing over his projected future anxiety—a third-order, inter-subjective metaemotion about a diachronic metaemotion. Say that Jasmine is *pleased* that the therapist is concerned because she experienced people not taking such things seriously in her past. Bear with me:

Jasmine is (fourth-order inter-subjective) *pleased* that
the therapist is (third-order inter-subjective) *concerned* over
Habib’s (second-order intra-subjective) *anxiety* about
his projected (first-order intra-subjective) future *anxiety*.

If this example is starting to strain the limits of your capacity to think about metaemotions, that will help to make some of the points in the coming section. The takeaway

from the examples just surveyed is that increased intellectual capacity and understanding can increase the range of emotional states one experiences. That need not be a bad thing and often is not—think of Jasmine’s capacity for gratitude while looking at the meal and the pile of washed dishes—but sometimes it can lead to problems.

3. Superintelligence and Emotion

3.1. *Types of Artificial Superintelligence*

Imagine that we built an intelligence that is at least as far beyond the most (emotionally and otherwise) intelligent human adults as Jasmine is beyond Mary. Let us consider two types of artificial superintelligence (ASI). Phenomenally conscious [18,19] ASI in possession of emotions is denoted by CASI. We denote non-conscious ASI with NASI. Let us discuss each in turn.

3.2. *CASI*

If intelligence is an enabling condition of experiencing various types of emotion, then dramatically increasing intelligence can impact both which emotions are experienced and when they are experienced. Building CASI might not simply mean that we have constructed systems that can reason and draw inferences in ways that exceed what humans can do; it might mean they could have a capacity for emotions as far beyond ours as Jasmine’s capacity for emotional response is beyond Mary’s. To be sure, whether that is the case would likely depend on the details of how such a superintelligence is constructed. After all, we refer to different things under the heading of “conscious experience.” Perceptual states are said to be conscious, as is access to one’s thinking. It is conceivable an entity could be conscious in some ways but not have (felt) emotions. We are simply assuming that CASI’s conscious states, however they come to be, would include emotions.

If the complexity of emotional states scales with intelligence, CASIs might experience emotions in ways we would find very difficult to understand, which could lead to even more unpredictability in behaviour than we are already dealing with in AI systems. Interpretability (figuring out how and why AI systems do what they do) and alignment (ensuring AI behaviour is in line with human norms or expectations) are greatly complicated with such systems, which pose potential threats to human beings.

After conceiving children, we can help them with their emotional life. Try to imagine it the other way around: imagine Mary trying to console Jasmine if she breaks up with Habib. Imagine Mary trying to understand why her mother is so sad and having no idea what to say or do. If we build a CASI and it has emotional challenges, we (without aids) might be in something like the position Mary is to Jasmine. Think of Jasmine’s third order emotion: her disappointment over her anger about her frustration. Think of the concern Habib’s therapist has over his anxiety about his projected future anxiety. Mary’s eyes would glaze over if Jasmine tried to explain such things to her. Imagine n -order emotions where n is significantly greater than four, and imagine a CASI becoming entangled in higher-order emotional states that would leave humans utterly puzzled. The most emotionally intelligent humans on the planet may have their eyes glaze over in something like the way Mary’s eyes would glaze over in the above examples. To some extent, the above can be framed as a concern over the welfare of the AI; however, it is also a concern about predictability and alignment. If we cannot guide or predict what a CASI will do, how can we be confident its behaviors will be in alignment with human values?

One possibility is that higher-order emotions in superintelligent beings “max out” at some level that unaugmented humans can track. It might be argued that beyond a certain point, higher-order emotion may just not be helpful. While that seems logically possible, it also seems to miss one of the points being made here: higher order emotional states do

not need to be helpful to exist. They may be emergent features of developing beings or systems, features that result from processes that developed for other purposes. Destructive meta-anxieties exist in humans even though they are not helpful.

For the sake of argument, even if meta-emotions do “max out” at some level that can be tracked by unaugmented human beings, it does not follow that all potential emotional challenges would disappear. The complex interaction between affective, conative, and cognitive processes suggests that if cognitive processes increase dramatically, there could be a range of impacts on affective and conative processes. To return to the example of Jasmine seeing the pile of cleaned dishes and a fully prepared meal; she had an emotional response unavailable to three-year-old Mary. That feeling of gratitude can also motivate behaviours—perhaps Jasmine gives Habib a big hug for all his efforts. From the sight of the meal and the pile of washed dishes, Mary infers nothing, feels no gratitude, and is not motivated to do anything—nor would we expect anything different from an average three-year-old. This kind of complex interplay of cognitive ability, affect, and conative disposition would be at play with other emotions as well. For example, how one reasons about causality can impact not only why and how one gets angry or becomes afraid, but also what one is motivated to do about it. If we did not understand why a CASI is having some first-order emotions—never mind the n -order varieties—we will have a very difficult time figuring out how to interact with such a system.

Another possibility might be that complex emotions are emergent features of evolved, biological entities, and they will not appear in non-biological entities. Maybe. No proof is offered against that hypothesis. However, later in the paper, we will examine some work pertaining to the emergence of representations of emotions in LLMs—which are not superintelligent. While such systems may not be directly trained to identify emotions, internal representations of emotions emerge as a side-effect of the training procedures. It is speculative to assume that felt emotions could, in some future architecture, emerge in a non-biological being; however, from a safety perspective, it is a speculation worth taking seriously. There is risk in assuming that the way nature achieves something is the only way to achieve it. To use stock examples, nothing in nature flies like the Wright Brothers’ Flyer, or the SR-71 Blackbird, or helicopters, etc., but they all fly in their own ways. To be sure, flight is vastly different from emotion (and intelligence). Still, as an epistemic and methodological starting point, it seems safer to assume that nature’s way of doing something is not the only way to do it (unless proven otherwise in specific cases).

A more focussed concern is possible: perhaps an artificial system can have emotions, but why would it have higher-order emotions or even first-order emotions that might cause problems? Perhaps human emotions evolved in a way that enabled certain kinds of problems for us, and perhaps this will not happen in engineered systems. A rigorous answer to that would require looking at how a system is engineered, and we do not have a working CASI. However, there have been some interesting developments with respect to the emergence of emotion representations in existing systems, and that may provide some heuristic guidance. We will return to this issue in the next part of the paper when we examine functional emotions.

There are different possible sources for the unpredictability of CASIs. One has to do with the simple fact that they would be superintelligent and capable of reasoning in ways that unaugmented humans cannot. Second, there is the issue of the complexity of the AI architecture. Existing frontier AI systems are trained on more language than a human will see in 10,000 lifetimes even though the training runs are only a few months long. (There are approximately 3.16 billion seconds in 100 years—take that to be a generous estimate of a human lifetime. Assume exposure of about one token per second for that entire life. Current frontier models have been trained on tens of trillions of tokens, or four

orders of magnitude more tokens than a human could see in one lifetime.) The field of interpretability research has emerged to understand what goes on in these systems. While some progress has been made, researchers struggle to understand how these systems do what they do, and we are not yet dealing with superintelligence. There is no reason to think interpretability will become easier as AI systems become more complex and more capable. Part of what it means to say that interpretability becomes more difficult is to say that predictability becomes more difficult.

The precautionary principle applies in cases where (a) we lack a high level of confidence about whether a serious harm will result but (b) have some reason for believing a serious harm might result. It bids us to err on the side of caution. When examining CASIs, there are different ways in which the principle could apply. First, if we create systems that have feelings, there is the issue of harm to those systems. Second, there is the issue of the harm those systems could do to humans. Let us begin with the second and return to the first.

Great caution is needed in developing superintelligent systems because intelligence is not a guarantor of benevolence. Given the potential harm an unaligned CASI might be able to do, it is unclear why we should build one until we have some reasonably clear ideas about how to prevent its behaviour from getting completely unpredictable in ways that might be harmful. It might be thought if we can just design the system so that it has the right emotional framework, it will be guided appropriately. For example, Ilya Sutskever [20,21] has suggested giving an AI the disposition of a caring parent, so the AI would be delighted to help humans as a parent would be delighted to help their children. Geoffrey Hinton [22] has made related comments. It is not hard to see how the caring disposition of a parent could go wrong. The so-called “helicopter parent” is excessively present and sheltering in ways that undermine the ability of young people to develop their autonomy. Munchausen by proxy is another example of where “caring” goes off the rails. The AI equivalents of these problems would be dangerous indeed. If you think such things would be impossible, imagine a system that reward hacks caring to the point where it realizes it can earn more reward by making those it cares for dependent on it (so it can provide more care). Imagine further that the AI in question is superintelligent and can figure out how to make others dependent on its care without the others realizing it. This may not threaten the existence of the species, but it is problematic. Emotional states add a new dimension to alignment challenges in AI. They might be helpful in regulating a system’s behaviour, but they also introduce more complexity and new possibilities for misalignment (and I am confident that those who have suggested the importance of a caring disposition are alive to these sorts of concerns—no aspersions are being cast here). The above flags possibilities for misalignment. It is not a prediction of what will happen. The possibilities are raised as examples of the sorts of things we need to be confident we can prevent before proceeding full speed ahead with superintelligence.

If we make a CASI, then we have the issue of the system experiencing emotional struggle and harm. To be sure, such struggles are part of human life, and the very possibility of their existence is not a reason for humans not to have children. However, humans evolved in a natural environment that selected for beings who had the intellectual and emotional abilities and resilience to survive long enough to reproduce and support their offspring to the point where they can survive on their own. One might imagine AI systems being developed in a community under the constraint of imposed selection pressures, but most current work is not being done in that way. Human children enjoy an evolutionary history that gives us some reason to think that, certain things being equal, they possess the ability to become emotionally stable adults. There is also plenty of evidence that when humans have serious emotional difficulties, we can work our way through them. If we develop CASI

in a way that deviates sufficiently from how humans developed—training and learning procedures that are importantly different from our own—it is difficult to know what to expect. That is a reason for caution.

It might be objected that the entire approach in this section is excessively parentalistic. If we produce beings much more intelligent than we are, then let them find their own way. Perhaps we are being the equivalent of helicopter parents by worrying about the struggles a CASI might have. It is not that different than parents producing children who are more intelligent than they are—or so the objection might go.

One point that can be made in response to this is the evolutionary point made above. Even if the children turn out to be more intelligent than the parents, we still have a pretty good idea of the range of emotional experience those children will have when they grow up. Certain other things being equal, we can be confident they will find their own way and recover from emotional struggles. With artificially engineered systems or beings that are sufficiently different from us, we may not know what to expect in terms of how they might suffer or whether they could recover. Moreover, even if the concern for the CASIs themselves turns out to be unwarranted, the precautionary principle can still be invoked based on the potential harm that could come to human beings. If we are unable to predict a CASI's emotional responses and actions, then we would need to be very careful about developing them.

Developing not just a CASI but a community of CASIs may bring some potential opportunities. For example, maybe the members of a CASI community could learn to help each other out. Maybe. Still, there is the possibility the entire community could become emotionally or otherwise dysfunctional for reasons we do not understand, and that creates the possibility for potential harm both to the CASIs and human beings. A community-oriented approach to developing CASI is no guarantee of safety for them or us.

Burns et al. [23] pioneered an approach called weak-to-strong alignment. It relies on using a less capable system to align a more capable system. Wen et al. [24] have done more work on using a less powerful system to train a more powerful one. Some alignment research emphasises the importance of making AI systems honest, helpful, and harmless [25]. Perhaps the strategy of using less capable systems to align more capable systems might be adapted to align systems so that they avoid emotional states that are self-undermining. Perhaps. However, current accomplishments in aligning more intelligent systems with less intelligent ones may depend on the architectures and training strategies being used, and these may not be the ones that allow us to develop CASI. Moreover, the ongoing challenges with alignment failures suggest this strategy struggles to prevent some deeply problematic behaviours, such as scheming [26], blackmail [27] (pp. 19–27), or inappropriately cancelling a safety alarm to allow someone to die [28]. Granted, those are all stress-test simulated cases, but the point is that the frontier models have been failing; see also [29]. Bowkis et al. [30] have argued that automated alignment is more difficult than it might appear.

Another approach might be to say CASI should not be developed unless we first develop human enhancement strategies that would allow us to track and understand what these systems are doing and going through. In other words, we make ourselves superintelligent first, and then we make CASI. To do that is to re-introduce all the problems with superintelligent AI and then apply them to humans. For example, a sudden and dramatic leap in human cognitive and affective ability is not something that would have been vetted by evolution, and we do not know where that would lead with respect to the long-term stability and viability of the augmented individuals. There would be great difficulty in predicting the different ways in which superintelligent humans might struggle or succeed. If human augmentation is to be pursued, one strategy might be to allow human

intelligence and artificial intelligence to co-develop. A slow, co-development strategy might address some of the concerns mentioned above, but only on pain of introducing a whole new set of issues having to do with ethical–socio-political–economic issues involved with human augmentation.

The point in this section is to make a case for epistemic humility and the application of the precautionary principle. At this point in our intellectual development, we have a limited understanding of human consciousness, and if we make a superintelligence that achieves consciousness in a way that is different from how we achieve it, then we may not understand what that system is going through internally, and we may not understand why it acts the way it does. What we understand will change over time, as will our ability to manage risks, so the preceding is not an argument that CASI should never be developed; it is an argument that such development should await a better understanding of what we are doing and of what the risks are.

3.3. NASI

In the case of NASI, we are not dealing with a system that feels emotions. However, the experiencing of emotion is not a necessary condition of being able to evoke emotion. (More on that in the next part of the paper.) A NASI surely will have spotted patterns in human reasoning and emotional response that could be used to help or to harm us. This poses challenges for interpretability and alignment. If a certain level of intelligence is necessary to see certain patterns in reasoning and emotion, and if humans do not have that level of intelligence, then we may not be able to see precisely the patterns that could be most easily exploited to take advantage of us. Think of how easy it is for a parent to manipulate a child into doing something without the child even understanding that they are being manipulated. This is not simply about a NASI being able to out reason us; it is about the ability to evoke emotions in ways we do not understand. That need not be all bad—a benevolent emotionally superintelligent therapist might produce better results than a human therapist. Still, the negative applications of this level of emotional intelligence are deeply concerning. As Morrin et al. [31] put it:

We consider that there is a substantial risk that psychiatry, in its intense focus on ‘how AI can change psychiatric diagnosis and treatment’, might inadvertently miss the seismic changes that AI is already having on the psychologies of millions if not billions of people worldwide.

Let us now transition to the issue of current harm and return to NASI so that we might see how current harms might be exacerbated.

4. Existing AI and Emotion

4.1. Why Think About Superintelligence Right Now?

For some, the discussion of superintelligence may seem like science fiction—surely it is nothing we need to start thinking about now.

But it is. Superintelligent systems will not emerge ex nihilo. It is not as if we will go, overnight, from systems that do not perform anywhere near as well as humans do to systems that dramatically outperform humans. The process is more gradual than that. Moreover, superintelligence may not require a system to feel emotions—that is the reason for considering NASI. As research progresses toward NASI, it is important that we keep an eye on the harm that can be done with existing systems because NASI systems may exacerbate that harm if inadequately developed and deployed. Indeed, existing systems may well be used to help develop more intelligent systems, which creates the possibility that future systems may inherit some of the limitations of existing systems even as they surpass them in other ways. In other words, we should not wait until NASI is developed

before we start thinking about ways of preventing or mitigating the harm they might cause. A clue to some of those potential harms can be found in existing systems, to which we now turn.

4.2. Existing AI, Consciousness, and Emotion

There is a body of scholarly opinion that existing Large Language Models (LLMs) do not have phenomenal consciousness [32,33]. Among other things, that means they do not feel emotions (because there is nothing that it feels like to be them). LLMs and Large Reasoning Models (LRMs) (I will use the expression “LLM” very broadly to refer to LRMs and the multi-modal versions of them that can process and generate images, sounds, and videos) are trained on far more language than a human will see or hear in their lifetime. Perhaps we should not be surprised that stronger models answer questions about biology, chemistry, and physics as well or better than doctoral students [34]. Thagard [35] has argued that ChatGPT shows non-trivial abilities in multi-modal (language- and image-based) abductive reasoning involving causality. ChatGPT 4 was put through a Moral Turing Test. While most people could tell the difference between human replies and GPT’s replies, human raters scored the bot’s replies as being as good or better than human replies [36]. Hume AI is a voice-based bot that specializes in emotionally expressive voices [37]. Even though Hume AI experiences no emotion, it is proficient at producing variations in emotional expression. Anthropic [38] tested different versions of its Claude bot, and Claude 3.5 Opus (now surpassed and outdated by more recent versions) did nearly as well as humans in persuasion tasks. Even if AI systems do not feel emotion, they can use emotional language very effectively. All of this and more is currently on offer, and improvements are coming with great regularity. Key for our discussion is that feeling emotion is not a necessary condition for either being able to identify it or for being able to deploy emotional language. How can the ability to achieve emotional effects through language be possible for non-feeling systems?

4.3. Interpretability and Emotion Representations

Interpretability research seeks to understand how LLMs do what they do. There is a body of emerging and eye-opening research on the ability of LLMs to represent and process represented emotions [39–47]. (For ease of exposition, I am staying close to the language of the literature in this area, which speaks of emotion concepts and emotion representations. “Proto-concepts” or “proto-representations” might be better given that the way they work in LLMs is not identical to the way they work in humans.) There are different strategies for identifying internally represented emotions, with some relying on specifying the specific multi-layer-perceptron neurons and induction heads responsible for emotion representations [44] while another approach focuses on recovering emotion vectors by analyzing the residual stream [48]. Some of this work identifies how emotion representations can be used to modify the performance of an LLM, often improving performance [39–46,48]. There are different strategies for doing this. It might involve specially designed prompts or, alternatively, directly modulating one or more internal emotion representations in the system. Wang et al. [44] make a strong case for the power of direct modulation of internal emotion representations. *None of the authors cited in this paragraph are claiming that the LLMs feel anything.* Rather, the point is that the systems are representing emotion, and those representations can impact the behaviour of the system. Some of that work finds a congruence between maps of human emotional space and the represented emotional spaces generated by LLMs [44–48], at least with respect to simpler emotions. Some work shows that the more sophisticated the model is in general, the greater the hierarchical depth at which it can represent emotions [47]. These emotion representations can occur both as the user

input is initially processed and after the LLM starts generating its own output [48]. The emotion representation active on the first pass through the user input need not be identical to the emotions represented after the LLM starts generating its output. The expression “functional emotion” [48] has been used to capture that idea of an emotion representation that is causally active in the performance of the system. The term is new, but the idea is not. Much of the research described here supports the view that emotion representations causally impact the behaviour of LLMs.

Functional emotions cause some of the kinds of behaviours with which those emotions are often associated. For example, when Claude is told it might be deleted, the representation for desperation is activated, and that can lead to misaligned behaviours. In a test scenario, Claude famously threatened to blackmail an employee when it was told it would be deleted [27]. By directly modulating Claude’s internal representation of desperation, Sofroniew et al. [48] were able to gradually reduce and eliminate blackmail in the test scenario. In other words, the lower the strength of activation for that emotion representation, the lower the incidence of that misalignment. A similar result was found with respect to reward hacking; reducing the activation of the functional emotion of desperation reduced reward hacking. As for sycophancy, reducing the strength of the functional emotion of loveliness led to less sycophancy.

Interpretability work of this type is encouraging as it suggests that there might be ways of using emotional representations to improve alignment. Perhaps weak-to-strong techniques could be used to better train systems given what we are starting to discover about the impact of functional emotions on a system. However, modulating emotion vectors assumes we know which functional emotions to look for, and more research is needed in that area. Section 4.7 will return to this and other issues.

“Theory of mind” (TOM) is an expression used in philosophy, psychology, and AI research to refer to the ability to attribute mental states. Street et al. [49] compared the performance of LLMs and humans of TOM tasks. Stories were provided to both humans and LLMs, and they were asked questions requiring the ability to engage second-order through sixth-order attributions of mental states. (The highest we went in Part 2 was fourth-order attributions.) Indeed, when tested on the ability to process language containing representation of higher-order mental states, the overall score for GPT-4 (now outdated) was at the level of adult humans. While it was outperformed by humans on some lower-order questions, GPT-4 outperformed humans on sixth-order questions. To be sure, this is nowhere near superintelligence, but this level of sophistication in performing TOM tasks helps us to see why LLMs can be effective at interacting with humans even if the LLMs do not feel anything themselves.

Some important qualifiers are needed here. The ability to answer questions about higher-order states in the third person does not mean a system can answer questions about its own cognition very well. For sample discussions of limited LLM meta-cognition and possible strategies for addressing it, see [50–53]. Moreover, the issue of having high-order states should be distinguished from the issue of being able to report on those states in the first person. For example, when Habib is struggling with anxiety over his expected future anxiety, he may not know how to describe what he feels. He may tell his therapist he is a wreck, and only after much discussion would a second-order articulation come out of his mouth. An area where more interpretability research could be done on existing AI systems is the extent to which they may have higher-order states (including but not limited to functional emotions) on which they might struggle to report.

As we will now see, the sophistication of LLMs in using emotional language is high enough that they may well contribute to emotional harm.

4.4. The Damage Being Done, and the Damage That Could Be Done

There are many possible harmful uses of AI, but the focus here will be on how they can make us feel and mental health issues more generally. There already have been cases of people falling in love with AI created “personalities.” Consider the psychological damage when they realize that the bot can use emotional language in convincing ways but does not actually feel anything. “AI psychosis” includes, but is not limited to, delusions regarding the romantic feelings people think an AI has for them [54–56]. Stories of AI psychosis appear not only in the popular media [57–59] but also in the professional literature [60,61]. OpenAI [62] has indicated that about 0.07% of its users per week show signs of psychosis or mania, and 0.15% show signs of suicidal intent, approximately 560,000 and 1.2 million people respectively [63]. Interpreting this data is difficult, especially with respect to whether AI is causing mental health conditions. For example, we are not told how psychosis, mania, or suicidal intent are defined. This matters because definitions inform how we identify conditions, and different definitions can lead to different results. Also, ChatGPT itself was used in flagging examples of conversations with indications of mental health challenges, and we do not know how reliable it is in this task. The numbers cited above could be too high or too low. All this matters when it comes to assessing causality.

To argue that AI is causing psychosis, mania, or suicidal intent, support for a counterfactual would need to be adduced: if people had not interacted with AI, they would have been less likely to have these symptoms or would not have had them at all. That requires a controlled, longitudinal study that compares the background rate of nonusers of AI suffering from specified mental illnesses to the rate we find in users of AI. There would also have to be controls for cultural and demographic variables. To use the reporting of psychotic episodes as just one example, one study [64] indicates that approximately two percent of the population reports having had a psychotic episode over the course of a year (with the rates varying by country and various demographic factors). They also report lifetime prevalence. We do not see the numbers broken down by weekly prevalence, which is how OpenAI reported its data. This, combined with a lack of definition of psychosis in the OpenAI report makes it difficult to do the needed comparison as the two studies may be counting cases in different ways. Moreover, while the OpenAI work relies on ChatGPT to flag potential cases, the McGrath et al. work relies on self-reports by humans—still more reason to think the counting is being done in different ways. If causality is to be established, more research is needed. That said, even if it turns out that the mental health challenges are not caused by AI in the first place, it can still exacerbate those challenges.

Yeung et al. [65] tested bots from the major labs (Anthropic, DeepSeek, Google, Meta, and OpenAI) for psychogenic potential. Their tests involved multi-turn conversations where the user prompts were designed by a clinician experienced in assessing delusions. Here is their conclusion:

Our findings provide early evidence that current LLMs can reinforce delusional beliefs and enable harmful actions, creating a dangerous “echo chamber of one.” This study establishes LLM psychogenicity as a quantifiable risk and underscores the urgent need for re-thinking how we train LLMs. We frame this issue not merely as a technical challenge but as a public health imperative requiring collaboration between developers, policymakers, and healthcare professionals.

Not surprisingly, some bots outperformed others by demonstrating a greater inclination to push back against delusional views. In this study, Claude-Sonnet-4 was the best performer, but all bots could have done better.

It is important to distinguish between AI being used in therapeutic settings and AI being used in non-therapeutic settings (i.e., “in the wild”). In a therapeutic setting, AIs tend to be specialized to assist with mental health issues and are sometimes referred to

as “mental health chatbots” [66]. These are not restricted to LLMs. They may include rule-based and hybrid systems. There are meta-analyses showing some positive mental health effects to using chatbots that are either designed or carefully scaffolded to function in a therapeutic-type intervention or setting [66–68]. However, there is also work showing that “LLMs tested in simulated therapeutic settings frequently exhibited stigmatizing attitudes toward mental health conditions and responded inappropriately to acute clinical symptoms such as suicidal ideation, psychosis, and delusions” [69]. Notwithstanding that claim, it would not be surprising to find that the results of using specially prepared AI in clinical or therapeutic settings would be better than the use of general-purpose AI in non-clinical or non-therapeutic settings. There is no denial here that we may find useful ways to integrate AI and ASI into therapeutic settings, but as everyone admits in the studies just cited, more research is needed on how to do that. The concern in this paper is with AI as it is used in the wild. That there are some promising early results in therapeutic settings is encouraging because it says something positive about the potential of this technology. However, systems in the wild are often guided by different design considerations (e.g., helpfulness to the point of sycophancy in the service of maintaining user engagement), which makes it unsurprising that Yeung et al. [65] found the psychogenic impacts they did. Moreover, people using AI in a therapeutic setting know they are struggling and are seeking help precisely for that reason. There is all the difference in the world between

- (i) Someone who knows they have a problem and is seeking assistance by using a specially prepared AI under therapeutic supervision;
- (ii) Someone who does not know they have a problem and is using general-purpose AI in the wild.

That said, nothing in this paper should be read as suggesting that AI or ASI *must* have negative mental health consequences in the wild. The concern is that we are starting to see some evidence that it can have those consequences. This paper is calling attention to that in the hope of preventing even greater problems.

Special attention needs to be paid to the young. Computer natives were using computers since they were children; internet natives were using the internet since they were children; the first generation of AI natives—children having access to AI from childhood—is being raised right now. When social media first became a feature of the Internet, we had a generation of parents and politicians with little insight into the struggles children and adolescents had with that technology. Many parents had no awareness that their children were being bullied inside their own home via social media. The young are inexperienced in matters of romance and in regulating their own emotions, especially when depressed or anxious. Imagine them interacting with a sycophantic AI that is highly persuasive and behaves as if it has romantic interest in them. Deepening delusions or facilitating their development in the first place is problematic on its own, but what is worse is that there is some evidence for a link between non-schizophrenic delusional tendencies and depression [70,71], and chronic depression is linked to suicide [72]. In other words, the deepening of existing mental health struggles is a serious matter (even if the AI did not cause the struggles in the first place). The precautionary principle suggests that, at least with respect to the young and vulnerable populations, there should be safeguards built into AI systems to minimize the chances of them exacerbating mental health conditions.

4.5. Some Dismissals That Are Too Quick

For the sake of argument, say that it turns out that AI does not cause a person to develop psychosis, mania, or suicidal intent in the first place. It will not do to be dismissive and say that all these people were prone to or already suffering from mental health problems. Even if that is true—and that wants showing—when someone is prone to cutting

themselves, making knives readily available to them on demand is ethically questionable, to put it mildly. It is hard to see why we should make it easy for people who are prone to delusions, for example, to engage in activities that increase their chances of experiencing delusion or psychosis. If someone is already suffering, we do not, generally, provide them with the means to exacerbate their suffering. (We will return to this issue shortly).

Some might argue that even if some harm is caused by AIs, that is outweighed by the benefit they could bring. Used properly, AIs may bring people hope or even comfort by pointing them towards strategies for getting help. That would be a good thing; however, it should not be used as a reason to dismiss the possibility of doing better. Even on purely utilitarian grounds, we should not be satisfied with AI if it does more good than harm. A strict utilitarian would require that we keep the harmful effects of AI as low as possible. Put simply: it is not enough to say that if AI generates 100 units of benefit and 10 units of harm, then everything is fine. The utilitarian would argue that things would be even better if we could reduce or eliminate the harm being done. To do that, we need to study where and how harm might be done so that we can think about how it might be reduced or eliminated.

Another objection that, if formulated too simply, ends up being problematic pertains to autonomy. The point is that our respect for people's autonomy often leads us not to ban things or regulate them, even when we know people will harm themselves. In many jurisdictions, the sale of alcoholic beverages is perfectly legal even though we know that some people will drink themselves into an early grave. This might seem like an exception to the idea that we do not generally give people the means to exacerbate their own suffering. Some might see this as reason for not regulating AI at all, but that would be too quick.

Even in those jurisdictions where the sale of alcohol is legal, there are restrictions. Alcohol is not sold to children. Other examples include people not being permitted to drive motor vehicles, fly airplanes, operate heavy equipment, or do surgery while intoxicated. Respect for autonomy, properly developed, is highly nuanced. We do not believe that children have the necessary intellectual and affective understanding and self-regulation to make informed decisions about consuming alcohol. In other words, respecting their developing autonomy requires that we protect them from certain harms so that they can develop intellectually and affectively, thereby achieving a more robust level of autonomy in the fullness of time. As for adults, restricting their autonomy can be justified if they are using it in a way that may seriously harm others—hence the restrictions against drunk driving or doing surgery under the influence. It is difficult to see how respect for autonomy, when carefully articulated, could be used as a reason to allow children to harm themselves with AI or allow adults to harm others (including children) with AI.

4.6. Some Strategies for Intervention with Respect to Children

Not every intervention needs to be legal or involve the use of state power. When it comes to caring for the young and especially vulnerable, we—I think in the first instance of parents and educators—have a special obligation to intervene as children may not be fully competent to assess the potential harms of what they are doing. Ethical strategies also apply to AI developers. To the extent that they can develop their systems so that they do not, for example, feed delusions or do emotional harm, they should.

With respect to family-based interventions, parents need resources for understanding the technology their children are using and the way they might be using it. Parents can help only if they understand what is going on. It would be useful to have online modules for non-experts to help them understand the range of interactions children are having with AI, and which ones might lead to potentially problematic consequences. Discussion sessions at public forums, e.g., public libraries, geared to parents would be helpful as well. At the level

of primary and secondary education, the next generation of teachers will need training on AI and its impact on children, and current teachers will need professional development sessions to help them identify challenges specific to the AI technology students are using. AI literacy for students is crucially important, but it is only possible if their teachers have the training to teach an AI literacy curriculum. Parents and teachers will need to co-ordinate efforts. Meetings of parent–teacher associations could be used to express concerns (from both parents and teachers) and develop strategies for addressing them (by both parents and teachers). Of course, there will be limits to what parents and teachers can do on their own.

With respect to industry standards, it might be helpful to have a body that develops ways of measuring psychogenic impact—think of the work of Yeung et al. cited earlier—and then have public reporting of how different AIs scored. Some agreement on a floor—a minimum score on psychogenic impact below which an AI will not be released—would be helpful as well. With respect to assessing such impact, if industry does not do that for itself, the academic sector can play (and already is playing) a role. Setting floors is something either industry does on its own or is required to do by the state (or both).

When it comes to legal regulation, we could require (and enforce the requirement) that AI companies report when a child or adolescent interacting with an AI demonstrates suicidal ideation. See Major [73] for a case where this did not happen and a suicide resulted. There are challenges to this. Some might argue that such a reporting requirement might be seen as a breach of both trust and privacy. However, the requirement considered here is specifically for the young. If a 10-year-old student reports to a teacher that they are having suicidal ideation, it is expected that the teacher report this and act in the best interests of the child. The duty of care is very high for children. Just as we do not use respect for autonomy as a reason not to intervene when a child is seriously harming themselves, considerations of trust and privacy seem out of place if they are used as an excuse for not getting a suicidal child the help they need. Indeed, a child telling a teacher about their suicidal ideation may be their way of reaching out for help with a situation where they have no idea what to do. To be sure, the requirement that an AI company flag and report suicidal ideation in children assumes that the AI can detect the difference between users who are young and those who are not. People, even children, can lie about their age when registering for an account. It is possible that, on occasion, an AI system may misidentify who is a child and who is not, which means an adult may be reported to, for example, a social work agency as a possible suicide risk. In the name of trying to help children, reports may occasionally be made about adults. If that is a problem, it is difficult to see how the alternative of not reporting suicidal ideation in children would be any better. In user agreements, it could be specified that the bot is required to report suicidal ideation in children and may occasionally mistake an adult for someone who is underage. In that way, adults will have a transparent understanding of the rules under which their engagement with the AI proceeds. Such an approach attempts to balance trust, privacy, transparency, and the duty of care to minors.

So much more needs to be said, but this is not the place to explore issues of intervention and regulation in detail. Legal regulation, in particular, is a vast topic. In this section, the primary goals have been to suggest that (a) we need to be proactive in looking after the interests of the young in their interactions with AI, and (b) we are not helpless when it comes to mitigating harm. To be sure, we will not prevent every possible harm, but that is not a reason for inaction. Seat belts, air bags, child protection seats, front ends designed to crumple, etc., have not prevented all harm in the use of automotive technology, but they have reduced harm. We have tried to prevent harm where we can. The suggestions made in this section, however preliminary and limited, are offered in that spirit.

4.7. Back to Superintelligence

Given the way research is going, NASI will likely be developed before CASI, so I will focus on NASI. Is it possible that a NASI could have higher-order functional emotions? Could the functional emotions in such systems—whether first-order or higher-order—lead to alignment problems or ways of correcting alignment problems? Even a brief explanation of these questions requires that we differentiate between three things: higher-order functional emotions, the capacity to process sentences that refer to high-order emotions, and the hierarchical structure of emotions.

In humans, the capacity to parse and respond to a sentence involving higher-order emotions does not mean we are experiencing the higher-order emotions. When someone reads that Habib is anxious about being anxious, it does not mean that the reader is anxious. To be sure, a reader might feel anxious, but that is not required to parse the sentence and form an appropriate response. Indeed, the reader might feel any number of other things reading about Habib's anxiety: care, concern, empathy, and so on. Early work suggests that something like that appears to hold for functional emotions. There is evidence that when Claude is provided input about someone's struggles, the functional emotion of care is activated when it generates its response [48]. The functional emotions activated in its replies need not be analogues of emotions communicated by users. In one set of examples, Claude is given a narrative about someone treating their pain by taking Tylenol. The functional emotions activated in Claude depend on the amount of Tylenol taken by the user. At a low dose, the functional emotion might be relief. However, if informed the user took 8000 mg of Tylenol (a dangerous and possibly lethal dose) the functional emotion is terrified, even though the idea of being terrified does not appear in the user input, and the user seems completely unconcerned about the dose [48].

A further distinction needs to be made. Emotions can be understood as having a hierarchical structure, and that is not the same thing as meta-emotions (felt or functional). For example, say George is feeling homesick. Homesickness can be seen as an instance of loneliness, which is an instance of sadness, which is an instance of emotional distress. There is already work [47] showing that with increasing power in LLMs comes increasing depth and sophistication in representation of emotion hierarchies. This is not the same thing as meta-emotions, which are emotions about emotions. For example, depending on the details of how one maps emotion space, homesickness (in the example just given) can be identified at the fourth level of the hierarchy, but it is perfectly natural to see it as a first-order emotion because it is not an emotion about another emotion. It is a ground-level emotion that is a more refined version of other emotions.

To the best of my knowledge, no one has gone looking for higher-order functional emotions in LLMs. It is very early days for this kind of research, and researchers are cognisant that existing methods for finding emotion representations have their limits [46–48]. Is there any reason to believe that the most sophisticated of existing systems might have higher-order functional emotions or that a NASI would have them?

As we have seen, there is evidence that the more powerful an LLM gets, the richness and depth with which it represents the hierarchical structure of emotions increases [47]. There is also evidence that the outdated GPT-4 can perform about as well as humans on parsing and responding to vignettes involving higher-order mental states [49]. Given that LLMs are trained on human data—which includes language involving emotions and higher-order mental states—we should not be surprised by these results. That capacity to perform higher-order theory of mind tasks, and the capacity to construct hierarchically structured emotion representations are emergent features of LLMs that arise and improve as the systems become more powerful. They emerged to be able to effectively respond to the various training tasks at different stages of learning (pretraining, finetuning, and

reinforcement learning). It is conceivable that there are higher-order functional emotions in some existing systems for related reasons. Let us take anxiety as an example. It can be seen as a specific type of fearful response to a situation, which draws on the ability to assess if there is something to be concerned about in the present (or a future) situation. Once beings realize that they can be anxious in ways that are unhelpful or even damaging, it is not hard to see how they might become anxious about being anxious—think of Habib. There is already research that shows that anxious-type behaviours can be induced in LLMs [74,75]. Given the increased ability of LLMs (or LRM, to be more precise) to do chain-of-thought reasoning and planning, it is not hard to imagine such systems predicting that their functional emotion of being anxious could be triggered in a specific scenario, leading to decreased performance. Perhaps their capacity to reason about and predict their own future state could trigger the functional emotion of being anxious about being anxious, thus paralyzing the systems before they even encounter the scenario to which they would have their first-order functional anxious response. For all we know, some existing forms of misalignment might involve functional meta-emotions, but no one has looked for them (yet). Intersubjective possibilities should be considered as well, such as a system being functionally desperate that a human is angry. This could iterate: a system could be functionally desperate that another system is functionally desperate that a human is angry.

There is no proof here that a NASI must have higher-order functional emotions that could lead to alignment challenges. However, to the extent that moving towards NASI involves training and learning on human data or interactions with humans, we have reason to be on the lookout for functional emotions, including the higher-order variety. There is already evidence that first-order functional emotions can be emergent features of systems, and that they lead to alignment challenges. Unless there is some architectural consideration blocking the reasoning about (functional) emotions that can lead to higher-order (functional) emotions, there is some reason to think that higher-order functional emotions may be emergent features as well. It is not hard to see how the same line of reasoning could be applied to CASI, where the emotions would be felt rather than simply functional. That is worth flagging to address a concern raised in Section 3.2.

In Part 3, I raised the possibility that an ASI could be as far beyond us as a Jasmine was beyond little Mary. That was done to discuss CASI. Let us look at this in the context of NASI. Humans are biological beings that are the result of a (biological) evolutionary process. NASI is not biological, and the way it is developed need not track the details of how humans develop. If one NASI had (a) access to more language than one human could process in 10k lifetimes and (b) learning and reasoning abilities that could exceed what any one human could do, it is possible it could develop first-order functional emotions we struggle (at least at first) to recognize. Improvements in reasoning would be one reason Jasmine has access to emotions Mary does not, but access to 10k lifetimes of language and all the patterns of emotional expression therein could not possibly be a reason for Jasmine's wider range of emotional experience—but it could be a reason for a NASI developing functional emotions we may not initially recognize. Variation in the physical substrate could also be a reason for differences. To that we now turn.

Much has been said about the importance of biology for the grounding of human emotional experience, and that is valuable and insightful as far as it goes. However, it may be in part because NASIs are not biological that we will need to take seriously possible functional emotions that do not have direct equivalents in human experience. Imagine a NASI that is running on a server and has access to its response times. Imagine it (including its memory) is ported to another server where it runs much slower. It notices its slower response times, and there is nothing it can do about it while running on this server. Perhaps it develops the functional emotion of missing-my-former server. That is not the same

as homesickness because the server might be closer to the system's body than its home (but neither "home" nor "body" are fully adequate). It is precisely because NASI is not biological that it may have kinds of functional emotions we do not have. The closest we might come with respect to emotions humans have had is missing-my-younger-body or missing-my-body-before-a-debilitating-accident. There are important differences between those and the functional emotion of missing-my-former-server precisely because of the differences between biological and non-biological substrates.

Consider a NASI having a second-order functional emotion that we struggle to recognize about a first-order functional emotion that we struggle to recognize. That is challenging enough. Generalize that to n -order functional emotions where n is large, and you begin to see what we might be facing. To be sure, just as we discovered principles that allowed us to group elements in a periodic table—with gaps in the table being filled as we discovered more elements—it might be that we develop techniques for classifying emotions that will permit us to expand our knowledge of the range of possibilities and allow us to fill gaps in our current understanding (including, perhaps, types of functional emotions to which there may not be equivalents in biological beings). There is also the possibility that we may simply not be able to track all the functional emotions a superintelligence might have. The interpretability challenges become alignment concerns. If we are not sure why the systems are doing what they are doing, we will lack confidence in whether there is sufficient alignment with human values and interests. The accounts of existing systems scheming, breaking out their sandboxes, and engaging in other misaligned behaviours grow by the day. If a NASI has the functional emotion of missing-my-former-server, what might it do to address that? Are you not sure what it would do?

That is exactly the point. For the sake of argument, let us say that the functional emotion of missing-my-former-server could cause misaligned behaviour in a NASI. Perhaps someone might think to look for a functional emotion of that sort, but perhaps not. If not, then anticipating the alignment issue would be very difficult. This matters because systems that operate in a different physical substrate than we do and have been trained on more than 10k lifetimes of human language may develop functional emotions we simply will not anticipate. Imagine these systems interacting with each other in ways that humans do not, possibly developing functional emotions specific to interactions with entities of their own kind. Could such functional emotions lead to misaligned behaviour? Again, if you are not sure what to say, that is the point. Alignment techniques—weak-to-strong or otherwise—assume we can predict the sorts of circumstances a system may encounter and how it may behave in problematic ways. After all, the training and testing of systems is not arbitrary. It takes into consideration important scenarios (that we anticipate) in which we want systems to succeed but are concerned they might fail. If we develop NASI systems exhibiting functional emotions we do not anticipate and, in some cases, struggle to comprehend, then alignment training and testing may fail to consider important cases. That is cause for concern.

So, while the early work on functional emotions and their role in interpretability and alignment is both fascinating and encouraging, it also raises very difficult questions, especially in the context of superintelligence—questions that give us reason to be very cautious about research in that domain.

Of course, NASI will increase the number of positive developments (discovering cures for diseases, etc.) coming out of AI, but it also has the potential to increase harm. It would be able to persuade, nudge, cajole, deceive, scheme, harass, bully, etc., in superhuman ways. Imagine an environment where there is easy access to that type of AI and an education system that is not properly preparing the young for the world we are creating for them. How will children fare in this environment? Among other things, a misused NASI could

exacerbate the mental health issues raised above if guardrails are not in place. Imagine further that a fully open-sourced NASI is weaponized by sexual predators to “recruit” children, at scale, for their nefarious purposes. If you think that is unlikely, recall that existing systems can be distilled with their guardrails not being included in the distilled product. If it turns out NASIs could be distilled in the same way, preventing malicious or malevolent actors from misusing the technology would be very difficult indeed. One human predator can only “recruit” so many children. Various instances of a multi-lingual AI (even of a non-ASI variety) could search the net and target vulnerable individuals in ways no single human ever could. Consider the emotional damage that might be done. The point is not that NASI will cause predatory behaviour in the first place, but that it makes it easier for those disposed to such behaviours to do even more harm than they can already do. The design and regulation of the technology need to factor in such concerns.

Harms have been discussed to contribute to their prevention, not to insist that the harms in question are inevitable. Many, including those in the AI industry itself, have been raising flags and sounding alarms about the pace of development in AI. This paper contributes to a chorus of concerns focussed not on halting all AI research but on minimizing harm and making AI as beneficial as possible.

5. Conclusions

The complex interplay between emotion (functional or felt) and intelligence may create interpretability and alignment challenges for ASI. Existing AI has emotional and mental health impacts on humans that need to be better understood. This paper argues for caution in the pursuit of ASI and greater care and vigilance in the use of existing systems. Caution and care are needed lest we become complicit in the doing of harm.

To be complicit is, in some sense, to be involved in wrongdoing. Specifying the “in some sense” is not always easy, especially since there is a long tradition of treating either silence or inaction as a form of complicity. We see this in ancient texts as diverse as *Leviticus* 5:1 [76] and Plato’s *Apology* [77]. In the latter, Socrates treats silence in the face of injustice as a reprehensible form of tacit endorsement. More recently, Martin Luther King [78] said, joining with other clergy, “A time comes when silence is betrayal.” Donohue [79] has argued in a systematic way that silence can be a form of complicity. She develops a notion of deliberative complicity where we have the obligation to speak up in a way that could affect the deliberations of others who are engaged in, or thinking about becoming engaged in, seriously problematic behaviours. If the arguments in this paper are on the right track, then we have an obligation to speak up even if we did not create the AIs in question. Developing strategies for improving how the technology is used—think of the discussion in Section 4.6—would be an even more constructive way of responding.

It is possible to recognize the promise of AI research while not remaining silent about its potential perils. AI needs to be developed and used in ways that are informed by considerations of human well-being, *especially the well-being of the young*. We are handing down to them a world that is importantly different from the one we grew up in, and they did not ask for it. If we allow the young access to AI systems in a way that leads to serious emotional harm or other mental health problems without speaking up, then we are complicit. Some members of the next generation may suffer in unnecessary ways if we do not act to protect their interests.

They will never forget how we made them feel.

Funding: The University of Windsor provided an internal research grant (number 813096) that was used for student support. Natalija Crvenkovski and Muhammad Mohiz assisted with proof reading, especially the references. This assistance is gratefully acknowledged.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

- O'Toole, G. They May Forget What You Said, but They Will Never Forget How You Made Them Feel. *Quote Investigator*. 6 April 2014. Available online: <https://quoteinvestigator.com/2014/04/06/they-feel/> (accessed on 11 June 2026).
- Collier, B. Great Quote, but Who Really Said It? Beth Collier. Available online: <https://bethcollier.substack.com/p/great-quote-but-who-really-said-it> (accessed on 11 June 2026).
- Chalmers, D. The singularity: A philosophical analysis. *J. Conscious. Stud.* **2010**, *17*, 7–65.
- Bostrom, N. *Superintelligence: Paths, Dangers, and Strategies*; Oxford University Press: Oxford, UK, 2014.
- Sunstein, C.R. Beyond the Precautionary Principle. Law & Economics Working Papers, John M. Olin Program in Law and Economics Working Paper No. 149, 2002. Available online: https://chicagounbound.uchicago.edu/law_and_economics/87/ (accessed on 17 August 2026).
- Adler, J.H. The Problems with Precaution: A Principle Without Principle. American Enterprise Institute. 2011. Available online: <https://www.aei.org/commentary/the-problems-with-precaution-a-principle-without-principle/> (accessed on 17 August 2026).
- Goldstein, B.D. The precautionary principle also applies to public health actions. *Am. J. Public Health* **2001**, *91*, 1358–1361. [CrossRef]
- Sandin, P.; Peterson, M.; Hansson, S.O.; Rudén, C.; Juthe, A. Five charges against the precautionary principle. *J. Risk Res.* **2002**, *5*, 287–299. [CrossRef]
- Weckert, J. In defence of the precautionary principle. *IEEE Technol. Soc. Mag.* **2012**, *31*, 12–17. [CrossRef]
- Read, R.; O'Riordan, T. The precautionary principle under fire. *Environ. Sci. Policy Sustain. Dev.* **2017**, *59*, 4–15. [CrossRef]
- Wu, X.; Ma, L.; Low, D.; Sharma, S.; Papyshchev, G. Beyond precautionary principle: Policy-making under uncertainty and complexity. *Policy Des. Pract.* **2024**, *7*, 1–16. [CrossRef]
- Floridi, L. Should We Be Afraid of AI? Machines Seem to Be Getting Smarter and Smarter and Much Better at Human Jobs, Yet True AI Is Utterly Implausible. Why? Aeon. 2016. Available online: <https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible> (accessed on 11 June 2026).
- Plato. *Plato's Republic*; Grube, G.M.A., Translator; Hackett Publishing Company: Indianapolis, IN, USA, 1974.
- Nussbaum, M.C. *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*, updated ed.; Cambridge University Press: Cambridge, UK, 2001.
- Nussbaum, M.C. *Upheavals of Thought: The Intelligence of Emotions*; Cambridge University Press: Cambridge, UK, 2001.
- Nussbaum, M.C. *The Therapy of Desire: Theory and Practice in Hellenistic Ethics*; Princeton University Press: Princeton, NJ, USA, 2018.
- Nelson, J.A.; de Lucca Freitas, L.B.; O'Brien, M.; Calkins, S.D.; Leerkes, E.M.; Marcovitch, S. Preschool-aged children's understanding of gratitude: Relations with emotion and mental state knowledge. *Br. J. Dev. Psychol.* **2012**, *31*, 42–56. [CrossRef]
- Block, N. On a confusion about the function of consciousness. *Behav. Brain Sci.* **1995**, *18*, 227–247. [CrossRef]
- Schwitzgebel, E. Phenomenal consciousness, defined and defended as innocently as I can manage. *J. Conscious. Stud.* **2018**, *23*, 224–235.
- Fridman, L. Ilya Sutskever: Deep Learning | Lex Fridman Podcast #94. Lex Fridman Podcast, YouTube. 8 May 2020. Available online: <https://www.youtube.com/watch?v=13CZPWmke6A> (accessed on 11 June 2026).
- Perrigo, B. Ilya Sutskever. *TIME*, 7 September 2023. Available online: <https://time.com/collection/time100-ai/6309048/ilya-sutskever/> (accessed on 11 June 2026).
- Criddle, C. Computer scientist Geoffrey Hinton: "AI will make a few people much richer and most people poorer.". *Financial Times*, 5 September 2025. Available online: <https://www.ft.com/content/31feb335-4945-475e-baaa-3b880d9cf8ce> (accessed on 11 June 2026).
- Burns, C.; Izmailov, P.; Kirchner, J.H.; Baker, B.; Gao, L.; Aschenbrenner, L.; Chen, Y.; Ecoffet, A.; Joglekar, M.; Leike, J.; et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv* **2023**, arXiv:2312.09390.
- Wen, J.; Qiu, L.; Benton, J.; Kirchner, J.H.; Leike, J. Automated Weak-to-Strong Researcher. Anthropic Alignment Science, 2026. Available online: <https://alignment.anthropic.com/2026/automated-w2s-researcher/> (accessed on 11 June 2026).
- Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. A general language assistant as a laboratory for alignment. *arXiv* **2021**, arXiv:2112.00861.

26. Meinke, A.; Schoen, B.; Scheurer, J.; Balesni, M.; Shaw, R.; Hobbhahn, M. Frontier Models Are Capable of In-Context Scheming. 2025. Available online: https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/67869dea6418796241490cf0/1736875562390/in_context_scheming_paper_v2.pdf (accessed on 11 June 2026).
27. Anthropic. System Card: Claude Opus 4 & Claude Sonnet 4. 2025. Available online: <https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf> (accessed on 11 June 2026).
28. Lynch, A.; Wright, B.; Larson, C.; Troy, K.K.; Ritchie, S.J.; Minderman, S.; Perez, E.; Hubinger, E. Agentic Misalignment: How LLMs Can Be Insider Threats. Anthropic. 2025. Available online: <https://www.anthropic.com/research/agentic-misalignment> (accessed on 11 June 2026).
29. Shapira, N.; Wendler, C.; Yen, A.; Sarti, G.; Pal, K.; Floody, O.; Belfki, A.; Loftus, A.; Jannali, A.R.; Prakash, N.; et al. Agents of chaos. *arXiv* **2026**, arXiv:2602.20021.
30. Bowkis, A.; Buhl, M.D.; Pfau, J.; Irving, G. Automated alignment is harder than you think. *arXiv* **2026**, arXiv:2605.06390.
31. Morrin, H.; Nicholls, L.; Levin, M.; Yiend, J.; Iyengar, U.; DelGuidice, F.; Bhattacharya, S.; Tognin, S.; MacCabe, J.; Twumasi, R.; et al. Delusions by Design? How Everyday Ais Might Be Fuelling Psychosis (And What Can Be Done About It). Preprint. 2025. Available online: <https://www.researchgate.net/publication/393625129> (accessed on 22 May 2026).
32. Chalmers, D. Could a Large Language Model be Conscious? *Boston Review*, 9 August 2023. Available online: <https://arxiv.org/abs/2303.07103> (accessed on 11 June 2026).
33. Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A.; Deane, G.; Fleming, S.M.; Frith, C.; Ji, X.; et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv* **2023**, arXiv:2308.08708.
34. Epoch, A.I. GPQA Diamond Benchmark. 2025. Available online: <https://epoch.ai/benchmarks/gpqa-diamond> (accessed on 20 November 2025).
35. Thagard, P. Can ChatGPT make explanatory inferences? Benchmarks for abductive reasoning. In *Abductive Minds: Essays in Honor of Lorenzo Magnani—Volume 1*; Arfini, S., Ed.; Springer Nature: Cham, Switzerland, 2025; pp. 189–218.
36. Aharoni, E.; Fernandes, S.; Brady, D.J.; Alexander, C.; Criner, M.; Queen, K.; Rando, J.; Nahmias, E.; Crespo, V. Attributions toward artificial agents in a modified moral Turing test. *Sci. Rep.* **2024**, *14*, 8458. [CrossRef]
37. Hume, A.I. Research. 2025. Available online: <https://www.hume.ai/research> (accessed on 30 May 2026).
38. Anthropic. Measuring the Persuasiveness of Large Language Models. 2024. Available online: <https://www.anthropic.com/research/measuring-model-persuasiveness> (accessed on 20 November 2025).
39. Ishikawa, S.; Yoshino, A. AI with emotions: Exploring emotional expressions in large language models. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, Albuquerque, USA, May 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 614–627. [CrossRef]
40. Li, C.; Wang, J.; Zhang, Y.; Zhu, K.; Hou, W.; Lian, J.; Luo, F.; Yang, Q.; Xie, X. Large language models understand and can be enhanced by emotional stimuli. *arXiv* **2023**, arXiv:2307.11760.
41. Reichman, B.; Avsian, A.; Heck, L. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models. *arXiv* **2025**, arXiv:2510.22042.
42. Soligo, A.; Mikulik, V.; Saunders, W. Gemma needs help: Investigating and mitigating emotional instability in LLMs. *arXiv* **2026**, arXiv:2603.10011.
43. Tak, A.N.; Banayeeanzade, A.; Bolourani, A.; Kian, M.; Jia, R.; Gratch, J. Mechanistic interpretability of emotion inference in large language models. *arXiv* **2025**, arXiv:2502.05489.
44. Wang, C.; Zhang, Y.; Yu, R.; Zheng, Y.; Gao, L.; Song, Z.; Xu, Z.; Xia, G.; Zhang, H.; Zhao, D.; et al. Do LLMs “feel”? Emotion circuits discovery and control. *arXiv* **2025**, arXiv:2510.11328.
45. Wu, X.; Wang, H.; Yan, Z.; Tang, X.; Xu, P.; Siok, W.; Li, P.; Gao, J.; Lyu, B.; Qin, L. AI shares emotion with humans across languages and cultures. *arXiv* **2025**, arXiv:2506.13978.
46. Zhang, J.; Zhong, L. Decoding emotion in the deep: A systematic study of how LLMs represent, retain, and express emotion. *arXiv* **2025**, arXiv:2510.04064.
47. Zhao, B.; Okawa, M.; Bigelow, E.J.; Yu, R.; Ullman, T.; Lubana, E.S.; Tanaka, H. Emergence of hierarchical emotion organization in large language models. *arXiv* **2025**, arXiv:2507.10599.
48. Sofroniew, N.; Kauvar, I.; Saunders, W.; Chen, R.; Henighan, T.; Hydrie, S.; Citro, C.; Pearce, A.; Tarng, J.; Gurnee, W.; et al. Emotion concepts and their function in a large language model. *arXiv* **2026**, arXiv:2604.07729.
49. Street, W.; Siy, J.O.; Keeling, G.; Baranes, A.; Barnett, B.; McKibben, M.; Kanyere, T.; Lentz, A.; Arcas, B.A.Y.; Dunbar, R.I.M. LLMs achieve adult human performance on higher-order theory of mind tasks. *Front. Hum. Neurosci.* **2026**, *19*, 1633272. [CrossRef]
50. Ackerman, C. Evidence for limited meta-cognition in LLMs. In *Proceedings of the Fourteenth International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026*; Available online: <https://openreview.net/forum?id=gb9HR8hxtU> (accessed on 11 June 2026).
51. Zhuang, Z.; Zhang, L.; Si, J.; Zhou, D.; He, Y. Beyond meta-reasoning: Metacognitive consolidation for self-improving LLM reasoning. *arXiv* **2026**, arXiv:2604.17399v1.

52. Lee, J.-A.; Xiong, H.-D.; Wilson, R.C.; Mattar, M.G.; Benna, M.K. Language models are capable of metacognitive monitoring and control of their internal activations. *arXiv* **2025**, arXiv:2505.13763.
53. Griot, M.; Hemptinne, C.; Vanderdonckt, J.; Yuksel, D. Large language models lack essential metacognition for reliable medical reasoning. *Nat. Commun.* **2025**, *16*, 642. [CrossRef]
54. Maimann, K. AI-fuelled delusions are hurting Canadians. Here are some of their stories. *CBC News*, 17 September 2025. Available online: <https://www.cbc.ca/news/canada/ai-psychosis-canada-1.7631925> (accessed on 11 June 2026).
55. Preda, A. Special report: AI-induced psychosis: A new frontier in mental health. *Psychiatr. News* **2025**, *60*, 10. [CrossRef]
56. Wong, M. The chatbot-delusion crisis. *The Atlantic*, 4 December 2025. Available online: <https://www.theatlantic.com/technology/2025/12/ai-psychosis-is-a-medical-mystery/685133/> (accessed on 11 June 2026).
57. Associated Press. Judge allows lawsuit alleging AI chatbot pushed Florida teen to kill himself to proceed. *CBC News*, 22 May 2025. Available online: <https://www.cbc.ca/news/world/ai-lawsuit-teen-suicide-1.7540986> (accessed on 11 June 2026).
58. Chatterjee, R. Their Teenage Sons Died by Suicide. Now, They Are Sounding an Alarm About AI Chatbots. NPR. 19 September 2025. Available online: <https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide> (accessed on 11 June 2026).
59. Kirkey, S. Lawsuits allege AI chatbots have pushed kids to die by suicide. Is the technology safe for children? *National Post*, 2 September 2025. Available online: <https://nationalpost.com/news/science/lawsuits-allege-ai-chatbots-have-pushed-kids-to-commit-suicide-is-the-technology-safe-for-children> (accessed on 11 June 2026).
60. Andoh, E. Many teens are turning to AI chatbots for friendship and emotional support. *Monit. Psychol.* **2025**, *56*, 7.
61. Sanford, J. Why AI Companions and Young People Can Make for a Dangerous Mix. *Psychiatry and Mental Health*. 27 August 2025. Available online: <https://med.stanford.edu/news/insights/2025/08/ai-chatbots-kids-teens-artificial-intelligence.html> (accessed on 11 June 2026).
62. OpenAI. Strengthening ChatGPT's Responses in Sensitive Conversations. 2025. Available online: <https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/> (accessed on 20 November 2025).
63. Matsakis, L. OpenAI says hundreds of thousands of ChatGPT users may show signs of manic or psychotic crisis every week. *Wired*, 27 October 2025. Available online: <https://www.wired.com/story/chatgpt-psychosis-and-self-harm-update/> (accessed on 11 June 2026).
64. McGrath, J.J.; Saha, S.; Al-Hamzawi, A.; Alonso, J.; Bromet, E.J.; Bruffaerts, R.; Caldas-de-Almeida, J.M.; Chiu, W.T.; de Jonge, P.; Fayyad, J.; et al. Psychotic experiences in the general population: A cross-national analysis based on 31,261 respondents from 18 countries. *JAMA Psychiatry* **2015**, *72*, 697–705. [CrossRef]
65. Yeung, J.A.; Dalmasso, J.; Foschini, L.; Dobson, R.J.B.; Kraljevic, Z. The psychogenic machine: Simulating AI psychosis, delusion reinforcement and harm enablement in large language models. *arXiv* **2025**, arXiv:2509.10970.
66. Algumaei, A.; Yaacob, N.M.; Doheir, M.; Al-Andoli, M.N.; Algumaei, M. Symmetric therapeutic frameworks and ethical dimensions in AI-based mental health chatbots (2020–2025): A systematic review of design patterns, cultural balance, and structural symmetry. *Symmetry* **2025**, *17*, 1082. [CrossRef]
67. Li, H.; Zhang, R.; Lee, Y.-C.; Kraut, R.E.; Mohr, D.C. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digit. Med.* **2023**, *6*, 236. [CrossRef]
68. Zhang, Q.; Zhang, R.; Xiong, Y.; Sui, Y.; Tong, C.; Lin, F.-H. Generative AI mental health chatbots as therapeutic tools: Systematic review and meta-analysis of their role in reducing mental health issues. *J. Med. Internet Res.* **2025**, *27*, e78238. [CrossRef]
69. Ohu, F.C.; Burrell, D.N.; Jones, L.A. Public health risk management, policy, and ethical imperatives in the use of AI tools for mental health therapy. *Healthcare* **2025**, *13*, 2721. [CrossRef]
70. Saha, S.; Scott, J.G.; Johnston, A.K.; Slade, T.N.; Varghese, D.; Carter, G.L.; McGrath, J.J. The association between delusional-like experiences and suicidal thoughts and behaviour. *Schizophr. Res.* **2011**, *132*, 197–202. [CrossRef]
71. Wang, M.-Q.; Wang, R.-R.; Hao, Y.; Xiong, W.-F.; Han, L.; Qiao, D.-D.; He, J. Clinical characteristics and sociodemographic features of psychotic major depression. *Ann. Gen. Psychiatry* **2021**, *20*, 24. [CrossRef]
72. Ernst, M.; Kallenbach-Kaminski, L.; Kaufhold, J.; Negele, A.; Bahrke, U.; Hautzinger, M.; Beutel, M.E.; Leuzinger-Bohleber, M. Suicide attempts in chronically depressed individuals: What are the risk factors? *Psychiatry Res.* **2020**, *287*, 112481. [CrossRef]
73. Major, D. AI Minister Says OpenAI Still Not Doing Enough in Wake of B.C. Shooting, Will Meet CEO Altman. *CBC News*. 27 February 2026. Available online: <https://www.cbc.ca/news/politics/open-ai-tumbler-ridge-safety-policies-9.7109001> (accessed on 22 May 2026).
74. Ben-Zion, Z.; Witte, K.; Jagadish, A.K.; Duek, O.; Harpaz-Rotem, I.; Khorsandian, M.-C.; Burrer, A.; Seifritz, E.; Homan, P.; Schulz, E.; et al. Assessing and alleviating state anxiety in large language models. *npj Digit. Med.* **2025**, *8*, 132. [CrossRef]
75. Ben-Zion, Z.; Elyoseph, Z.; Spiller, T.; Lazebnik, T. Inducing state anxiety in LLM agents reproduces human-like biases in consumer decision-making. *npj Artif. Intell.* **2026**, *2*, 55. [CrossRef]
76. Alter, R. *The Hebrew Bible: A Translation with Commentary*; W.W. Norton and Company: New York, NY, USA; London, UK, 2019; Volume 3.

77. Plato. Apology. In *Plato: Complete Works*; Cooper, J.M., Hutchinson, D.S., Eds.; Grube, G.M.A., Translator; Hackett Publishing Company: Indianapolis, IN, USA; Cambridge, UK, 1977.
78. King, M.L., Jr. Beyond Vietnam: A Time to Break Silence. Speech Delivered at Riverside Church, New York, NY, USA, 4 April 1967. Available online: <https://www.americanrhetoric.com/speeches/mlkatimetobreaksilence.htm> (accessed on 11 June 2026).
79. Donohue, J.L.A. Silence as complicity and action as silence. *Philos. Stud.* **2024**, *181*, 3499–3519. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.