Reasons to Respond to AI Emotional Expressions

Rodrigo Díaz [1,2] and Jonas Blatter [3]

[1] Institute of Philosophy, CSIC.

[2] Centre for Research in Ethics, University of Montreal.

[3] Institute for Philosophy I, Ruhr University Bochum.

(forthcoming in *American Philosophical Quarterly*, please cite the published version)

Abstract

Human emotional expressions can communicate the emotional state of the expresser, but they can also communicate appeals to perceivers. For example, sadness expressions such as crying request perceivers to aid and support, and anger expressions such as shouting urge perceivers to back off. Some contemporary artificial intelligence (AI) systems can mimic human emotional expressions in a (more or less) realistic way, and they are progressively being integrated into our daily lives. How should we respond to them? Do we have reasons to reply to the appeals made by AI emotional expressions? In this paper, we examine the conditions under which AI emotional expressions could give us prudential or even moral reasons to change our behavior. We argue that these conditions do not depend on whether the emotional expression is genuine or not, but rather on the presence of features some of which can be implemented in emotive AI given our current level of technological development. We extract recommendations and warnings for the development of emotive AI.

Keywords

Artificial Intelligence; Emotion; Emotion Expression; Practical Reasons; Moral Reasons

1. Introduction

Some contemporary artificial intelligence (AI) systems can mimic human emotional expressions in a (more or less) realistic way, and they are progressively being integrated into our daily lives. How should we respond to AI emotional expressions? Picture a scene where a child, forming a close bond with an AI-driven toy robot, accidentally steps on its foot while playing together. In response, the robot winces and lets out a sorrowful sigh. The child, overwhelmed with concern, immediately apologizes with watery eyes. Something here might seem amiss. But a similar discomfort might arise in the converse example, where a callous child repeatedly stomps on its robot companion without any hint of regret, while the robot continually shows signs of sadness and pain over this treatment. As observers, we are left to grapple with the question: how should we respond to the robot's expressions?

The examples above illustrate the complex challenges that arise as AI and robotics become increasingly sophisticated in their ability to display and evoke emotions. Services like ChatGPT, Google's Bard, or other AI-based chatbots such as Replika are very capable of generating text that includes affective language, emojis, and other forms of writing that can be seen as emotional expressions. Humanoid social robots such as Pepper or Jibo are also capable of emotionally expressive behaviors. As a collective group, we will refer to AIs and robots with the capacity to generate emotional expressions as emotive AI. While much work has discussed the ethical implications of AI systems that can *detect* human emotional expressions (Baumann & Döring 2011; Crawford 2021; Smith & Miller 2022), much less has been said about the normative implications of AI systems that can *generate* emotional expressions that address humans in specific ways.[1] This paper aims to fill this gap.

Our starting point is the widely acknowledged fact that emotional expressions involve appeals to perceivers (see § 2). When interacting with someone, an anger expression warns you to stop what you are doing, a guilt expression prompts you to forgive, and an expression of sadness requests that you provide emotional support (see § 2). As addressees of such emotional expressions, most of us are more or less capable of immediately understanding what they signal and know how to respond. In our interactions with emotive AI, we might intuitively respond to emotional expressions the same way we are used to from human interactions. If a robot makes a sad face, you might want to comfort it. And if ChatGPT expresses regret that it cannot help you with a problem, you might feel the impulse to type "no worries." But should you? Does it make sense to respond to AI emotional expressions?

Here, we will investigate under which conditions we have practical reasons (both prudential and moral) to respond to emotional expressions' appeals, and whether those conditions could apply in the case of AI emotional expressions. Our discussion will benefit from a nuanced understanding of the nature of emotion and emotional expression. The paper will proceed as

[1] There have been some discussions of more specific areas of application, such as care (see, e.g., Coeckelbergh 2018) or love robots (see, e.g., Nyholm & Frank 2017), that can involve emotional expressions.

follows. In Sections 2 and 3, we will present the idea that emotional expressions prompt perceivers to act in certain ways. We argue that whether we should respond to those prompts does not hinge on whether the expression is genuine, but rather whether some features typical (but perhaps not constitutive) of emotion are present. In Section 4 we will introduce those features. In Sections 5 to 9, we will examine in detail which emotional features give us reasons to adapt our behavior in light of emotional expressions, and whether those features are or can be implemented in AI systems. We will conclude with some upshots for the development of emotive AI.

2. Emotional Expressions as Appeals to Perceivers

Emotional expressions (e.g. anger expressions) can involve facial (e.g. frowning), vocal (e.g. shouting), or bodily signs (e.g. clenching fists). Because these signs do not unequivocally communicate emotion, their interpretation is likely to depend on the context in which they arise, and on the perceiver. For example, someone who's frowning might be angry or concentrated, and someone who shouts can be angry, ecstatic, or just trying to speak to someone in a loud bar. For the sake of simplicity, we will talk about emotional expressions to refer to things such as frowning and shouting, but note that those might sometimes not communicate emotion.[2] It is not at issue here whether the emotional meaning of signs such as frowning or smiling is universal or culturally specific, determined by evolution or convention, and so on (see Cordaro et al. 2015, for a discussion on these topics). In any case, these signs can be interpreted as expressions of emotion at least in some contexts.

For our purposes here, it is important to distinguish between emotional expressions' intentionality and genuineness. Emotional expressions can intentionally or unintentionally communicate that the subject or speaker is experiencing a certain emotion at the moment. Under normal circumstances, unintentional expressions are accompanied by an emotion, while intentionally generated expressions are not. However, this is not always the case. For example, one can unintentionally smile without being happy, just out of habit. Conversely, one can intentionally frown to make their anger more apparent to others.

To distinguish between emotional expressions that are accompanied by emotion from those that are not, we talk about genuine emotional expressions and non-genuine emotional expressions.[3] Genuine emotional expressions require that the expresser actually experiences the

[2] Although some authors prefer to talk about "behavioral displays" rather than "emotional expressions" to avoid a strong association between the relevant signs and the expression of emotion (Fridlund 1994 2007), we will talk about emotional expressions for the sake of simplicity.

[3] We choose "non-genuine" to avoid the negative connotations of adjectives such as "fake" or "feigned"

emotional state associated with the expressions, for example: sadness when crying or anger when shouting. In contrast, non-genuine emotional expressions communicate something false, for example: that the expresser is sad or angry when they are actually not.[4]

Note that there does not have to be anything nefarious or malicious about non-genuine emotion expression. However, intentional emotional expressions can sometimes be done with questionable intentions. For example, someone can intentionally generate a smile to be nice or out of habit, but also to deceive or manipulate others. This is possible thanks to the complex information that emotional expressions can communicate in social interaction.

Emotional expressions can communicate that the agent is experiencing an emotion but also many other things. Paul Ekman (1993 2004) claimed that emotional expressions carry up to seven types of information, including information about the emotion that the expresser is experiencing, but also their thoughts, their internal physical state, what they are likely to do next, metaphors applicable to them, the antecedent circumstances, and, most important for our purposes here, what the expresser wants the perceiver to do. While not everyone shares Ekman's particular views, many researchers agree that emotional expressions communicate requests or appeals to perceivers.

It is widely acknowledged that emotional expressions communicate both (1) the emotional state of the expresser, and (2) an appeal to perceivers to behave in a certain way. These appeals are different depending on the particular emotional expression at hand. For example, anger expressions (e.g. shouting) communicate an appeal to back off, sadness expressions (e.g. crying) communicate a request for aid and support, and happiness expressions (e.g. smiling) solicit affiliation or celebration (Fridlund 1994, p. 129; Parkinson 1995, p. 286; Scarantino 2017, p. 181). Our paper will build on this basic premise.

But before moving on, some clarifications are due. First, remember that how emotional expressions obtain their meaning is not at issue here, as long as this meaning can be effectively communicated in certain interaction contexts. Second, we use terms such as "appeal" or "request" to refer to pieces of information that can take the form of an imperative sentence, such as: "back off", "help me", or "celebrate with me." Those imperatives do not necessarily carry entitlement or authority. Thus, other factors will determine whether addressees have reasons to comply with them. Our task in the next sections will be to determine what those factors are, and whether they apply in the case of AI emotional expressions. Finally, note that

[4] It might be confusing to call something an expression, when there is nothing to be expressed. Here, we choose to call fake emotional expressions "emotional expressions" to avoid more tedious formulations, such as "facial, vocal, or bodily signs that could be interpreted as the expression of an emotion."

we will sometimes talk about "responding to emotional expressions" rather than "responding to the appeals communicated by emotional expressions" for the sake of simplicity.

3. Reasons to Respond to AI Emotional Expressions: A First-Pass View

In the previous section, we have discussed that emotional expressions communicate appeals to perceivers to act in certain ways. But when do we have a reason to comply with those appeals? A first-pass answer is that we should only respond when the emotional expression is *genuine*. According to this answer, we only have a reason to respond to the request communicated by the emotional expression when the expresser is actually experiencing the emotion. For example, if someone is expressing sadness, you only have a reason to provide support if they are sad; and if someone is expressing anger, you only have a reason to back off if they are angry.[5]

While the first-pass answer has some intuitive appeal, it has important problems. Indeed, we contend that genuineness is neither necessary nor sufficient to have reasons to respond to emotional expressions.

Sometimes, we lack reasons to respond to genuine emotional expressions. Tantrums provide good examples of this. Imagine that your child throws a tantrum because you refuse to give them candy for dinner. Your child is angry. Thus, the tantrum is a genuine anger expression. And this genuine anger expression urges you to stop giving them healthy food and serve candy for dinner instead. But should you respond to this request? Probably not. Genuineness is not sufficient reason to respond to an emotional expression.

Genuineness is also not necessary. In some cases, there are good reasons to respond even when the expression is non-genuine. Consider an anger expression such as shouting. In the right context, shouting can communicate both "I am angry" and "back off!" But should you, as the perceiver, back off if the expresser is not angry? While it is true that someone angry might be more likely to engage in physical aggression than someone who is not angry, it is also true that someone who is shouting might be more likely to attack than someone who is not. Thus, you should probably back off from a shouting individual, even when their shouting is not a genuine expression of anger. Unless you want to get hurt.

[5] Furthermore, one might assume that an emotional expression is genuine only if the expressor experiences certain feelings or phenomenally conscious states. Both claims, that reasons require genuineness and that genuineness requires consciousness, are contestable. In Section 3, we focus on arguing against the former (but see Section 4 for a hint on possible challenges to the latter).

The anger case might be even more suggestive when considering an example that concerns the emotional expressions of emotive AI. Imagine a robot that is programmed to, under certain circumstances, (1) simulate an anger expression such as frowning and (2) attack the nearest target. In this case, it does not matter whether this robot can be truly said to be angry when frowning. If you see it frowning, you better back off, because it will attack you!

The main takeaway from these cases is that, when trying to assess whether we have reasons to respond to emotional expressions, we have to consider the features that are associated with those expressions (or at least more likely to be present when the expression is in place), and not whether the emotional expression is genuine. Indeed, it is irrelevant whether the features that are associated with the expression are part of an emotion or not.[6]

It is important to note that we deny that genuineness *by itself* is irrelevant to whether we should respond to emotional expressions. But we agree that genuineness could *indirectly* provide reasons to respond as an indicator of something else. For example, the genuineness of an angry expression might provide reasons to back off as an indicator that the expresser is going to attack us. But here, it is not the genuineness that provides reasons, but the tendency to attack. Something that, as we have seen, can be present in non-genuine emotional expressions.

Finally, while we maintain that genuineness simpliciter (without considering the different features of emotion) is irrelevant to whether we should respond to emotional expressions, the intention of the expression might be relevant in some cases. In § 2, we saw that genuineness and intention are separate aspects, and that people can intentionally generate emotional expressions to manipulate their interaction partners. Even if non-genuineness by itself is irrelevant, an intention to manipulate might, under certain circumstances, constitute a reason against responding to someone's emotional expression.[7]

In the next section, we will provide an exhaustive list of features that are typically associated with emotion. This will allow us to later on examine what features provide reasons to respond to emotional expressions and whether those could be implemented in AI systems.

4. Emotions versus Emotional Features

Emotional episodes can involve experiential, bodily, motivational, and cognitive features.[8] While most researchers agree that these are typical features of emotion, they disagree around which one(s) *constitute* the emotion, and which one(s) merely *accompany* the emotion, e.g., as

[6] This is something that, as we will see in § 4, depends on contentious issues about what emotions are.

[7] We would like to thank an anonymous reviewer for raising this issue.

[8] Psychological theories also frequently count emotional expressions themselves among the features typical for emotion, while philosophical theories of emotions almost always see them as their effects.

elicitors or consequences (see Prinz 2004). For example, Cognitive theories claim that emotions are cognitive states (e.g., Nussbaum 2001), and physiological changes are mere consequences of the emotion. But Somatic theories claim that emotions are bodily states (e.g., Damasio 1994), and cognitive states are contingent elicitors.

As we saw in the previous section, whether emotional expressions are genuine is, by itself, irrelevant to determine whether we should respond to it. Thus, we don't have to think here about what makes an emotional expression genuine according to different theories of emotion (a cognitive state? a physiological change? both? neither?). As long as a feature is correlated with the emotional expression, it can be the grounds under which we have reasons to respond, independently of whether that feature is consider constitutive of an emotion or not. In this section, we will briefly introduce each of the features that typically accompany emotions.

Let's take fear as an example. A textbook example of fear involves feeling uneasy (subjective experience), having an accelerated heartbeat and cold sweat (bodily changes), a tendency to run away or hide (action tendencies), thoughts about the dangerousness of the situation (evaluative cognition), and wide open eyes (expression).

Subjective experiences, or feelings, are an important aspect of emotion (though perhaps not essential, see Díaz 2023). Feelings can also feature in other types of mental states, most notably sensory states. You can feel the wind in your skin, experience the vividness of the colors in a painting, etc. Emotional experiences are usually distinguished from these other experiences because they feel good or bad. In other words, they have a positive or negative hedonic tone.

Bodily changes are a broad category. In the context of emotion, bodily changes refer to things like alterations in breathing, perspiration, or heart rate (i.e., autonomous nervous system activity, see Siegel et al. 2018), rather than alterations in weight, height, or hair length. Emotional bodily changes also typically exclude muscle reactions that are under our control, such as moving our face or limbs.

Emotional *action tendencies* are the motivation[9] or impulse to perform behaviors that are specific to different emotion types: escaping in fear, confronting in anger, etc. We might have these motivations without the relevant emotion. One thing that distinguishes emotional from non-emotional action tendencies is that they are highly prioritized over other motivations (Scarantino 2014).

[9] Note that "motivation" here is not to be confused with acting for reasons, volition, or intention.

Evaluative cognitions refer to representations of things and events as positive or negative in different ways.[10] It is usually assumed that each emotion type is associated (either causally or essentially) with a particular cognitive evaluation or appraisal (Moors et al. 2013). For example, fear is associated with representations of something as dangerous, and anger with representations of something as offensive.

Motivational bases are widely recognized as a cause and perhaps necessary precondition for emotion.[11] Motivational bases are the desires (Schroeder 2004), goals (Moors et al. 2017), concerns (Roberts 2003), and other pre-existing motivational states[12] that determine which emotions we have. For example, if I want to play basketball this evening, I'll be joyful to do so. But if you don't want to play basketball this evening, you might be annoyed if you have to play.

While it might be interesting to consider whether AI can have emotions depending on different theories of emotion, this is not the main focus here.[13] In this paper, we take the features above as *typical* of emotion, and we remain agnostic regarding whether they are *essential* to emotion. What we do want to focus on is whether these features, when correlated with emotional expressions, give us reasons to respond to them, and whether AI (can) have them. We will take this endeavor in the next section.

5. Reasons to Respond to AI Emotional Expressions: A More Nuanced View

In this and the following four sections, we discuss what reasons different emotional features can provide us to respond to the appeals made by emotional expressions, given they are integrated with those expressions. We leave out the question whether mere bodily changes can

[10] These evaluative cognitions are sometimes characterized as (analogous to) perceptions (Tappolet 2016) or beliefs (Nussbaum 2001).

[11] Some consider that other features are emotional features only when they are appropriately related to motivational bases. For example, a pounding heartbeat or a tendency to flee are features of fear if they are triggered by our concern for safety (Prinz 2004), but they are not emotional if they are triggered by intense physical exercise. Similarly, what distinguishes a *cold* assessment of the danger of a situation from a fearful evaluation of danger is that the latter is appropriately connected with our concern for safety (Nussbaum 2001; Solomon 1976).

[12] More technically, they are states with a "world-to-mind" direction of fit (Anscombe 1957).

[13] But, as it will become evident by our later discussion, some theories are more likely to allow AI emotions than others. For example, while evaluative representations seem easily within the reach of AI systems, something like bodily changes or even feelings of bodily changes seem more unlikely to be replicated in AI.

give us reasons to respond.[14] Thus, we will consider AI action tendencies, cognitive evaluations, feelings, and motivational bases.

We will consider two types of reasons: Prudential and moral reasons. Prudential reasons are reasons that relate to one's own self-interest or well-being. For example, someone may have prudential reasons to exercise regularly, eat a healthy diet, or save money for the future. Prudential reasons are often contrasted with moral reasons, which relate to broader ethical principles and considerations of right and wrong. For example, someone might have moral reasons to avoid unnecessary harm to others.

Note that, at least when it comes to prudential reasons, those can vary widely across individuals depending on their particular desires and concerns. For example, if you want to become a philosophy professor, you have reasons to publish papers in the American Philosophical Quarterly. But if you are committed to a career in competitive badminton, then you have no reason to do so. In the same way, if you want an AI robot to be happy, your reasons might be different from those of someone who wants an AI robot to be sad. Our discussion will be based on reasons that are relatively universalizable, barring niche desires and concerns.

We will take a contextual, feature-by-feature approach to elucidate what provides reasons to respond to emotional expressions' appeals. But before starting this endeavor, we would like to flag that there might be a reason to respond to *any* appeal made by *any* emotional expression in *any* context, which has to do with preserving one's (moral) character as a person that is very sensitive or "compliant" to other's emotional expressions.[15] It is also important to note that all reasons considered in the following sections are pro-tanto reasons, rather than sufficient reasons. Thus, they can and should be weighed against other reasons.

6. Action Tendencies as Reasons to Respond

In humans, certain emotions are associated with certain behaviors. For example, typically angry behavior includes such things as being confrontational or even aggressive towards others, and

[14] Can AI undergo bodily changes typical for human emotions? It seems like the answer is no, unless we fully recreate the human body including the biological functions of our respiratory, electrodermal, and cardiovascular systems. And there might be little incentives to do so. However, it is an open question whether the workings of non-biological bodies, e.g., ventilation, cooling of the motor, etc. could count as analogues to the workings of humans' autonomous nervous system. Consider, for example, artificial hearts in humans. At best, bodily changes embodied AIs can provide reasons to expect emotional behavior. If an AI is embedded in a robotic body and needs to start a certain motor to perform fast and forceful movements, then the reliable correlation of starting that motor whenever it expresses anger could be a necessary condition to take the expression seriously. In the case where the motor does not start, even if the robot would be motivated to act aggressively, we would not have reason to fear it doing so, since it would not be capable of quick and forceful actions.

[15] We would like to thank an anonymous reviewer for pointing this out.

sadness is associated with retreating to oneself. Therefore, if you don't want to be attacked, or you want the other person to do something other than retreating to themselves, then you have reasons to comply with the appeals communicated by anger and sadness expressions: "back off" and "provide support", respectively.

Note that there is an important difference between the two examples we provided above. When it comes to reasons stemming from the emotion's associated action tendency, you might have a reason to provide support to someone expressing sadness only if complying to this appeal is going to stop the expresser from retreating to themselves (e.g. by lifting the expressers' mood). In the anger case, however, you have a reason to back off even if your backing off will not stop the expresser from being aggressive. While this difference might sound trivial, it will be important when considering AI emotional expressions. For us to have reasons to respond to AI emotional expressions, sometimes it is only necessary that the expression is correlated with certain behavior. But other times, it is also necessary that our response to the expression can influence subsequent behavior. The latter case involves a further level of sophistication in the AI system at hand.

The relevant expression-behavior integrations can range from very simple implementations (like our example in § 3) to complex and intricate ones. For a simple example, a vacuum robot could have a display that shows a sad face, and make whining noises, whenever it runs low on battery, requesting this way to be put into its loading station and indicating that it will cease activity, a behavior that is somewhat typical for sadness or help-seeking in humans.

For a more complex example, an AI assistant might express being offended by how you have been treating it and while it does not stop being of assistance, it will be less pleasant or forthcoming towards you, and even get your smart toaster to pester you about it as well, until you have changed your behavior in the relevant way to resolve the grudge that it *seems* to be holding against you.

Keep in mind that, in these scenarios, we assume the emotionally expressive behavior of the vacuum robot or AI assistant negatively affects the well-being of the humans interacting with them. Therefore, these individuals have pragmatic reasons to respond to the AI agent's emotional expressions.

7. Cognitive Evaluations as Reasons to Respond

Emotional expressions in humans are typically tied to how they evaluate the situation, and that can be a source of reasons to reply to them. For example, when, as a child, your parents express disappointment with your school grades, you might feel the weight of the associated negative judgment and the need to change their opinion of you. And if everyone expresses fear at you when you are trying to have friendly interactions, you might want to change the way you portray yourself so they don't perceive you as dangerous, which is the evaluation typically associated with fear.[16] Situations like this point towards the importance people give to how others evaluate them and act in response to this evaluation.

Note that, in some cases, responding to others' emotional expression in the relevant way will not really change their evaluation of what you already did. Consider anger. The evaluation typically associated with anger is that of offensiveness. When targeted at someone, that amounts to evaluating them as a person who does something offensive. And the appeal implicit in the expression is to stop what you are doing. In this case, to not behave in an offensive manner. So when the person targeted by the angry expression desires not to be seen as offensive (or there is a moral reason to not act offensively), they would have reasons to respond to the anger expression. But this might only change the expresser's evaluation of subsequent behavior, and not necessarily their evaluation of what you already did.

All we said so far in this subsection holds only if emotional expressions can be taken to be a reliable signal of the expresser's evaluations and, crucially, those evaluations are reliable themselves. In humans, we can be confident that expressions of anger at least sometimes express reliable judgements of offensiveness, assuming that humans are relatively good at tracking offensiveness. This can of course be either abused or unreliable. In the case of AI, similar questions arise.

AI systems can perform cognitive evaluations in the sense of processing information and assessing it according to different parameters. Now, can AI evaluations, if integrated with their emotional expressions, provide reasons for us to respond to those emotional expressions? We will in turn consider personal and moral reasons for this.

When it comes to your *personal reasons* why to take emotional expressions in AI seriously, we only need to focus on whether there is a good integration between emotional expression and evaluation, without considering whether the evaluation itself is reliable or justified. You might, for example, have a relationship with your AI personal assistant such that you care about what it thinks of you independently of whether its judgments are any good. In that case, you should take your AI personal assistant's angry expressions seriously if they correlate with it representing your behavior as offensive.

[16] An important issue here is the possible gap between the responses that the expression-underlying evaluations provide reasons for and the appeals made by the emotional expression. The appeal implicit in fear expressions is along the lines of a "protect me" message. There is a long way from, e.g. concealing your tattoos from the elderly and protecting the elderly. While there might be a route to be traced between the two, note that the connection might not be a direct one.

When we examine *moral reasons* for why to stop potentially offensive behavior in light of an AI-produced angry expression, we need to consider moral epistemology. It is unclear whether and under what conditions an AI could produce its own, autonomous moral judgements, and we don't want to go into speculating about this. However, a well-trained AI could in theory provide access to a generalized moral evaluation of your situation, trained on, and thereby relating the evaluations of other people, which are pertinent to your situation. This could either be a way to get the moral judgements that people in general have about a given topic, or it could reflect the judgements of people especially affected by this type of behavior.

Imagine that a "morally-trained" Alexa expresses anger at a comment you made. This might give you good reasons to think twice about your language use, the same way in which you might have good reasons to react to fear expressions from a robot that is very accurate in detecting danger.[17] AI emotional expressions can, under certain circumstances, be a very valuable source of information and even a learning device, and there can be moral and prudential reasons to respond to them.

8. Feelings as Reasons to Respond

As we have seen in § 4, emotional feelings have a hedonic tone, that is, they are pleasurable or painful. Some important moral theories, like hedonic utilitarianism in the vein of Bentham (1789) or Mill (1861), take pain and pleasure as a central concern. According to hedonic utilitarianism, we should maximize pleasures and minimize pains. While we do not want to promote or adopt a utilitarian framework in this paper, we nonetheless want to explore the sorts of reasons connected to AI emotional expression that are relevant to such a framework.

Crucially, we can help diminish others' negative emotions by complying with the appeals of their emotional expressions. For example, if we provide support to someone expressing sadness, this can help alleviate their sadness; and stopping what we are doing can help an angry person calm down, especially if our behavior is what caused their anger in the first place.

If we ought to minimize pain, and we can minimize pain by adapting our behavior in line with the appeals of others' expressions of negative emotions such as anger and sadness, then we should indeed respond to others' emotional expressions.[18] This definitely applies in the case of humans. But can AI have subjective experiences?

[17] As long as you trust the designers, and there might be good reasons for that as well.

[18] The same reasoning also gives us moral reasons not to act in such a way as to elicit the same emotional responses in the first place.

Whether AI can have subjective experiences is a very contentious issue (see Chalmers 2022). Some authors have proposed behavioral tests for AI consciousness (Schneider 2019) but those have been criticized on the basis that much AI is designed precisely to imitate human behavior while working in different ways (Andrews & Birch 2023; Udell & Schwitzgebel 2021). Another approach is to take the physical processes responsible for consciousness in humans, and see whether those are present in AI. However, this approach faces the so-called "indeterminacy problem" (see Birch 2022; Carruthers 2019; Cutter 2017; Simon 2017). Let us briefly explain this problem.

There is an ongoing debate on whether consciousness can be identified or fully explained by physical processes, and what those physical processes might be. But even if we agree that consciousness is realized by a specific physical process, or set of processes, we might not be able to determine when an entity is conscious or not. This is so because the processes that realize consciousness are likely complex, meaning, they have various features, and it is not clear which of those features, or which subset of them, are necessary for the realization of consciousness.

We could be certain that entities with the same features as humans are conscious. But it is indeterminate whether entities are conscious when they are similar to us in some aspects but different in others. Given that these conditions apply in the case of AI, it is indeterminate (there is no fact of the matter to) whether AI is conscious. And, in turn, it is indeterminate whether we should respond to AI emotional expressions in a way to alleviate negative emotion or foster positive emotions based on their feelings.

However, note that even if the status of AI as conscious is indeterminate, an argument from moral precaution could be made (see, e.g., Knutsson & Munthe 2017). A precautionary stance would tell us that it would be better to give a being due consideration, even if it does not actually feel, rather than disregard it when it is able to feel. Hence, if we have at least some degree of likelihood that an AI could have subjective experiences of pain and pleasure, we have some morally precautionary reason to respond to its emotional expressions. How to determine the likelihood that AI can have feelings, however, is unlikely to be an easy task (but see Butlin et al. 2023).

9. Motivational Bases as Reasons to Respond

The motivations that usually underlie emotional expressions can also provide reasons to respond to them. For example, imagine a librarian gives you an angry look across the room when you start talking. Given that you are used to the norms of libraries, this can easily be interpreted as a plea for silence. You might already have sufficient reasons to comply with this plea, based on the likely behavior that would follow non-compliance. The librarian could take steps to enforce the rule to be quiet, given by the means available to them, by either confronting you, alerting someone to your talking, or even removing you themself. But you might also just want to react to the fact, expressed in the angry look, that the librarian does not want you to be loud.

We have claimed, in § 4, that the emergence of emotions in humans depends on our motivations, like desires, preferences, goals, or concerns. In the above example, the angry look can be interpreted to reflect that the librarians' desire for silence is frustrated. Can AI have desires and other motivational states?

The fact that very different types of motivational states (desires, preferences, goals, concerns, etc.) can cause the emergence of emotions in humans suggests that whether something can be a motivational base for emotion does not depend on its particular features, such as their neural implementation, but on what's common to all of them. Contrary to what happened with the conditions for phenomenal states, the conditions for having motivational states might not lead to an indeterminacy problem (see § 8).

Arguably, what is common to all motivational states is their *world-to-mind* direction of fit (they aim at changing the world, rather than accurately representing the world) or functional role in pushing a certain type of behavior. Motivational function can thus be *the one thing* that we can point towards when trying to identify motivational states.

If motivational states can be implemented in an AI system, and they can be connected to emotional expressions in the same way as they are in humans, then they might provide us reasons to reply to AI emotional expressions. The question is, can motivational states be implemented in AI? It seems hard to deny that reinforcement learning AI has motivational states as functionally described above.[19] Indeed, the "building blocks" of reinforcement learning agents are representations of the state of their environment, and goals regarding the "desired" state of the environment (Sutton & Barto 2014).

Preference Utilitarianism (Hare 1981, Singer 2011) posits that we should maximize the satisfaction of preferences, and minimize their frustration.[20] If, as it happens with humans, AI negative emotional expressions emerge when things go against their preference-like states, and if changing our behavior can change the extent to which these motivational states are

[19] Note that the issue here is not whether reinforcement learning systems can be called "agents" in the thick sense of acting for reasons (Butlin 2023), but whether they have motivational states, i.e. "end states" that determine their behavior in a flexible way (in contrast to fixed stimulus-driven behavior).

[20] A similar point can be made under a non-utilitarian preference-based theories of the good, e.g., Rawls (1971). As mentioned in § 8, page 11, we do not argue for or adopt either a utilitarian nor a preferecne-based ethical framework, but focus on the conditional claim that under such frameworks, these emotinoal features become relevant bases for moral reasons to respond to AI emotional expressions.

frustrated, we might have a moral reason to do so, if we accept this normative framework. However, this might also depend on what kinds of motivations become implemented in AI. For example, whether it simply has the goal to win chess games, or whether it has the motivation to survive and thrive itself.

In humans or animals, we don't take just any motivation as morally relevant. For example, being a fan of a specific basketball team might be important for an individual, but it does not seem necessary for them to live a good life. In effect, such a concern might have a lesser importance, or only a derived moral importance, since it only seems to have an indirect moral value. Emotions relating to these types of concerns might have some moral importance as well and should be taken somewhat seriously, but just like the value of the motivations, their importance is likely derived from something like basic needs.

Even more clearly problematic are motivations that run counter to the good life or well-being of a being itself or others. Humans, but also other animals, might have self-destructive impulses or ideologies. Predators desire to hunt and thereby destroy the lives of other beings. These concerns, while some might have initial moral importance, raise hard problems for how to weigh conflicting concerns and desires against each other.

The more clear cases of morally important motivations are what we might call basic needs. The desire to retain bodily integrity, sustenance, belonging, etc. are commonly counted among the core morally relevant goods in list theories of well-being or justice (Brock & Miller 2019; Pölzler 2021). Emotions that relate to these basic motivations, such as fear for your life, joy in social interaction, or anger at someone disrespecting your basic rights, would therefore count among those we should take most seriously.

We can now turn to the case of AI and see whether the same basic, derived and problematic motivations might apply, and whether we should take them seriously to the same degree. The most relevant class seems to be the basic needs. For example, if an AI cares for its own integrity, continued existence, or well-functioning, that might be a sufficient analogue to humans' fear for their life. Or, at least, provide prima-facie or defeasible reasons against disregarding the emotional expressions that are connected to those concerns.

It has been argued that, even though plants' and other living entities' need for self-preservation can give us reasons to not wrong them, those reasons are easily overridden by other considerations (Gibert & Martin 2022). However, note that self-preservation is not the only basic need that could be implemented in AI systems. Things like companionship, social acceptance, and freedom from harassment are also considered basic needs (Braybrooke 1987) and could be implemented as goals in emotive AI systems. A reliable integration between AI emotional expressions and the achievement or frustration of these goals would make interaction with emotive AI much more realistic.[21] But it would also give us strong moral reasons to respond to AI emotional expressions.

10. Conclusion

In this paper, we have argued that the normative force of AI emotional expressions does not depend on whether those expressions can be considered genuine (and thus on one's preferred theory of emotion). Instead, different features that are merely associated with emotion can provide prudential and even moral reasons to comply with the appeals communicated by AI emotional expression, given that they are properly integrated with them. We hope that our discussion paves the way for further work on this topic that goes beyond simple "AI equals no emotion" claims. In concluding, we would like to suggest some practical implications of our results.

Regarding prudential reasons, we have argued that there are prudential reasons to act in response to AI emotional expressions when those are integrated with behaviors or evaluations that are relevant to our goals (§ 6-7). The association between particular signs and behaviors is already exploited in systems as simple as an engine light. And the possibility of using realistic emotional expressions for the same purposes could open up concerns about (emotional) manipulation. In the case of evaluative cognitions, integrating realistic emotional expression with the computational power of AI systems can guide the development of educational tools that help humans learn about (typical) evaluations of their behavior and facilitate social interaction for those humans who struggle with it.

Regarding moral reasons, some prominent moral theories entail that there are moral reasons to adapt to AI emotional expressions if those are integrated with subjective feelings, but the status of AI as conscious is problematic (§ 8). However, we have shown that there might be moral reasons to adapt to AI emotional expressions if those are grounded on (functionally defined) motivational states which push AI agents to pursue things that can be considered basic needs, such as self-preservation or companionship (§ 9). While motivational states can be implemented in AI computationally, no AI to our knowledge is equipped with motivational states concerning basic needs. But note that there might be incentives to do so.

The desire to make AI more human-like could prompt developers to make AI that is motivated to pursue fundamental human needs. Consider Dolores from Westworld, which is endowed

[21] Someone might want to think that it is desirable to create, for example, a Replika that really has a goal for social connection (basic need), because that would make it more sophisticated (intelligent and flexible) and less scripted (if-then behavior). Then we might have (moral) reasons to adapt, it won't be about playing games any more.

with the motivation to care for her child. Or Replika, the romantic chatbot, is advertised as "the AI companion who cares." It is an open question whether Replika really cares, but what is sure is that romantic robot like Replika might get a totally new fascinating dimension in terms of realism and user engagement if programmed with a desire for affection or emotional support. If our considerations are on the right track, this will also gives us reasons to respond to its emotional expressions. We should no longer consider our romantic interactions with it as mere theatrical play. Indeed, we might have moral reasons to respond to it in certain ways.

This raises a tension or even a dilemma between two opposing goals. On the one hand, we could want to create AI that can emotionally express itself and give us reasons to respond to these expressions accordingly. This goal might be based on the aim to make human-AI interactions more fluid and feeling more natural, for AI to serve us more effectively. On the other hand, we might not want to create a group of beings with such sophisticated workings, including goals such as self-preservation or social acceptance, that we have to take them as morally relevant. Because then, we could no longer simply use those beings as servants, completely subjected to our own goals, needs and desires (Schwitzgebel & Garza 2020).

Affiliations of Authors

Rodrigo Díaz

Institute of Philosophy, CSIC. Albasanz 26, 28037 Madrid, Spain.

Centre for Research in Ethics, University of Montreal. 2910 Boulevard Édouard-Montpetit, H3C 3J7 Montréal, Canada.

rodrigodiaz.philosophy@gmail.com

Jonas Blatter

Institute for Philosophy I, Ruhr University Bochum. Universitätsstraße 150, D-44801 Bochum, Germany.

jonas.blatter@ruhr-uni-bochum.de

Acknowledgements

We would like to thank Martin Gibert and participants in Eva Weber-Guskar and Sebastian Weydner-Wolkmann's Colloquium Digitale for their comments on early drafts of this paper. This work is funded by the Ministry of Culture and Science of North Rhine-Westphalia (grant number: PROFILNRW-2020-135) and the Spanish Ministry of Science and Innovation (grant number: JDC2022-048913-I).

References

Andrews, K. & Birch, J. (2023). What has feelings? Aeon. Retrieved from https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals

Anscombe, G. E. M. (1957). Intention. Harvard University Press.

Braybrooke, David 1987, *Meeting Needs*, Princeton: Princeton University Press.

Baumann, H., & Döring, S. (2011). Emotion-Oriented Systems and the Autonomy of Persons. In R. Cowie, C. Pelachaud, & P. Petta (Eds.), *Emotion-Oriented Systems: The Humaine Handbook* (pp. 735–752). Springer. https://doi.org/10.1007/978-3-642-15184-2_40

Bentham, J. (1789/1907). An Introduction to the Principles of Morals and Legislation. Oxford: Clarendon Press.

Birch, J. (2022). Materialism and the Moral Status of Animals. The Philosophical Quarterly, 72(4), 795–815.

Brock, G., & Miller, D. (2019). Needs in Moral and Political Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/needs/

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., … VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. Retrieved from https://arxiv.org/abs/2308.08708

Carruthers, P. (2019). Human and Animal Minds: The Consciousness Questions Laid to Rest. https://doi.org/10.1215/00318108-9264069

Chalmers, D. (2022). Reality+: Virtual Worlds and the Problems of Philosophy. W. W. Norton & Company. https://doi.org/10.5840/tpm20229875

Coeckelbergh, M. (2018). Why Care About Robots? Empathy, Moral Standing, and the Language of Suffering. *Kairos. Journal of Philosophy & Science* 20(1), 141–158. https://doi.org/10.2478/kjps-2018-0007

Cordaro, D. T., Fridlund, A. J., Keltner, D., Russell, J. A., & Andrea Scarantino. (2015). Debate: Keltner and Cordaro vs. Fridlund vs. Russell. *Emotion Researcher: ISRE's Sourcebook for Research on Emotion and Affect*. https://emotionresearcher.com/the-great-expressions-debate/

Crawford, K. (2021). Affect. In *The Atlas of AI* (pp. 151–179). Yale University Press; JSTOR. https://doi.org/10.2307/j.ctv1ghv45t.8

Cutter, B. (2017). The Metaphysical Implications of the Moral Significance of Consciousness. Nous-Supplement: Philosophical Perspectives, 31(1), 103–130.

Damasio, A. (1994). Descartes' Error. Emotions, Reasons and the Human Brain. New York: Avon Books.

Díaz, R. (2023). Against emotions as feelings. Towards an attitudinal profile of emotion. Journal of Consciousness Studies. Volume 30, Numbers 7-8 2023, pp. 223-245(23). https://doi.org/10.53765/20512201.30.7.223

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*, 384–392. https://doi.org/10.1037/0003-066X.48.4.384

———. (2004). Expression or communication about emotion. Uniting Psychology and Biology: Integrative Perspectives on Human Development., 315–338. https://doi.org/10.1037/10242-008

Fridlund, A. J. (1994). Human facial expression: an evolutionary view. San Diego: Academic Press.

Gibert, M., & Martin, D. (2022). In search of the moral status of AI: why sentience is a strong argument. AI and Society, 37(1), 319–330. https://doi.org/10.1007/s00146-021-01179-z

Hare, R. M. (ed.) (1981). Moral Thinking: Its Levels, Method, and Point. Oxford: Oxford University Press.

Knutsson, S. & Munthe, C. A. (2017). Virtue of Precaution Regarding the Moral Status of Animals with Uncertain Sentience. *Journal of Agricultural and Environmental Ethics, 30*, 213–224. https://doi.org/10.1007/s10806-017-9662-y

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. Emotion Review, 5(2), 119–124. https://doi.org/10.1177/1754073912468165

Moors, A., Boddez, Y., & De Houwer, J. (2017). The power of goal-directed processes in the causation of emotional and other actions. Emotion Review, 9(4), 310–318. https://doi.org/10.1177/1754073916669595

Mill, J. S. (1861/1998). Utilitarianism. Roger Crisp (ed.). Oxford: Oxford University Press.

Nyholm, S., & Frank, L. E. (2017). From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible? In J. Danaher & N. McArthur (Eds.), *Robot Sex: Social and Ethical Implications* (pp. 219–243). MIT Press.

Nussbaum, Martha C. (2001). Upheavals of Thought: The Intelligence of Emotions. Cambridge University Press.

Parkinson, B. (1995). *Ideas and Realities of Emotion*. *Ideas and Realities of Emotion*. London: Routledge. https://doi.org/10.4324/9780203135648

Prinz, J. J. (2004). Gut Reactions: A Perceptual Theory of the Emotions. Oxford University Press.

Pölzler, T. (2021). Basic needs in normative contexts. Philosophy Compass, 16(5), 1–14. https://doi.org/10.1111/phc3.12732

Rawls, J. (1971). A Theory of Justice, Cambridge, MA: Harvard University Press.

Roberts, R. C. . (2003). Emotions: An essay in aid of moral psychology. Cambridge: Cambridge University Press.

Russell, J. A., Miguel, J., Dols, F., & Fridlund, A. J. (2017). 5 The Behavioral Ecology View of Facial Displays, 25 Years Later. In J. A. Russell & J. M. Fernandez-Dols (Eds.), The Science of Facial Expression (pp. 77–92). Oxford: Oxford University Press.

Scarantino, A. (2014). The Motivational Theory of Emotions. In J. D'Arms & D. Jacobson (Eds.), *Moral Psychology and Human Agency* (pp. 156–185). Oxford University Press.

———. (2017). How to Do Things with Emotional Expressions: The Theory of Affective Pragmatics. Psychological Inquiry, 28(2–3), 165–185. https://doi.org/10.1080/1047840X.2017.1328951

Schneider, S. (2019). Artificial you : AI and the future of your mind. Princeton University Press.

Schroeder, T. (2004). Three Faces of Desire. New York: Oxford University Press.

Schwitzgebel, E., & Garza, M. (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom. In Ethics of Artificial Intelligence (pp. 459–479). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0017

Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., … Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. Psychological Bulletin, 144(4), 343–393. https://doi.org/10.1037/bul0000128

Simon, J. A. (2017). Vagueness and zombies: why 'phenomenally conscious' has no borderline cases. Philosophical Studies, 174(8), 2105–2123. https://doi.org/10.1007/s11098-016-0790-4

Singer, P. (2011). The expanding circle: ethics, evolution, and moral progress. Princeton, NJ: Princeton University Press.

Smith, M., & Miller, S. (2022). The ethical application of biometric facial recognition technology. *AI & Society*, *37*(1), 167–175. https://doi.org/10.1007/s00146-021-01199-9

Solomon, R. C. (1976). The Passions. University of Notre Dame Press.

Sutton, R. S., & Barto, A. G. (2014). An introduction to reinforcement learning. MIT Press. https://doi.org/10.4018/978-1-60960-165-2.ch004

Tappolet, C. (2016). *Emotions, Values, and Agency*. Oxford University Press.

Udell, D. B., & Schwitzgebel, E. (2021). Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed. Journal of Consciousness Studies, 28(5), 121–144.