AI & SOCIETY
https://doi.org/10.1007/s00146-024-02022-x

MAIN PAPER

Can machines have emotions?

Anand Jayprakash Vaidya[1,2]

Received: 9 April 2024 / Accepted: 15 July 2024

Abstract
In this paper I articulate the question of whether machines can have emotions. I then reject a common argument against why they cannot have emotions based on the lack of a capacity for feelings. The goal of this paper is not to decisively show that machines can have emotions, but to decisively show that the naïve argument for the conclusion that they cannot needs to be critically examined. I argue that machines that have artificial general intelligence can have emotions based on having the capacity to make judgments that are essential and constitutive of certain emotions, such as anger. I argue against the view that phenomenological or physiological profiles are essential to anger on the basis of emotion regulation. I consider a long list of objections to the position that machines can have emotions.

Keywords AI · Intelligence · Emotions · Understanding · Judgement · Emotion Regulation

Part I: An example for reflection

Interviewer: Do you believe that HAL has genuine emotions?

Frank Poole: Well he acts like he has genuine emotions. Of course, he is programmed that way to make it easier to for us to talk to him. But as to whether or not he has real feelings is something that I don't think anyone can truthfully answer.

————-

HAL: Dave, stop it. Stop it, will you. Stop, Dave…

HAL: I am afraid.

HAL:Dave my mind is going. I can feel it. I can feel it. My mind is going.

2001: A Space Odyssey –Stanley Kubrick.

1 Science fiction, philosophy, and the question

Works of science fiction, like Stanley Kubrick's 1968, *2001: a Space Odyssey*, have explored the question: can machines have emotions? In the first piece of dialog above, for example, a human, Frank Poole, tries to assess whether a machine can have an emotions. In the second, a machine, HAL, asserts that it has the specific emotion of fear, and that it can feel its mind going, which involves the capacity to have a phenomenological state of feeling.

In the *Twilight Zone*, *Season 2* episode, *The Lateness of the Hour*, which originally aired in 1960, we meet Jana, who lives with her parents, Dr. and Mrs. Loren, and the set of robot servants that Dr. Loren has built to serve their needs. Jana resents the robots, blaming them for the cloistered life she is forced to live. She wants a family of her own and insists that her father destroy the robots, which he does. When she threatens to leave anyway, he is forced to tell her the truth. Just as his butler served his need for his pipe to be refilled, and the maid served Mrs. Loren's need to be massaged, Jana served their need for offspring: Jana too is a robot. It is a truth that Jana cannot handle, and Dr. Loren reprograms her to be a maid.

In *Star Trek VII: generations*, Season 4 episode, *Decent Part 2*, which originally aired in 1993, data installs an emotion chip that enables him to have emotions. The episode raises the question: what is essential for having an emotion?

I would like to thank Noam Cook for extensive and valuable discussion of this work. In addition, thanks go to audiences at IIC, Delhi; University of Hawaii, Manoa; University of Stockholm, Sweden; and University of Hong Kong. I would also like to thank Susan Schneider's research group for helpful encouragement as well as Anandi Hattiangadi and Manjula Rajan for critical feedback.

✉ Anand Jayprakash Vaidya
anand.vaidya@sjsu.edu

[1] San Jose State University, San Jose, USA

[2] University of California, Los Angeles, USA

What capacities, must *x*, whether *x* is a plant, a single cell organism, a non-human animal, a machine, or indeed, a human, possess, such that *x* can have emotions or a particular range of emotions?

The question of this inquiry is: can machines have emotions? In order to make progress on this question I will turn to Turing's (1950) classic, *Computing Machinery and Intelligence*, because he considers a similar question, concerning thought rather than emotions. His inquiry concerns: can machines think? In his work he proposes a methodology that on one hand I agree with and on the other hand I disagree with. In tackling his question he first investigates the terms 'machine' and 'think' before proposing a modified version of the question based on the imitation game. Roughly, the imitation game tests whether a system is intelligent by testing whether it could imitate a human being in a conversational exchange such that a human would judge the system to be human.[1] After describing the imitation game, Turing says:

> We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?' (Turing 1950, p. 434)

Following Turing's method of investigating the terms in his question, I will begin my inquiry by offering an account of the terms in the question: can machines have emotions? However, I will not follow Turing in changing the question to an alternative question involving a behavioral test, such that we are justified in saying that a machine has emotions just in case we cannot tell the difference behaviorally between it and a human.

Can: there are at least three different uses of 'can' that one can philosophically engage with. Often philosophers are discussing the logical use of 'can.' On this use, to say that machines can have emotions is to say that from a logical point of view machines can have emotions. That is, there is no contradiction between the sentences '*x* is a machine' and '*x* has an emotion', they are logically compatible, so it is not impossible for a machine to have an emotion. For example, the sentences '*x* is a circle' and '*x* is a square' are contradictory, and thus, the sentences jointly express a logical impossibility. One might argue that the definitions of 'machine' and 'emotion' do not lead to a contradiction as one finds in the case of 'circle' and 'square' and thus in the logical sense machines can have emotions. I will not explore this use of 'can.' It is too wide. Another use of 'can' is the physical use. On this use, to say that machines can have emotions is to say that from the point of view of the laws of physics, chemistry, biology, neuroscience, cognitive science, and psychology there is no contradiction to be found in a machine that has emotions. However, the physical sense of 'can' is problematic in three ways: completion, reduction, and essence. First, it is not clear that the laws of any of these special sciences are complete.[2] Are all of the laws of physics currently known? Second, it is not clear whether the laws of some special sciences are reductively determined by the laws of other special sciences.[3] Do the laws of chemistry reductively determine the laws of neuroscience? Third, it isn't clear that a scientific account of a phenomenon offers us an account of the essence of the phenomenon that is open to multiple realizability in a wide sense.[4] So, I will forgo exploring the question with respect to the physical use of 'can.' I am interested in the metaphysical reading of 'can' where it expresses metaphysical modality that is tied to the nature of kinds.[5] In this sense, to say that machines can have emotions is to say that what it is essentially for something to be a machine is not inconsistent with what it is for something to essentially be an emotion. In short, that the essence of being a machine is compatible with the essence of being an emotion. So, when I ask: can a machine have an emotion? I am really asking: given what is essential to being a machine and what is essential to being an emotion, is it metaphysically possible for something to be a machine and have an emotion?

Machine: in terms of the question I am asking it is more common to hear people ask: can an AI, an LLM in particular, an Android, or Robot have emotions? The emphasis is not on simply being a machine, but on being an intelligent machine of a certain kind. So, why is the focus here on machines? The main reason why is to focus on the material that the machine is made out of as opposed to the kind of

[1] I borrow this description of the imitation game from Anandi Hattiangadi.

[2] For example, we might take note of the position known as promissory note physicalism, on which although all the laws of physics are not currently known, physicalism is nevertheless true, in the sense that no law of the universe that will be discovered would refute that claim that the physical facts determine all the facts.

[3] See the classical debate over this issue between Fodor (1974) and Kim (1992).

[4] Kripke (1980) argues that for some mental phenomena, such as pain, the essence is not given by any neurological property that co-varies with pain experience, whose essence is phenomenologically given.

[5] Although Kripke (1980) does say that physically modality might turn out to be necessity *tout court* (metaphysical modality), I take it that he introduces the idea of metaphysical modality as kind of modality tied to the essences of kinds and particulars, which is distinct from logical and physical necessity. See Fine (1995, 2002) for an expression of this picture.

architecture that renders it intelligent in some sense. On my view, there are two kinds of machines that need to be distinguished. First we can define a machine as having two components, hardware and software. Some machines, such as 1974 Mustang's engine, have no software. Other machines, such as a 1982 Apple 2e, have both hardware and software. Based on that distinction, a dry machine is a machine whose hardware is all inorganic. No part of the machine's hardware is composed out of organic matter. By contrast, a wet machine has hardware composed out of organic matter. For some philosophers, humans are simply wet machines.[6] I am interested in inquiring into whether or not dry machines can have emotions. So from here on out, when I speak of machines, I mean dry machines. There is a lot research that shows that artificial intelligence can be made to work within wet systems, either cellular or by placing a chip inside of an already existing creature. For example, we can look at recent work by Neuralink, shows a patient is able to play chess due to an implanted chip in their brain.[7] In order to make the inquiry into whether machines can have emotions tackle the issue of how something inorganic can have an emotion, I am taking on the hard case of dry machines. In fact one might say that *the hard problem of emotions* is to explain how something completely inorganic can have an emotion. I am assuming that the case of dry machines having emotions is much more controversial than wet machines having emotions. Again, if a human can have an emotion, and part of their brain is replaced by chips, it would seem that unless the implanting of a specific chip blocked emotions from occurring, the presence of chips that added functionality would not be a hindrance to realizing emotions. The background idea is that since we are wet machines, and we have emotions, it is harder to tackle the issue of whether or not dry machines can have emotions. For it may be the case that having an emotion depends on the hardware being realized in a certain substrate, the entity in question being alive, or even having a certain evolutionary history. Finally, in terms of the underlying mechanisms in the hardware, I am not concerned with whether a neural net is involved or simply classical deductive logic. Large Language Models, LLMs, housed completely in inorganic matter are machines on my view. So, another version of my question is: can LLMs housed in inorganic matter have emotions?

Have: there are at least two different notions of 'have.' On the property instantiation view, for x to have P is for x to instantiate P. On the property exemplification view, for x to have P is for x to exemplify P. The difference is the following. Anything that exemplifies a property instantiates the property. However, some instantiated properties are not exemplified. Exemplification is a metaphysical account of what it is to have a property. Instantiation is a logical account of what it is to have a property. In asking whether or not machines can have emotions, I am using 'have' in the exemplification sense, and not in the instantiation sense. I am drawing this distinction on the basis of an analysis of a Cambridge change. A Cambridge change happens when a predicate P is true of an object O at one time, but not another; however, there is no intrinsic change rather only a relational change. Anand has the property of being such that his mom lived in Orange County in 2013. Anand has the property of being such that his mom did not live in Orange County in 2014. Using the distinction above, we can say Anand instantiates the property of being such that his mom lives in Orange County in 2013, but he does not exemplify the property. By contrast, Anand exemplifies and instantiates the property of having black hair in 2013. Property exemplification is a form of metaphysical realization. Property instantiation, by contrast, is merely a form of logically possessing a property.

Emotions: in his classic, *What is an Emotion?*, James (1884) takes on the nature of emotions. He is quick to clarify the scope of his inquiry. He says:

> I should say first of all that the only emotions I propose to expressly consider here are those that have a distinct bodily expression. That there are feelings of pleasure and displeasure, of interest and excitement, bound up with mental operations, but have no obvious bodily expression for their consequence, would I suppose, be held true by most readers. (James 1884, p. 189)

He continues by maintaining that the term "standard emotions" is to be used for those emotions for which there is a distinct bodily expression. He says.

> One natural way of thinking about these standard emotions is that the mental perception of some fact excites the mental affection called the emotion, and that this latter state of mind gives rise to the bodily expression. (James 1884, p. 189)

In contrast to this view, he says:

> My thesis is that *the bodily changes follow directly the* PERCEPTION *of the exciting fact, and that our feeling of the same changes as they occur* is the emotion. (James 1884, p. 189-90)

I will soon turn to an evaluation of this theory of emotions for the purpose of investigating whether machines can have emotions. For now, we can simply note, what many

[6] The idea of a natural born cyborg explored by Clark (2003) is one account of humans as organic machine of a certain kind.

[7] See https://www.msn.com/en-us/health/other/video-shows-first-neuralink-brain-chip-patient-playing-chess-by-moving-cursor-with-thoughts/ar-BB1klUQg

philosophers in the philosophy of emotions have noted, there are at least three different properties that emotions are said to possess or are typically associated with. The phenomenological aspect of an emotion is the what it is like aspect of the emotion. Sadness doesn't feel the same way as anger or love. What it is like to undergo sadness is not the same as what it is like to undergo anger or love. The physiological aspect of an emotion is the embodied aspect of the emotion which pertains to how the emotion is associated with physical changes in the body and how it is expressed. A person's body undergoes different physiological changes when the person is sad as opposed to when they are happy or angry. Although there are some physiological changes that are in common between different emotional states, in general the overall physiological state is said to be different. Sadness and anger can both bring about a change in one's heart rate, but supposedly there are other physiological characteristics that are different. Finally, there is the cognitive aspect of an emotion that captures what kind of attitude the emotion is or what judgment is involved in the emotion. Fear and anger differ in terms of the kind of attitude each involves. Love is also different from joy in terms of the attitude each involves. With these three different aspects on the table the leading question about emotions in this inquiry is: which, if any of these, alone or in combination, is essential to emotions? Is it the phenomenological and somatic aspects that are essential in combination or is it, for example, the cognitive aspect alone that is essential?

The question: having defined all of the terms, I will now reduce the scope of the question from an inquiry over emotions in general to a specific emotion, *E*. The general question 'can machines have emotions?' now becomes: is it metaphysically possible for a (dry) machine to exemplify *E* in virtue of exemplifying what is essential to *E*? By 'essential' I mean what is exhaustively requisite for the presence of the emotion, not simply necessary conditions that are not themselves jointly sufficient. Part of what it takes to answer this question is an account of what is essential to a specific emotion *E* across the dimensions of phenomenology, physiology, and cognition either alone or in some combination. I now turn to the argument against the possibility of machines having emotions.

2 The naïve argument against machines having emotions

In order to explore the question, 'Is it metaphysically possible for a machine to exemplify *E* in virtue of exemplifying what is essential to *E*?', it will be instructive to lay out an argument for the most common answer to the question: no!

2.1 *The naïve argument*:

1. Emotions are essentially tied to feelings. It is essential to anything that has an emotion that it has the capacity to feel.
2. Machines cannot feel.
3. So, machines cannot have emotions.

The naïve argument is often not stated in the context of debates about the essential nature of emotions, since those debates concern emotions and not machines. Nevertheless, it is the most common position one finds on the possibility of machines having emotions. That is why I started with, science fiction. In fact, it can be defended by taking James's theory of emotions as the truth about emotions. However, it is the purpose of this paper to argue that the naïve argument is exactly what it is: naïve. In fact the naive argument is also naïve in its more mature form which I will discuss in Sect. 8. My defense does not require a refutation of James's view of emotions, but a defense of how to make sense of an alternative view that can be used to account for how machines can have emotions.

The naïve argument can be voiced as an inconceivability argument based on logical possibility. Just as it is inconceivable for there to be something that is both a square and a circle, it is also inconceivable for there to be something that is a machine and exemplifies emotion *E*. The assumption is that just as there is a contradiction between something being a square and a circle, there is a contradiction between something being a machine and exemplifying emotion *E*. However, as I noted in my discussion of the logical use of 'can,' this is incorrect unless one fills out what other notions are in play where one finds a contradiction. My goal is to give a positive conceivability argument that shows how a machine can metaphysically exemplify emotion *E*. It is to be contrasted with a negative conceivability argument that aims merely to show that there is no logical contradiction in the statement *m* is a machine and *m* exemplifies emotion *E*.[8]

2.2 What is essential to an emotion? The argument from emotion regulation

The literature on what is essential to an emotion is both deep and wide. It is deep in the sense that it stretches back far. For example, in Western philosophy, it goes back at least to ancient Greek Philosophy.[9] And it is wide in the sense

[8] See Chalmers (2002) for discussion of the distinction between negative and positive conceivability.

[9] See Nussbaum (2001) and Shivola and Enberg-Pedersen (2010) for discussion of emotions in Ancient Greek philosophy.

that we find accounts of emotions in non-Western traditions, for example, in Chinese and Indian philosophy.[10] It is impossible to say with any certainty what is essential to an emotion in a way that won't be subject to responses from other accounts of emotions. However, this should not stop one from inquiring into and arguing about whether or not machines can have emotions in the way the question has now been defined: is it metaphysically possible for a machine to exemplify emotion *E* in virtue of exemplifying what is essential to *E*? I will now set out my main argument against the view that either phenomenology or physiology is essential to emotions. I will focus on the emotion of anger to be consistent with the setup of the question. After setting out the argument I will proceed to explain every premise before entertaining a number of objections followed by responses.

It should be noted from the outset that the position I defend on the possibility of machines having emotions is closely related to Cappelen and Dever (*forthcoming*). However, there are important differences. First, Cappelen and Dever defend the possibility of emotional life for LLMs; my target is not focused on LLMs because it is not focused on the typical architecture of LLMs, statistical text prediction. Second, Cappelen and Dever appeal to a cognitivist account of emotions found in the stoic tradition and defended by Nussbaum (2001). While my view is consistent with the cognitivist approach, my view is not focused on emotions as a kind, but is about specific emotions that have a cognitivist profile to them. For example, while I do think machines can have anger, I don't think that they can have rage because rage is essentially connected to a phenomenal state that anger does not require. Third on my approach artificial general intelligence, AGI, and access consciousness are important. However, on their approach it is not. Fourth, Cappelen and Dever want to set aside the question of what is essential to an emotion. Their approach to attributing emotions to LLMs, such as ChatGPT, is different from mine. In general they take a much more holistic approach to attributing mental states of all kinds, including beliefs, understanding, and agency, to LLMs. My approach is more focused on specific mental states. Fifth, the main argument I appeal to for defending the cognitive view of anger as opposed to the phenomenal + physiological view is not an argument from subtle variation of physiology or absent phenomenology as we find in Nussbaum. Nussbaum says:

Beliefs of the right kind are central to emotions:

> In order to have anger, I must have an even more complex set of beliefs: that there has been some damage to me or to something or someone close to me; that the damage is not trivial but significant; that it was done by someone; that it was done willingly; that it would be right for the perpetrator of the damage to be punished. It is plausible to assume that each element of this set of beliefs is necessary in order for anger to be present: if I should discover that not x but y had done the damage, or that it was not done willingly, or that it was not serious, we could expect my anger to modify itself accordingly or recede. (Nussbaum 2004, p.190)

There is a possibility of irrelevant subtle physiological variations:

> There usually will be bodily sensations and changes involved in grieving, but if we discovered that my blood pressure was quite low during this whole episode, or that my pulse rate never went above sixty, there would not, I think, be the slightest reason to conclude that I was not grieving. If my hands and feet were cold or warm, sweaty or dry, again this would be of no criterial value. (Nussbaum 2004, p. 195)

There is a possibility of absent phenomenology:

> there are feelings without rich intentionality or cognitive content—for instance, feelings of fatigue, of extra energy. As with bodily states, they may accompany emotion or they may not—but they are not necessary for it. (In my own case, feelings of crushing fatigue alternated in a bewildering way with periods when I felt preternaturally wide awake and active; but it seemed wrong to say that either of these was a necessary condition of my grief.) (Capellen et al. al., 2024, p. 152)

Cappelen and Dever are correct in my view to point to the cognitivist tradition in Western philosophy as an easy way to defend the view that machines can have emotions. They cite a number of figures who defend the view: Pitcher (1965), Roseman (1984), Solomon (1993), and Lyons (1980). They argue for emotions in machines by.

> [Relying] on well-established theories of what it is to have an emotion. Specifically, there are leading theories according to which the capacities we have already established that ChatGPT has, would suffice for the presence of emotions. According to cognitivists about emotions, they are to be understood as certain kinds of belief clusters. Nothing, we'll argue, prevents ChatGPT from having those kinds of beliefs. (Capellen et al. al., 2024, 150 in *unpublished manuscript*)

Here, it is, again, important to distinguish my approach from theirs. Consider the difference between a *conditions-of-satisfaction approach* and a *conditions-of-ascription approach.* Cappelen and Dever avoid a

[10] See Virág (2017) for a presentation of emotions in early Chinese philosophy. See Bilimoria and Wenta (2015) & Heim, Ram-Prasad, and Tzohar (2021) for a discussion of emotions in Indian philosophy.

conditions-of-satisfaction approach under which determining whether or not a machine can have emotions requires determining whether or not it satisfies certain conditions that are constitutive of having an emotion. Rather, they favor a conditions-of-ascription approach on which we should focus on the circumstances in which we are willing to ascribe emotions to an entity. They like Turing-style tests for when we would ascribe emotions.

My approach is squarely in the conditions-of-satisfaction approach. I think focusing on conditions-of-ascription takes us back to an approach where there could be a gap between *what a machines exemplifies* and *what a machine instantiates*. For example, passing any number of Turing-like tests would not be sufficient on a conditions-of-satisfaction approach even if conditions-of-ascription have been satisfied. While my approach shares features in common with Nussbaum, who Cappelen and Dever appeal to, my argument deploys variation across physiology and phenomenology due to practices of emotion regulation. I take inspiration for this approach not from Stoicism as Nussbaum does, but from Buddhism and Yoga philosophy—even if the position on emotions is not a Buddhist or Yogic one. Here is the main argument I will develop and defend:

2.3 The argument from emotion regulation

1. In the case of anger, either the phenomenology, physiology, or cognition associated with anger is essential alone or in some combination with the other aspects.
2. If feeling is essential, then it is impossible to have anger and regulate the phenomenology.
3. If physiology is essential, then it is impossible to have anger and regulate the physiology.
4. It is possible to have anger and regulate phenomenology and physiology in various ways.
5. So, neither the phenomenological nor physiological aspect of anger are essential to it.

Premise 1 highlights the fact that there are two cases. Either one of the three components alone is essential, but the other two are not, or any two or three of the components are essential in combination. For example, perhaps phenomenology and physiology are essential together, but not independently. While premise 1 might not seem important, it is important because some might argue that no single aspect is essential, rather all three are or some pair is essential. My argument strategy will aim to show that if phenomenology and physiology are not individually essential in any sense, then they cannot be essential in some combination. That is, if physiology is not essential alone, then physiology and phenomenology in combination cannot be essential. If F is not essential to Ks, then F and G cannot be essential to Ks.

Premise 2 is the heart of the main argument from emotion regulation. The basic phenomenon that is relevant is the following. It is possible to become angry at someone and then through techniques of emotion regulation change both the phenomenology and physiology associated with the cognitive judgment. More formally,

1. Consider anger.
2. Through techniques of emotion regulation, such as meditation (as in Buddhism) and breathing exercises (as in Yoga), one can change the phenomenology and physiological characteristics that co-occur with the onset of a given episode of anger without ceasing to have anger.
3. If an aspect A of an emotion E is essential to E, then E cannot occur without A.
4. Both the phenomenological and the physiological characteristics associated with anger can be changed through emotion regulation without ceasing to have anger.
5. So, E is not essentially either the phenomenological or physiological characteristics associated with it.

It is well understood that emotions can be managed, and that the management of emotions involves emotion regulation.[11] However, as I will discuss below, it is controversial as to whether emotion regulation involves changing an emotion or only regulating inessential features of the emotion. One component of "emotion regulation" involves changing how it feels to undergo the emotion at a time, and that in turn can be taken to involve changing the physiological features that are present when an emotion is exemplified in a person. What the argument here aims to do is show that some of the features associated with an emotion, such as the phenomenology and the physiology can change while a person is undergoing one and the same emotion because those features are being regulated through meditative or breathing techniques. These techniques often lead, in the case of anger, to the dissipation of phenomenology as well as physiological features such as tension in the body. This in turn is argued to show that phenomenology and physiology are not essential to the emotion of anger. In order to best understand this argument it is necessary to consider several objections to it, and responses to them. In each case the responses do not intend to be decisive, but to offer a reasonable response.

[11] See Gross (2007) for discussion of the phenomenon of emotion regulation.

2.4 Objection 1: the many emotions objection (or are we even regulating as opposed to changing)

One way to block the argument in favor of phenomenology and physiology being inessential to an emotion, such as anger, via emotion regulation is to simply say that there are many sub-types of anger. Simply put, there is no emotion regulation only changing emotion. There are two versions of how this objection can be presented.

According to the emotion regulation argument the reason why the phenomenology and physiology associated with anger are not essential to it is because one can be angry while regulating down the phenomenology (how it feels) and the physiology (the tension in their body). According to the first version of the many emotions objection, to the regulation argument is that it will not successfully show that phenomenology and physiology are not essential. The general model of why the argument cannot show that is as follows. Where 'A' stands for anger, 'PH' stands for phenomenology, and 'PY' stands for physiology, we can illustrate the essential dependence as follows:

A1-PH1-PY1 — A2-PH2-PY2 — A3-PH3-PY3

The general idea is that as one is regulating their anger it is not the same kind of anger that stays the same while the phenomenology and physiology change. That is, it is not the case that the mental state is one of anger where the phenomenology and physiology are contingent because they are changing through regulation. Rather, while the phenomenology and the physiology are changing, so is the anger. Rather than having just anger, we have anger 1, anger 2, and anger 3 where each anger is essentially tied to the phenomenological and physiological changes. On this model all three components are essential because any time any one of them changes the whole emotion changes. It isn't the case that we experience one and the same anger over time because the judgment is the same but the phenomenology and physiology are changing. Rather, as any one aspect changes the whole complex, which is the anger, changes.

According to the second version of the objection, some type of coarse-grained feeling must be present for anger to be present. However, it is possible to down-regulate the fine-grained feeling without getting rid of the coarse-grained feeling. On this version of the objection, the main point is that some feeling is always present at the coarse-grained level, all that we regulate is the fine-grained anger. So, feeling is essential.

2.5 Response to 1: over-generation fallacy

The main response to the first version of the many emotions objection is that it over-generates emotions. The truth behind the objection is that there are different grades of anger, and we do note that by drawing a distinction between irritation, on one hand, and rage on the other. However, to accept that while the judgment stays the same and moment by moment some subtle phenomenological or physiological component changes would lead to a vast over-generation of types of anger. There are two reasons for this.

First, within a subject it is possible that while the judgment stays the same, the physiological changes due to emotion regulation do not map one to one with phenomenological changes. For example, a person can hold in mind that they are angry at another because they wronged them and regulating down, through breathing, their heart rate. However, for every drop in heart rate there is no corresponding difference in feeling. According to the many emotions view when this happens there is still a new emotion. Even if one holds the same judgment and feels the same, the slightest drop in heart rate means that they have a new form of anger. This is not theoretically viable because it leads to the over-generation of types of anger because a new anger is present even when the subject cannot detect a difference either in the judgment or the phenomenology. New anger arises just because one's physiology changes while all else remains the same. Second, across subjects there can be variation in how anger is regulated through breathing. For example, we can imagine that Susan and Harry can regulate their anger in roughly the same way. When they get angry they can breathe and regulate down their heart rate at a certain speed and to a certain degree while their phenomenology follows. But suppose Carlos is different, and he can hold in mind a certain judgment but due to breath practice he can regulate down his anger quicker and he can do so too a much greater degree. Almost to the degree where he feels nothing, but still holds in mind the judgment that the person he is angry at has harmed him. In this case we will generate distinct types of anger across subjects. There will be anger that Carlos can feel, that neither Susan nor Harry can feel because we have made phenomenology essential to anger.

The value of the many emotions objection, as I noted, is that it does make sense to say that there are different kinds of emotions. But to now tie phenomenology and physiology to the emotion in order to account for that leads to a vast over-generation of emotions that is theoretically a violation of Occam's razor.

The response to the second version of the many emotions objection requires considering the real objection it relies on.

2.6 Objection 2: the impossibility of zero phenomenology and physiology emotions

The view that no specific phenomenology or physiology is essential to a given type of emotion leads to a simple problem. It leads to the view that it is possible to be angry, for example, without feeling anything at all that we typically associate with anger. And it leads to the view that one can be angry while

one's body is completely calm and not in any state of tension. This appears to be an absurd conclusion. To be angry requires feeling a sensation that is not associated with love or fear or melancholy, for example. Thus, when no specific phenomenology or physiology is essential to an emotion we are led to the confusing result that one can be angry, for example, without having any *angry feelings*. In addition, the impossibility of zero phenomenology and physiology emotions is what supports the second version of the many emotions objection by supporting the view that a person can regulate the fine-grained physiology and phenomenology of an emotion without eliminating the coarse-grained phenomenology and physiology.

2.7 Response to 2: zero phenomenology and physiology emotions are a feature not a bug

If we consider the phenomenon of emotion regulation to be wide and real in the sense that some people can regulate phenomenology and physiology wider than others, then the result that one can be angry and not feel anything or be completely not tense in one's body is a sound result. As already noted, Nussbaum appears to endorse this view on her cognitivist account. The difficulty that is pushed in this objection is the fact that one wants to say that when the phenomenology of anger is gone and the physiology or tension in the body is gone, so is the judgment. That is, there is a causal relation. In regulating down the physiology and the phenomenology the judgment should then be caused to also disappear. But there is no reason to accept this, when one realizes that a person can attend directly to the judgment while regulating down the phenomenology and physiology.

That is, one can remain justifiably angry at someone because they really harmed them, but then learn through years of breathing exercises and meditation and training in emotion regulation how to get rid of the phenomenology and physiology associated with anger. And there can be good reason for one to want to learn this. Often times the phenomenology and physiology associated with anger gets in the way of performing in a context. By learning to down-regulate the phenomenology and physiology associated with anger one is able to perform in a context where their anger is triggered. The important thing to remember is that when the person down-regulates they don't lose their judgment that a person harmed them. In fact, later one can even choose to act on their judgment once they are in a different situation. If in fact the judgment was taken away by the down-regulation they would lack a reason to act.

2.8 Objection 3: physiological generation conditions

Suppose that emotion E is typically caused in humans by physiological process P as part of its causal profile. That is, P plays a causal role in the production of anger. One might infer from this that P is essential to E. The basic idea is that if anger is caused by a physiological process in us, then that physical cause is essential to anger.

2.9 Response to 3: generation-individuation fallacy

There is a distinction between what causes something to occur and what is essential to it for the purposes of individuation. Even if anger in humans is caused by physiological process P it doesn't follow that P is essential to anger.

Recall that the manner in which we are discussing the essential features of an emotion has to do with metaphysical modality. Is it possible in the metaphysical sense for something to be angry yet not have physiological process P be the cause? One way to see how this can happen is through an analogy with spectrum inversion. In spectrum inversion cases one person sees red from a typical cause c, another person sees green, because the two are spectrum inverted. In emotion inversion cases, one gets angry from typical cause c, another person gets sad because the two are emotion inverts. If spectrum inversion is conceivable, then emotion inversion is conceivable. Thus, if the former is metaphysically possible on the basis of conceivability, so is the latter.

2.10 Objection 4: recalcitrant emotions

Suppose I feel afraid of falling from a bridge but judge that my situation is not fearsome (because I have reason to believe that the bridge is perfectly safe).

In this case I have what is called a *recalcitrant emotion*: I remain afraid even though I judge that there is nothing fearsome about my situation. If emotions, such as fear, are simply judgments, then my fear of falling is the judgment that falling is fearsome, and it follows that (i) I have two inconsistent judgments and (ii) I am presumably irrational because of my inconsistent judgments. Thus, there is a *prima facie* problem for the view that emotions are essentially cognitive judgments.

2.11 Response to 4: there need not be recalcitrant emotions

The so-called phenomenon of recalcitrant emotions aims to put pressure on accounts of emotion that take them to be essentially cognitive judgments by showing that if they are cognitive judgments then it will be the case that there is an

inconsistency when one has a recalcitrant emotion. However, there is more than one way to analyze the situation.

If I judge that there is nothing to fear, then what I experience is simply the feeling of fear or I experience a fear-like feeling. But it need not be the case that I have the emotion of fear. There is no reason why one cannot analyze the phenomenon of recalcitrant emotions simply by denying that there are two emotions present that are contradictory. Rather, one can simply point out that as long as one has the judgment that there is nothing to fear, they can simply have feelings that are typically associated with the emotion of fear without having the emotion. This analysis carries over to anger.

I can judge that I am no longer angry at a person, and still have feelings that typically co-occur with the emotion of anger. The reason why is because I can rationally decide that my judgment that another person wronged me was based on a false belief. Now that I know that the belief is false, I have no reason to be angry. As a consequence, I retract the cognitive judgment that is essential to the emotion of anger. Once I have retracted that judgment I no longer have the emotion of anger. Nevertheless, I can have feelings that typically co-occur with anger because their dissipation requires a greater amount of time than the time required to have a rational change in view about whether or not the person I was angry at really deserves my anger.

Part II: Is the capacity for emotions tied to consciousness?

It is hard to explore the question of whether machines can have emotions without taking on the question: in what sense, if any, can machines be conscious? The naïve argument against emotions that I laid out in **2** actually rests on the relation between feeling and consciousness, which is left suppressed in the naïve argument. It is actually made clear by using James's view of emotions as feelings and a premise connecting feelings with consciousness. I call the argument that adds in these connections: the mature argument.

3 *The mature argument*:

1. Emotions are essentially tied to feelings. It is essential to anything that has an emotion that it has the capacity to feel.
2. If a creature can have feelings, then it must have the capacity for consciousness.
3. Machines lack the capacity for consciousness.
4. So, machines cannot have feelings.
5. So, machines cannot have emotions.

However, in the last section I defended the view that emotions, such as anger, are not essentially tied to a specific phenomenology or physiology. That phenomenology and physiology are present in us when we have emotions doesn't make them essential to emotions as a mental state type that is multiply realizable across various kinds of beings. As a consequence, it will be useful to clarify the mature argument and then explore different kinds of consciousness and how they might be related to the possibility of emotions in machines. Following Block (1995) it is common to distinguish between two types of consciousness.

Phenomenal consciousness is the familiar type of consciousness that goes along with the phrase, coined by Nagel (1974), of experience having a *what it is like aspect* for a subject. There is something that it is like to see red vs. blue, taste tea vs. coffee, hear C major vs. F major, smell jasmine vs. rose, and touch sandpaper vs. velvet. P-consciousness, for short, captures the what it is like aspect of experience. When an emotion theorists talks about the phenomenology or feeling associated with anger they are talking about phenomenal consciousness. And they are saying that phenomenal consciousness, which is where feelings are to be found, is essential to having an emotion. Why? Because all affective consciousness (emotional consciousness) depends on phenomenal consciousness (what it is like consciousness). If you don't know what anger feels like, you cannot have anger.

Access consciousness is not familiar outside of philosophy and psychology. It is an idea introduced by Block, when he explores the phenomenon of blindsight. In blindsight, a person reports that they cannot "see" for example out of their left eye, yet when an experimenter asks the person to pick up what is in front of them or to navigate down a hallway, they are able to do so. A-consciousness, for short, captures the way in which a state is poised for rational control of action and speech and inferentially available. The actual definition of A-consciousness is controversial. One way to draw the distinction at a high level of generality is to say that P-consciousness goes with phenomenology, and access consciousness goes along with capacity. A representation of a content is A-conscious when a subject can do something with it, when the representation is poised for use in some way. The central issue that Block raises in his paper is whether it is possible for A-consciousness to come apart from P-consciousness. There are two directions here. I will only be looking at the direction related to the possibility of machines having emotions – Why? Because all affective consciousness (emotional consciousness) depends on phenomenal consciousness (what it is like consciousness). If you don't know what anger feels like, you cannot have anger.

The possibility of A without P-consciousness is supposed to be supported by the case of blindsight. A person with blindsight can navigate down a hallway using only their "blind" eye, where they report not being able to "see" anything. Their ability to navigate suggests that representations of the hallway and obstacles in it are accessible for use in the rational control of action, even though there is no

phenomenology. A blindsight subject will say that they cannot "see" anything. Yet they are doing things that typically require and depend on them being able to "see."

In the mature argument the kind of consciousness that is tied to feelings is phenomenal consciousness. So, if A-consciousness without P-consciousness is possible, then perhaps machines can have an emotion, *E*, when *E* is tied only to a cognitive judgment.

3.1 Objection 1: from seeming states

However, there is an important objection to the possibility of A without P-consciousness. It is the objection from the dependency on seeming states.

1. A-consciousness for a being depends on its external environment *seeming to it a certain way*.
2. *Seeming a certain way* to a being depends on P-consciousness. It is a phenomenal notion.
3. So, A-consciousness depends on P-consciousness.
4. So, it is impossible for there to be A-consciousness without P-consciousness.

3.2 Response 1: probability and seeming

There is a way of making this objection stick in the case of blindsight. In the setup of one of the experiments the sighted right eye of the subject is closed, and the blind left eye is left open. The subject is asked, "can you see anything in front of you through your left eye?" The subject reports back that they do not see anything. Now if the subject is then told to, "pick up the pen in front of you." The subject often responds by saying, "I don't see anything there to pick up." However, if the experimenter continues to prompt the subject by saying, "just try and pick up anything that is in front of you." It turns out that the subject reliably picks up the pen in front of them. The problem is that without the experimenter prompting the subject that the *world is a certain way such that there is something in front of them* the subject is not inclined to do anything. The argument above crystalizes this point as an objection by attempting to draw the result that A-consciousness needs P-consciousness in order for any action to actually occur. If the world doesn't seem a certain way, nothing is going to happen. There are two points that can be used to reduce the force of this objection.

First, the fact that in humans and other non-human animals it is the case that phenomenal consciousness plays the role of presenting the world to us in a way that makes it actionable fails to support the conclusion that phenomenal consciousness is necessary for the world to be actionable to some being. P-consciousness is a path to representing the world, and it is the one we use, but it might not be necessary.

Second, and related to the first, there is a clear way in which the world can be represented to a creature or system that lacks phenomenal consciousness. A self-driving car has representations of its environment that allow it to navigate from place to place. The representations are another way in which the world can be presented to a creature or system. It is a way of presenting the world where phenomenology is absent because the world is represented conceptually and categorically as opposed to phenomenologically. In addition, the world can be represented in a non-phenomenal way in terms of probabilities. For example, in a self-driving car a person in a crosswalk can be represented non-phenomenally through the probability that a sensory input falls under a category —a category that might be tied to a halt function. There is nothing it is like for the self-driving car to receive a probability distribution that presents the world such that there is figure in a crosswalk. Nevertheless, it can act on the basis of that sensory input.

Thus, while it is true in us that A-consciousness might depend on P-consciousness because P-consciousness is how we represent the world or have the capacity to represent the world, it does not follow that the world seeming some way depends on phenomenal consciousness across different creatures or systems.

Let's take stock of where we are at. I am arguing for the conclusion that machines can have emotions. And my argument for that conclusion so-far has depended on a defense of two claims.

(a) Some emotions, for example anger, are not essentially tied to a specific phenomenology or physiology.
(b) There are two kinds of consciousness, P-consciousness and A-consciousness.

Given that I am dispensing with the claims that makes up the naïve argument in its mature form –emotions are tied essentially to feelings and feelings require phenomenal consciousness, I will simply move on to the following questions:

What does it take for a machine to have access consciousness?

Is it possible for a machine that has access consciousness to make judgments that are of the right complexity and kind for having an emotion, such as anger?

4 The path to machine emotions

Let me set the path forward to the conclusion that machines can have emotions by simply offering the argument for the conclusion. I will then explain the premises before turning to a series of objections followed by responses to them.

1. Machines will eventually have artificial general intelligence.
2. Artificial (or natural) general intelligence is a benchmark for access consciousness.
3. Access consciousness is tied to the ability to make judgments, reason from them, and act rationally in relation to them.
4. Some emotions are simply judgments at their core, the phenomenology and physiology associated with them are not essential.
5. So, machines can have emotions.

Premise 1: Current machines demonstrate domain specific intelligence that surpasses that of humans, for example, by beating the best human chess or go players, or solving protein folding problems. However, no machine has the capability to hold a conversation, and fold clothes at the same time while dancing. Many humans can do this. In principle, there is no barrier to move from domain specific to general artificial intelligence in machines. So, it is likely that artificial general intelligence in machines is on its way.[12]

Premise 2: Mindt and Montemayor (2020) have defended the view that there is a connection between general intelligence and access consciousness. Because they are focusing in their work on LLMs, they focus on the case of artificial general intelligence, and not natural general intelligence. They argue that a necessary condition on access consciousness in LLMs is the capacity for artificial general intelligence. Their argument does not entail that artificial general intelligence is required in non-human animals for access consciousness. Their main argument is focused on the space of artificial intelligence, where we need to locate a benchmark for access consciousness, just as we look for benchmarks or markers for phenomenal consciousness. Their argument is the following.

1. General intelligence, GI, yields differential attention.
2. Differential attention distribution is necessary for access consciousness because attention is a kind of action that prepares the agent to access and use information, which is essential to access consciousness.
3. So, access consciousness requires GI, in machines AGI.

It is important to clarify this argument by pointing out that AGI is not inconsistent with modularity. Some AI researchers hold the following:

A. Minds are massively modular.
B. AGI is inconsistent with massive modularity.
C. So, machine minds that are sufficiently similar to human minds with human-level intelligence will not have AGI.

However, the argument has three problems.

First, it is very unlikely that human minds are completely modular the way massive modularity proposes. Fodor (2001) persuasively shows that a mind can be modular to a large degree without being completely modular.

Second, AGI is consistent with levels of modularity, where lower modular systems feed more general, but still modular systems. If there is no massive complete modularity, then AGI is consistent with levels of modularity in a hierarchy.

Third, a machine can achieve human level intelligence in one domain without achieving it in all domains.

Clarification: It is important to note that this argument is consistent with modularity. Some researchers hold that because minds are massively modular, and AGI is inconsistent with massive modularity, machines minds that are sufficiently similar to human minds with human-level intelligence will not have AGI. There are three things wrong with this argument. First, it is very unlikely that human minds are completely modular. In, "The Mind Does not Work that Way" Fodor (2001) effectively argues against the massive modularity hypothesis of Stephen Pinkers' (1999), "How the Mind Works." Second, AGI is consistent with levels of modularity where lower modular systems feed more general, but still modular systems. Third, a machine can achieve human-level intelligence in some domains without achieving it in all domains.

Premise 3: Simply note the definition of access consciousness offered by Block (1995) as well as others who have built on his seminal work.

Premise 4: Recall that the view on offer here is *that some emotions are essentially judgments and not feelings. The view is not that all feeling states are essentially judgments*. Some feeling states are essentially feeling states, for example, sympathy, empathy, and compassion. Since it is judgments and not feelings that are essential to some emotions, if a machine can make the right kind of judgment, then the machine can have some emotions.

[12] For evidence of improvements toward AGI, see https://www.livescience.com/technology/robotics/ai-powered-humanoid-robot-figure-01-can-serve-you-food-stack-the-dishes-and-have-a-conversation-with-you?utm_medium=social&utm_campaign=socialflow&utm_content=livescience&utm_source=facebook.com&fbclid=IwAR3TMhW4E4kQXE6_p3Kea11_yVkur1GCX4DI-3OtU71H6wR4zFFF2f_P4UU

In order to better understand this argument, let us consider some objections to it.

4.1 Objection from understanding and meaning

The argument depends on the question of whether machines can really make judgments? However, there are reasons for thinking that machines cannot make judgments. One simple way to defend this point is by assuming that making a judgment requires intelligence and understanding a language. However, given Searle's (1980) Chinese Room Thought Experiment, there are reasons to doubt that an LLM, for example, understands the language in which its outputs to queries are expressed. Since Searle's argument focuses on GOFAI (good old fashioned AI) programming, and LLMs are stochastic text prediction machines, the point of the Chinese Room argument has to be altered. Here is an altered version.

1. Machines only engage in symbol manipulation or stochastic text prediction when they yield an output to a query.
2. Neither symbol manipulation nor stochastic text prediction is sufficient for understanding meaning or meaning something via an output representation.
3. Understanding meaning, and meaning something via an output representation are necessary for judgment.
4. So, machines cannot make judgments.

What is salient about the Chinese Room is that Searle himself does not understand Chinese, yet by following a simple look-up table that converts input symbols he does not understand to output symbols he does not understand he is able to answer all queries correctly. Thus, the argument is since he does not understand, the machine which he is a part of does not understand, since no other part could yield understanding. For example, the look-up table cannot yield understanding, if Searle doesn't already understand.

One issue in Searle's example concerns symbol grounding. Searle has no idea of what any symbol is related to non-symbolically. He does not know, for example, that a given symbol, 'cat' is correlated with CATS.

One might think that Searle's argument can easily be extended to LLMs because the only difference is that LLMs use stochastic text prediction and not GOAFI. However, there is a slight issue. In Searle's thought experiment there is the problem of symbol grounding in addition to the claim that Searle doesn't understand Chinese. However, depending on how an LLM is trained there is symbol grounding in an LLM because semantic items are grounded in images. The question is whether or not the symbol grounding in LLMs yields any genuine understanding. One could argue that because the neural net does not understand any of the data it crawls over, the fact that the data have certain connections, the LLM cannot be said to understand. Connecting 'cat' to cats won't help if the LLM doesn't know what cats are.

Hattiangadi (*forthcoming*) develops a compelling argument against LLMs understanding the natural languages in which their outputs occur. Her argument goes beyond Searle's thought experiment, and it engages an important issue. She uses a more vivid case than that of Searle's Chinese Room and the look-up table. She engages Blockhead from Block (1981) to show the lack of intelligence and understanding in LLMs.

Before we consider the thought experiment from Block and her use of it, let us develop one of her main considerations against intelligence and understanding in LLMs:

1. Understanding a natural language and intelligent use of a natural language requires satisfying the productivity property of natural languages. Productivity is the capacity to understand and produce indefinitely many novel sentences one has never encountered before. Productivity, in a sense, is a "superficial" property of a system. This is explained by a hypothesis about the architecture that gives rise to productivity. That architecture involves ascribing to a system that understands a language, understanding of the meanings of the expressions in a finite lexicon and a set of compositional syntactic-semantic rules for the construction of complex expressions.
2. LLMs have no capacity for satisfying productivity because in order to satisfy the constraints on cognitive architecture necessary for compositional understanding of a natural language, a system must have a causal-explanatory structure that mirrors the syntactic structure of the language. Since stochastic text prediction does not involve an architecture that mirrors the compositional structure of the language, systems that engage in stochastic text prediction do not satisfy the compositionality requirement on cognitive architecture.
3. So, LLMs don't understand the natural languages that their output strings occur in, and do not intelligently use them.

From this aspect of her work we can then add in (4) to get to a relevant negative result with respect to LLMs possessing understanding.

4. Compositionality requirements are a necessary condition on intelligent and competent judgment.
5. So, LLMs don't make judgments.

Now consider her thoughts on Block's thought experiment.

> [It] seems to me that the failure of productivity constitutes one good reason to think that Blockhead, the affectionately named machine described in Block's (1981) thought experiment, is unintelligent, and does not understand natural language. Blockhead has been programmed to give responses to a set of questions that mimic the responses that would be given by Block's Aunt Bertha. The programmers encode in Blockhead's inner workings strings of questions and answers in English, arranged in a tree structure. Given the input A provided by the interrogator, Blockhead searches through the pre-programmed sensible responses to A, and randomly selects B as its output. The interrogator then inputs C, and Blockhead searches through all of the strings which start with ABC in its initial sequence and then types out a fourth sensible sentence, and so forth (Block 1981, 20). Intuitively, Blockhead neither understands English, nor displays intelligence. There may be several good reasons for thinking this. At least one of them is that the whole pathway of questions and answers have been in a sense memorized in advance. So, Blockhead displays no productivity at all, since no sentence that is either input or output is genuinely new. (Hattiangadi 2024, p. 5-6)

4.2 Response to understanding

What does the capacity to understand meaning or simply to bestow meaning to a representation require?

First, it isn't clear that we need to prove that machines satisfy productivity and thus compositional architecture the way we do. The setup of the objection is fallacious.

1. We satisfy productivity and compositional architecture via X.
2. LLMs cannot do X.
3. So, LLMs cannot satisfy productivity and compositional architecture.

One response to this kind of argument is simply to say that different kinds of creatures or systems count as understanding even if they do it in different ways. That is, understanding is multiply realizable not only across different material substrates but also across different functional relations. Hattiangadi, for example, is arguing that there is no understanding where there is no productivity and compositional architecture, or at least, that there is a lack of a certain kind of intelligence when productivity, and thus compositional architecture, is missing.

Another response to this kind of argument is to simply concede that different kinds of creatures and systems *do not* count as understanding because they do not do what we do, but nevertheless, they count as understanding*, where there is a similarity relation between understanding and understanding*.

Both of these responses blunt the force of Searle's and Hattiangadi's arguments. Searle actions of responding to a query through a look-up table constitute either a way of understanding – note he gets all the answers correct – or Searle understands*. Likewise Blockhead understands in a way that we don't or he understands*.

In addition, we should consider the interaction between understanding and intelligence. My view is that it is much easier to concede that neither Searle nor Blockhead are intelligent for the very reason that Hattiangadi argues for, lack of productivity as a capacity. But the fact that something doesn't have productivity doesn't mean it doesn't understand in any sense; it shows that it isn't that intelligent. An adult human learning a second or third language late in life doesn't learn the same way a young child learns multiple languages in a bilingual home. In the case of the young child we might see lots of productivity as they try out new strings, but we might also say they don't understand what they say. In the case of the adult we might see less productivity and more understanding because they know what they want to say and what words are used to communicate that. The child is just trying out stings, which is evidenced by the frequent nonsense present in children, but absent in adults.

Following this point, we can note that the link between intelligence and judgment can be broken. Does the child that is trying out new strings make judgments? I would say yes because the utterance 'Fido sits' while staring at the family dog involves subsumption of an individual under a property – a hallmark of judgment. However, does the child understand? Perhaps not, if they are just trying out strings to get feedback from parents. How does this bear on understanding?

Perhaps, someone can judge only with a kind of understanding that is distinct from what we find in normal adult first-language use. This suggests that there are different ways of understanding or different kinds of understanding, such as understanding*.

Second, the argument against understanding cannot require phenomenal consciousness at this stage of argument. For if understanding meaning requires phenomenal consciousness than access consciousness without phenomenal consciousness cannot be a type of consciousness that involves the ability to report on one's own states, which requires making judgments, and is definitional of what access consciousness involves as a capacity. So, I will assume that understanding a representation do not require the capacity for phenomenal consciousness. To see this point, let's consider Jackson's (1986) thought experiment about Mary in relation to understanding as opposed to phenomenal consciousness. Reconstructing the thought experiment this way leads to the view that

understanding is open to an incompleteness phenomenon where it is appropriate to say that someone understands even if they do so incompletely or only to some degree. Assume, as Jackson sets up in his thought experiment, that Mary does not know what it is like to see red because she has never experienced red through phenomenal consciousness.

Recall Mary is sighted but raised in a black and white environment with no exposure to color experience. Through reading and taking classes with teachers, she learns that when someone asks her, 'what color is the sky on a non-cloudy day?' she should answer 'blue,' while if they ask her 'What color is a cloud on a non-rainy day?' She should answer 'white.' Given that she has never seen these colors, does she mean or understand what she is saying, when she uses color terms, such as 'blue' and 'white?' My intuition is that she does because understanding is open to incompleteness. Just as one can understand what is meant by using the word 'two' for the numeral 2 without knowing every aspect of the latter, such as that it is the square root of 4, one can understand the word 'blue' for blueness without understanding every aspect of it, even that it essentially looks a certain way. If we switch Jackson's case to the one Russell discussed, where a person is blind, we could say a blind person understands color words incompletely because they do not know what qualities they pick out even though they know many facts about the qualities picked out by color words.

Likewise, one can argue that Searle and Blockhead incompletely understand because their understanding is dependent on a look-up table, and it doesn't facilitate knowing how to use Chinese words to say something independently of a query in relation to a look-up table, as in Searle's case. More importantly, it does not allow for answers to questions outside of the memorized decision tree in Blockhead's case. But, as I just argued, this suggest that neither are intelligent. It does not suggest that they have no understanding. Their understanding is limited and incomplete, just as someone who understands geometry but not trigonometry could be said to have an incomplete understanding of the mathematics of triangles.

4.3 Response to meaning

Finally, putting aside the case of understanding, how about the case of bestowing meaning upon a representation? Concerning this question we come into contact with Searle's distinction between original and derived intentionality. Consider the following argument.

1. Machines only have derived intentionality because they are programmed.
2. Only things with original intentionality can mean anything via a representation, such as speech or symbolization.
3. Therefore, machines cannot mean anything via speech output or symbol output.

The response to this objection comes from noticing that derived intentionality is a way of having intentionality not a way of being robbed of intentionality. Searle portrays the distinction as if original intentionality is the only form of genuine intentionality. However, derived intentionality is genuine when we see that it is a way of bestowing meaning. One can say, "I mean by 'Joan of Ark,' what my teacher meant when they taught me it in class." And this chain can be traced back all the way to the baptism of the name. What this shows is that the speaker need not be engaged in any act of original intentionality in order to mean something. It can simply be derived or parasitic intentionality. The main point that Searle is trying to make is that the system cannot mean anything through its outputs if it has no original intentionality. It sounds a bit like Lady Lovelace / Charles Babbage's analytic engine objection. Machines cannot mean anything because everything is programmed in. There is no freedom involved in the response it makes, and freedom is required for meaning. But this is at best controversial. Why should the absence of freedom block us from meaning?

4.4 Objection from truth

Yet another argument that suggests that machines cannot make judgments is tied to truth as opposed to meaning or understanding. Although Cappelen and Dever take up the issue of truth, they do not discuss the argument below. *Assertoric force argument.*

1. Making a non-internal judgment requires putting something forward as true with assertoric force as a speech act.
2. Putting something forward as true via a speech act requires having a conception of truth, and understanding norms of truth.
3. Machines do not have a conception of truth, or understand norms of truth.
4. So, the outputs of a machine are not judgments.

Cappelen and Dever discuss whether an LLM needs a conception of truth. It could be the case that no requirement on truth is necessary for judgment. So, what support is there for (3). One reason to think that (3) is true is that LLMs hallucinate, and hallucination in LLMs has been associated with the fabrication and putting forward as fact something that is wholly made up with no foundation. One might even wonder, as Vaidya (2024) argues, whether LLMs are

bullshitters in the sense advocated by Frankfurt. To bullshit is not just to lie and deceive, but to have no concern for the truth.

In general,the argument above captures, what I call, the intellectualist position on making judgments because at premise (2) it requires that for any *x* to make the judgment that *j*, it must be the case that *x* has a conception of truth.

4.5 Response to objection from truth

I will concede that if intellectualism is true, the conclusion holds. But intellectualism about judgment is false.

First, note that it only requires a conception of truth on the part of the subject, but not the true conception of truth. Second, intellectual theories of judgment make it the case that young children and non-human animals don't make judgments because they lack a conception of truth. Third, a self-driving car, for example, can be said to make a "judgment" as to whether to apply the brakes, based on input fed to its algorithms from its sensory apparatus because it can either apply the breaks or not in the space of available actions. This is similar to what happens in the case of young children and animals. In all three cases an action space is available and a move in the action space can be made where the agent is taken to have been instrumental in the action taken.

5 Conclusion

It is best to state my conclusion as an argument based on conceivability that aims to counter the naïve argument against machines having emotions. As I said at the outset, the naïve argument maintains that (SQ) and (ME) are true rather than contradiction.

(SQ) Nothing can be both a square and have the properties of a circle at the same time.

(ME) Nothing can be both a machine and have the properties of an emotion at the same time.

The naïve argument aims to show that just as (SQ) is true (ME) is true, it does so by using the further claim in the mature argument that emotions depend on feelings and feelings depend on the capacity for phenomenal consciousness.

My position is that the naïve argument and the mature argument are both naïve. I have defended a positive conceivability argument that shows machines can have some emotions.

1. Some judgment involving emotions are correctly and exhaustively captured by the judgment component of the emotion as realized by a normal human, as opposed to the phenomenological or physiological components associated with the emotion, since those unlike the former vary across subjects and over time.
2. Access consciousness depends on artificial general intelligence.
3. Machines can achieve artificial general intelligence.
4. Judgment-involving emotions that are exhausted by the judgment component depend on access consciousness for their realization.
5. So, machines can have some emotions.

Clearly, one need only hold that all emotions depend on phenomenology by being feelings essentially to refute the argument above and to defend the naïve argument. The exercise of this inquiry has been to help establish how machines can have emotions.

Funding There are no funding sources.

Declarations

Conflict of interest There are no conflicts of interest or data reported.

References

Bilimoria P, Wenta A (2015) Emotions in Indian thought-systems. Routledge Press, London
Block N (1981) Psychologism and behaviorism. Philos Rev 90(1):5–43
Block N (1995) On a confusion about a function of consciousness. Behav Brain Sci 18(2):227–247
Cappelen H, Dever J (2024) A defense of artificial intelligences: an essay in inhuman philosophy. Cambridge University Press, Cambridge
Chalmers D (2002) Does conceivability entail possibility? In: Gendler TS, Hawthorne J (eds) Conceivability and Possibility. Oxford University Press, Oxford, pp 145–200
Clark A (2003) Natural born cyborgs: minds, technology, and the future of human intelligence. Oxford University Press, Oxford
Fodor J (1974) Special sciences (or the disunity of science as a working hypothesis). Synthese 28(2):97–115
Gross J (2007) Handbook of emotion regulation. Guilford Press
Hattiangadi A (2024) Why large language models don't understand natural language (and probably won't anytime soon). In: Cappelen H, Sterken R (eds) Communication with AI: Philosophical Perspectives. Oxford University Press, Oxford
Heim M, Ram-Prasad C, Tzohar R (2021) The Bloomsbury Research Handbook on emotions in classical Indian philosophy. Bloomsbury Publishing, London
Jackson F (1986) What mary didn't know. J Philos 83(5):291–295
James W (1884) What is an emotion? Mind 9:188–205
Kim J (1992) Multiple realizability and the metaphysics of reduction. Philos Phenomenol Res 52(1):1–26
Kripke S (1980) Naming and Necessity. Harvard University Press, Cambridge
Lyons W (1980) Emotion. Cambridge University Press, Cambridge
Mindt G, Montemayor C (2020) A roadmap for artificial general intelligence: intelligence, knowledge, and consciousness. Mind Matter 18(1):9–37
Nussbaum M (2001) Upheavals of Thought: The Intelligence of Emotions. Cambridge University Press, Cambridge

Nussbaum M (2004) Emotions as judgments of value and importance. In: Solomon RC (ed) Thinking about feeling: Contemporary philosophers on emotions. Oxford University Press, pp 183–199
Pitcher G (1965) Emotion. Mind 75:326–346
Roseman IJ (1984) Cognitive determinants of emotions: a structural theory. In: Shaver P (ed) Review of Personality and Social Psychology. Sage, Beverly Hills, pp 11–36
Searle J (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–457
Solomon RC (1993) The passions: Emotions and the meaning of life, 2nd edn. Indianapolis, Hackett
Turing A (1950) Computing machinery and intelligence. Mind 59:433–460
Virág C (2017) The Emotions in Early Chinese Philosophy. Oxford University Press, Oxford