**Are current AIs moral agents? Why?**

Essay submitted for the examination of Subject 198 in Philosophy

**Candidate Number**: 1042151

**Candidate School**: Honour School of Mathematics and Philosophy (Part C)

**Word count**: 4924

**Are current AIs moral agents? Why?**

In the following essay, I will argue that the current AIs are not moral agents. I will first criticize the influential argument from sentience accounted by Véliz. According to Véliz, AIs are not moral agents because AIs can not feel pleasure and pain. However, I will show that moral agency does not necessarily require the ability to be sentient and refute Véliz’s argument. Instead, I will propose an argument from responsibility. First, I will establish the truth that moral agents necessarily have the ability to bear moral responsibility. Next, I will explain three major accounts of responsibility introduced in the *Stanford Encyclopaedia of* *Philosophy*. Then for each of the three accounts, I will show that current AIs do not have the ability to bear responsibility. In conclusion, this argument shows that current AIs can not be moral agents. At last, this essay will also discuss the possibility of AIs being partial moral agents and institutional moral agents.

**Contents**

[1. Argument from sentience 3](#_Toc133500464)

[2. Argument from responsibility 6](#_Toc133500465)

[2.2. Form of the argument 6](#_Toc133500466)

[2.3. Can AI bear moral responsibility? 7](#_Toc133500467)

[3. Partial and institutional moral agents 12](#_Toc133500468)

[3.1. Partial moral agents 12](#_Toc133500469)

[3.2. Institutional moral agents 12](#_Toc133500470)

[4. Conclusion 14](#_Toc133500471)

# Introduction to the question

Moral responsibility

* Ethical goodness attribution
	+ Not well-designed
	+ But you can not say it is a morally good AI
	+ The problem of responsibility
* AI-human Trust
	+ AI-human Relationship – instrumental - but not trustful and authentic
	+ The problem deception/exclusion/loneliness because it is lack of meaning.

Argument of Veliz

* Very Intuitively appealing to some, but extremely unappealing to others
	+ (over-generation of pain and pleasure)
	+ Physicalists even do not care about sentience…
* Avoid this problem to find a solution that is appealing to those who are sceptical of whether sentience can be known at all
* Difference in strength, the arguments will have different answers to the question of what AIs are important?
	+ Veliz is meant to have a very strong negative conclusion
	+ The argument of . As I will explain, it is possible for AIs to be moral agents.

# Argument from sentience

Roughly speaking, Véliz's (Véliz, 2021) argument can be presented as the following:

1. Sentience is necessary for moral agency.
2. Current AIs are not sentient.
3. (Conclusion) Therefore, the current AIs are not moral agents.

In Véliz's words. sentience is 'the capacity, at a minimum, to have a subjective experience of pleasure and pain. ' (Véliz, 2021, p. 487) So, in other words. Véliz argues that AIs are moral agents only if AIs have the capacity to experience pleasure and pain. However, because current AIs do not seem to experience pleasure and pain, it means that the current AIs are not moral agents. While the argument looks simple and intuitively attractive, this argument is not sound.

The point of contention is premise (a). To argue for premise (a), Véliz conceives of a possible being called a 'moral zombie'. A moral zombie can not experience pleasure, pain, empathy, intimacy, remorse, guilt, shame, or any other moral emotion. Furthermore, a moral zombie cannot be empathetic towards others' satisfaction and suffering (Véliz, 2021, p. 493). According to Véliz, this means that even if moral zombies have a list of moral goals and act successfully to reach these objectives, they cannot act with 'coherent' moral considerations (Véliz, 2021, p. 494). Véliz claims that this is because empathetic feeling of others' pain and pleasure is the foundation of moral considerations (Véliz, 2021, p. 494). Véliz then draws the conclusion that moral zombies are not moral agents. According to Véliz, the possibility of moral zombies demonstrates that sentience, as a necessity for moral considerations, is necessary for moral agency.

To take stock, the argument for (a) can be broken down into the following premises:

1. Sentience is necessary for empathetic feelings of others’ pain and pleasure.
2. Empathetic feeling is necessary for moral considerations.
3. Moral considerations on empathetic feelings are necessary for moral agency.

However, all three sub-premises for premise (a) can be challenged. First, one can challenge premise (a.1) that sentience is necessary for the empathetic feeling of others' pain and pleasure. When we reflect on our everyday life, we can notice that empathetic feelings arise for others by imagining others in pain or pleasure. But it is far from true that we need to experience actual pain (or pleasure) by imagining others in pain (or pleasure). For example, when we see someone being physically beaten, we may feel bad for his pain, but we are not experiencing pain by virtue of feeling bad for his pain. To better demonstrate the challenge, I will give a sharp counterexample to this necessity claim. In this example, we conceive of a future where technologies can impede a person from feeling pain or pleasure. I believe this is well conceivable. In this future, there will be permanent pain-killing and pleasure-killing treatments. Suppose Joe, who was a normal agent capable of feeling pain and pleasure, decides to permanently kill his pain and pleasure to follow the teaching of Buddhism. Suppose Joe is determined to live without any pain or pleasure. Slowly with the passing of time, he can not remember a single incidence in his life in which he is pleasant or painful. But Joe does know and remember that pain is something people want to avoid, and pleasure is something people want to pursue. When someone complains to Joe that he is in great pain, we would still think Joe understands what it means and feels bad for the person, despite Joe lacks feeling of pleasure and pain. If this is possible, then it shows that empathetic feelings do not need sentience.

Second, we can challenge premise (a.2) that an empathetic feeling for others' pain and pleasure is a necessity for moral considerations. Similar to the above case, I will give a counterexample to this claim. This example is constructed upon a medical condition called ‘congenital insensitivity’ to pain. Like Joe, patients with this condition lacks experience of pain. Because they are so insensitive to pain, some patient may bite their mouth into deformation. However, since pain-insensitive individuals can not experience pain, we can not coherently imagine them being in pain; so by definition, we can not be empathetic for pain-insensitive patients. Does this mean that we can not have moral considerations to treat and protect them? I think not. It is intuitive that we have moral considerations to treat and protect pain-insensitive individuals. We care about them as members of our community and the embodiment of human dignity. These moral considerations are not based on the empathetic feelings of the insensitive individuals but on our obligations to humans. Hence regardless of whether we can feel for pain or pleasure of the pain-insensitive individuals, we have an obligation to maintain their human dignity.

Third, one can challenge (a.3) that moral considerations based on empathetic feelings are necessary for moral agency. We can use psychopaths to construct a counterexample. As we know, psychopaths have an impaired ability to be empathetic to others. But psychopaths are moral agents. In fact, when psychopaths do something wrong, we hold them responsible for there actions.[[1]](#footnote-2) Psychopaths can be punished when they do something wrong. For example, we sentence psychopaths to jail. In fact, a survey even shows that psychopathy is an aggravating factor for imprisonment in sentencing.[[2]](#footnote-3) Psychopaths also have the obligations to act morally, because it is possible to teach psychopaths to be pro-social, even if the probability plummets as they grow up[[3]](#footnote-4). So depite psychopaths lack the ability to have empathetic feelings, they are moral agents.[[4]](#footnote-5)

In addition, based on the above thought experiments, we can conceive of a counterexample case directly against premise (a). We can continue with the example of Joe, who is no longer sentient by taking the treatment to remove feelings of pain and pleasure. For convenience of reference, we can call him the Buddhist Joe. In this example, Buddhist Joe maintains his moral principles and his ability to reason and act based on his principles. Suppose now Joe does something wrong. Joe assassinates the head of a crucial institute because he falsely deduces from his principle that killing the head of the institute benefits humanity. But in fact, he causes a period of chaos in the world with the absence of a crucial figure. Should Joe be held responsible for this chaos? I think yes. We will hold Joe responsible for his action: we will blame and punish Joe for his wrongdoing, and we will reason with Joe about the immorality of homicide. This example of Buddhist Joe shows that non-sentient beings can be moral agents, and premise (a) is wrong.

To conclude this section, Véliz's positive argument for premise (a) does not stand. There might be other ways of arguing for premise (a) or other conceptions of sentience that can circumvent these counterexamples. But the burden of proof is on the defenders' side for this premise.

# Argument from responsibility

## 2.1. Form of the argument

This section will propose an alternative argument against the moral agency of current AIs from responsibility. This argument follows the same general form as the argument from sentience:

* 1. Moral agency necessarily requires the ability to X.
	2. Current AIs do not X.
	3. (Conclusion) Current AIs are not moral agents.

In the previous section, we have tried to use sentience for X and showed that it does not work. To give another argument in this form, one must determine an X that satisfies both premises 1 and 2.

A good starting point for the analysis will be the concept of the agent. In its minimal notion, an agent is conceived to be a thing that can initiate events. This widely recognised definition of an agent will be the foundation of the following analysis. (Schlosser, 2019, Sec 2.2) From this minimal definition, philosophers have developed many definitions of specific agents. These definitions differ on how agents initiate an event. For convinience of discussion, we can call them conditions for initiation (sometimes referred to it as *condition* for simplicity). As we will show, the difference in condition gives rise to specifications of agents, such as physical agents, mental agents, and moral agents.[[5]](#footnote-6)

One way of initiating events is through physical causation. Because physical causation captures the dynamics of events under the laws of physics, we can specify the concept of physical agent by taking physical causation as the condition for its initiation. For example, a human is a physical agent. A human can initiate the movement of a ball by pushing the ball. The push physically causes the transfer of momentum from the human to the ball. As the effect, the ball moves. However, physical causation is obviously too loose as a condition for moral agency. Almost all physical things seem to initiate physical events in some way. If one accpets that an agent is a moral agent only by virtue of being a physical agent, then almost everything is moral agent. But this can not be the case. So, X can not be physical causation.

Another option of X is mental causation. Some mental events are said to be initiated with consciousness (i.e. intentionality, sentience, phenomenology). If an agent can act with consciousness, then the agent can be specified as a mental agent. For example, when a human dances, she is also consciously experiencing the event: she has an intention about her body, she feels her body moving and, perhaps, enjoys the pleasure in dancing. However, consciousness is likely too strict to be a condition for moral agency. For example, the case of Buddhist Joe shows that sentience is not the condition for moral agency. Similar counterexamples might be conceived against other aspects of consciousness. [[6]](#footnote-7) This essay will not investigate these counterexamples. As the following alternative is clearly more accurate as a condition for moral agency, I believe there is also no need to dive in.

So, instead, this essay suggests that the additional condition for moral agency is the ability to bear moral responsibility. Moral agents’ actions are different from non-moral agents’ actions by virtue of the responsibility attribution for the action. When a moral agent acts, its action allows it to be morally responsible for the moral significance of the event. This suggestion is highly intuitive and plausible. For example, if an adult needs to bear responsibility for killing an endangered bird for fun, then it means that the adult is a moral agent. Conversely, when an agent is not responsible for any moral significance of its action, it is not a moral agent. For example, if a recommender system (on its own) does not bear responsibility for its misinformation, then it is evident that the recommender system is not a moral agent. Therefore, the ability to bear (some) moral responsibility is a sufficient and necessary condition for moral agency. As we settle the X in the general argument as bearing responsibility, we can move on to the question of whether the current AI models can bear moral responsibility.

## 2.2. Can AIs bear moral responsibility?

As the philosophical investigation continues, this essay will now focus on the issue of moral responsibility. According to *The Stanford Encyclopedia of Philosophy*, there are three major accounts of moral responsibility: reason-responsiveness account, forward-looking account, and reactive attitude account (Talbert, 2022). This essay does not aim to adjudicate among the three accounts of responsibility.[[7]](#footnote-8) Instead, the focus is to reject the claim that the current AI models are moral agents. To do this, I will try to find a necessary condition for moral agency under each of the three accounts of moral responsibility. I will use empirical cases to demonstrate how current AI models fail each of these necessary conditions. As the current models fail all accounts of bearing moral responsibilities, I will conclude that the current AIs are completely not moral agents.

First, let us suppose that we adopt the reason-responsiveness account of moral responsibility. According to the reason-responsiveness account of moral responsibility (Talbert, 2022, Sec 2.3), an agent bears responsibility for its action when its action is generated with a reason-responsive mechanism. To understand the reason-responsive mechanism, we should first understand reasons. Reasons are linguistic messages of motivation; upon understanding the message behind the reason, one possesses a motivation to act in a certain way. The reason-responsiveness mechanism is the mechanism for this understanding. [[8]](#footnote-9)

The reason-responsive mechanism is an important pre-condition of bearing responsibility, because without this mechanism, an agent can not act reasonably and bear responsibility for its unreasonable action. To demonstrate this, let us suppose that to have fun in dance is a reason to dance. Now, suppose that Kay has the reason-responsive mechanism. Kay is motivated to dance because she understands and accepts the reason that dancing is fun. In this case, when Kay neglects her duty to do homework because she spends too much time having fun dancing, we will think that Kay should be responsible for her neglect. In comparison, Paul developed a condition of involuntary dancing[[9]](#footnote-10). This condition causes Paul to engage in dancing unexpectedly; there is no reason that can motivate Paul to do otherwise. For instance, even if he understands that behaving well is a good reason to stop, he can not make this motivation effective. Because of this, we will not see Paul responsible for his unexpected dance. These contrary examples show the intuition that one only bears responsibility for reasonable actions. But at the same time, the necessary precondition to act reasonably is to have the mechanism to understand reasons and act from reasons. Hence it is necessary for an agent to have a reason-responsive mechanism to bear responsibility.

However, merely being changed by reasons is insufficient to show the ability to bear moral responsibility; the agent needs to act in accordance with the motivations behind the actions. For example, a cat can change its course of action because the cat is scared by the sound made in our reasons, but it does not show that the cat's action is reasonable, and the cat does not bear responsibility for its response in actions. Only when an agent acts in accordance with the content of the reasons (and the agents accept the motivations), the agent's action is reasonable, and the agent can bear moral responsibility for the action. Following this argument, one can see that reason-responsiveness requires understanding reasons and action from the understanding. Without understanding, there is no reason-responsive mechanism.

Can the current AI models have this reason-responsive mechanism? Some AIs appear to have this ability. For example, GPT-4 is one of the most advanced large language models that can respond to our questions and reasonings satisfyingly. Moreover, it possesses the ability to code directly according to our request. Many developers have given GPT-4 the ability to control robots directly. For example, if you ask GPT-4 to control a robot in code and tell the model that dancing is fun, it will give you an abstract of code to let the hypothetical robot dance. [[10]](#footnote-11) These robots embodied by GPT appear to understand our reasons and deliver behaviour determined by the reasons. However, beneath the appearance, the mechanism of large language models implies that they do not have the ability to understand reasons. According to the current technologies, large language models are trained to mimic human behaviour and speech with big data. When the model analyses the reasons we give to them, they process the reasons syntactically to generate codes that match the answer for the question in the data poll. Because they do not understand the content of the reason, their response in words or codes can be insensible in many cases[[11]](#footnote-12). This shows that although models of AI are changed by reason, they do not possess the ability to act according to the content of the reasons. So, the current AI models, as advanced as chatGPT, do not have the ability to bear moral responsibility under the reason-responsive account.

In comparison, the forward-looking account of moral responsibility states that an agent can bear moral responsibility only when the agent can be punished or rewarded for its actions (Talbert, 2022, Sec 2.1). Some may argue that this shows that sentience is a necessary condition for an agent to bear moral responsibility, for someone can only feel the punishment or reward only if it is sentience. But this does not need to be the case. As we have discussed**,** psychopaths (or Buddhist Joe) do not feel the punishment (or reward) as normal humans do, but in legal practice, we do punish psychopaths for their misconduct. From the perspective of consequentialism, [[12]](#footnote-13) one can give many justifications for punishments, including retribution, deterrence, rehabilitation, and incapacitation.[[13]](#footnote-14) For example, incapacitation aims at reducing the capacity of the agent (i.e. by putting it to jail) such that it can not harm others even if it wants to. Hence the consequences of incapacitating a psychopath with violent intentions as punishment are positive. So then, according to the forward-looking account, a psychopath (or Buddhist Joe) can bear responsibility by virtue of being punished.

A similar argument is applicable to AI. Since we can restrict an AI's usage and modify it when it acts inappropriately, could this be seen as a form of incapacitation and rehabilitation for AIs? There is room for debate; I will argue that it is not the case. For example, if a knife is dull and it does not cut as we intend it to, we will stop using it until it is sharpened again. But this is not a kind of punishment against the knife. Because when the knife is used to cut, the knife is not in control of the cutting; therefore, limiting the use of the knife does not incapacitate knife in terms of reducing its range of the knife is in control of. The knife's user is in control, and one incapacitates the user by removing the knife from it. Hence, following this intuition, only when something has rational autonomy, can it be punished (or rewarded). Do the current AI models have rational autonomy? They do not. To have rational autonomy, an agent must be able to pass independent judgements of a situation and possess the capacity to act according to the judgements it makes. But as AIs are not reason-responsive, the AIs lack the ability to understand the content of the judgements and act accordingly. Moreover, the judgements made by the AI models are highly dependent on the training data it has. According to consensus, the current AI models only have limited inference and causal investigation ability. This means that the current AI models lack the ability to investigate and conceptualize the situation on their own. Then according to the forward-looking account, because the current AI models lack the rational autonomy to be punishable (and rewardable), they cannot bear moral responsibility.

At last, according to the reactive-attitude account of moral responsibility, an agent bears moral responsibility only if it can be blamed for its wrongdoings (Talbert, 2022, Sec 2.2). When we blame someone, we respond negatively towards the others' attitudes towards us in their actions. Under a strong interpretation of attitudes, to demonstrate attitudes, one must be a mental agent; the agent must have intentions and sentience to invoke attitudes. Under this interpretation, as sentience is necessary formoral agency**,** the argument against the moral agency of current AIs will reduce to the argument for sentience. As we have discussed, this is a counter-argument against current AIs' moral agency. Thus, in the following discussion, I will assume a weak interpretation of attitudes. Under this interpretation, attitudes can be possessed by non-mental agents like AIs. In this notion, attitudes are dispositions to act in certain ways directed to someone.

Current AI models can be personalized to show dispositions towards different individuals. Does this show that current AIs can show attitudes towards a particular person? I think not. The current AI models do not have the ability to recognize an individual as a person despite its personalization. For example, recommender systems are developed to select personalized content for individuals. To select a piece of content, it either compares the content with other contents in the user's history or compares the user with other users to see what content they consume in general. However, the model does not recognize the user as an individual person in this process. It does not understand the user's preferences, thoughts, or reasons for each piece of content. So, suppose a software recommends content that is harmful to Gary because other users watch them. Should Gary morally blame the software for its recommendation? I think not. [[14]](#footnote-15) When someone blames an agent, it aims at adjusting and improving the dispositions of the agent towards itself, but because of the mechanism of the recommender systems, it is determined that the system cannot hold a directed disposition towards Gary as an individual. So, Gary cannot blame the recommender system for it has no attitude towards Gary. Therefore, even under the reactive-attitude account, the current AI models, such as the current recommender system, cannot bear moral responsibility.

To take stock, this essay argues that an AI can bear moral responsibility only if it has the ability to be reason responsive, or be rational autonomous, or recognize an individual person. Unfortunately, the current AI does not have any of the above abilities, so it cannot bear moral responsibility in any of the three plausible accounts of moral responsibility. Yet, as the argument from moral responsibility shows, moral agents necessarily have the ability to bear moral responsibility. Hence the current AI models are not moral agents. In the next section, I will discuss some compromises one may make to accommodate the intuitions of current AI's moral agencies.

# Partial and institutional moral agents

## 3.1. Partial moral agents

First, the scope of currency in the question can raise ambiguities. Some may maintain that AIs in the near future may be qualified as current AIs. So even under the investigation of current Ais, it is relevant to look further away in development to see if they can be moral agents. As we have discussed, the current AIs completely lack moral agency because it is not reason-responsive or rationally independent, nor does it have the ability to recognize an individual. As we have discussed, the mechanism of deep learning determines difficulties around reason-responsive or rationally independent. Unless there are major breakthroughs in AI models such that AI can understand language semantically, this essay does not expect AIs to be complete moral agents any time soon.

However, this essay maintains no unbreakable hurdles in making AIs moral agents. It is well conceivable that AIs can have the ability necessary to bear moral responsibility. Moreover, when AIs develop these abilities, they may not come simultaneously. Again, from a conceptual point of view, we may see three accounts of moral responsibility as parallel specifications. An agent can have partial moral agency when they develop the ability to bear moral responsibility in one account. For example, an 11-year-old child is normally old enough to recognize an individual and be aware of the reasons behind their action. Hence the child can be a partial moral agent because it can be blamed and reasoned with. But a child does not have rational independence as they learn what their peers, schools and parents teach. Hence they are not completely moral agents because they can not be punished for their lack of autonomy. A similar situation can happen in the development of AI. With the development of recommender systems, perhaps soon, AI will be able to recognize individuals and be capable of acting towards the preferences of individuals. When this happens, AI might first be a partially moral agent in the near future before it overcomes other hurdles to become a complete moral agent.

## 3.2. Institutional moral agents

There is one influential argument for the moral agency of AIs called the responsibility gap argument (Nyholm, 2020, pp. 51-78). In general, this argument goes like this:

1. In situation X(human), the human bears responsibility for its actions in X.
2. In situation X(AI), the AI acts the same as a human does in X(human), if the AI does not bear responsibility for its actions in X, then there is a responsibility gap.[[15]](#footnote-16)
3. There cannot be a responsibility gap.
4. Therefore, AI must bear moral responsibility for its action in X.

This argument does not work because premise 3 is not true. A responsibility gap can exist. For example, both bombs and earthquakes can cause the destruction of a city. But in an earthquake, no one can attribute the responsibility for the immediate destruction of the city to any particular moral agent; but one can attribute responsibility to the bombers if the city is destructed by bombs. There is a responsibility gap between accidents and human-caused events.

However, this argument does capture the intuition that we do like the responsibility gap. We like to think that an agent who can be reached and reasoned with is in control of the event, such that we can argue with, punish, and blame this entity when something wrong happens on a massive scale. To fill in the responsibility gap, one of the responses is to invoke the concept of institutional moral agents. An institutional agent consists of other moral agents and entities such that the institution can bear responsibility as a whole.[[16]](#footnote-17) For example, the developers, operators, regulators and the AI form an AI institution that can bear the responsibility that the AI cannot bear in an X(AI) situation. So, when the action of the AI is wrong in X(AI), the AI institution bears responsibility for the action of AI. How this responsibility is to be shared among the individuals depends on the constituting structure of the institutions. For example, the senior developer designated to supervise the development might bear more responsibility than the developers because she is in control of the event and knows more about the event. Nonetheless, this does not show that the current AI can be a moral agent because it does not share responsibility in the AI institution without the ability to bear responsibility. To argue that the current AI has some collaborative moral agency is misleading.[[17]](#footnote-18)

# 4. Conclusion

In conclusion, this essay argues from responsibility that the current AIs are not complete moral agents. However, it can be a partial moral agent in the near future. For problems in attributing moral responsibility in the action of AI, one should find the AI institution for developing, running and supervising the AI. The AI institution can be a moral agent, but it does not change the conclusion that the current AIs are not moral agents.

# Bibliography

Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism .* Rowman & Littlefield Publishing Group.

Schlosser, M. (n.d.). Agency. *The Stanford Encyclopedia of Philosophy (Winter 2019 Edition)*. (E. N. Zalta, Ed.) Retrieved from https://plato.stanford.edu/archives/win2019/entries/agency/

Talbert, M. (2022). Moral Responsibility. *The Stanford Encyclopedia of Philosophy (Fall 2022 Edition)*. (E. N. Zalta, & U. Nodelman, Eds.) Retrieved from https://plato.stanford.edu/archives/fall2022/entries/moral-responsibility/

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI and Society 36 (2)*, 487-497.

1. Perhaps, it is fruitless to (emotionally) blame them because they recognize no individual other than themselves. But psychopaths can be reasoned with and punished for their misconduct. [↑](#footnote-ref-2)
2. See [Psychopaths get a break from biology: Judges reduce sentences if genetics, neurobiology are blamed -- ScienceDaily](https://www.sciencedaily.com/releases/2012/08/120816121825.htm#:~:text=A) [accessed on 27/04/2023] [↑](#footnote-ref-3)
3. See <https://brainblogger.com/2016/12/06/the-science-of-raising-a-friendly-psychopath/>. [accessed on 27/04/2023] [↑](#footnote-ref-4)
4. In the paper, Véliz argues that it would be a stretch to argue that psychopaths are not sentient. I agree with her statement. But this essay does not claim that the existence of psychopaths is a counterexample to premise (a) directly as Véliz expects. This essay only use the case of psychopaths as a counterexample against the sub-premise (a.3), which relies on the claim that psychopaths lack empathetic feelings. [↑](#footnote-ref-5)
5. This essay will not adjudicate which specifications can give the so-called true notion of agency. However, in the analysis, these specifications are parallel notions developed from the minimal notion. [↑](#footnote-ref-6)
6. A short sketch of a counterexample: a paralysed person that can only communicate with the outside world with words against the necessity of phenomenology. Intentionality might be necessary, but only so far as it sustains reasoning. But experience without phenomenology and sentience can hardly be qualified as conscious experience. [↑](#footnote-ref-7)
7. For the problem of determinism and free-will [↑](#footnote-ref-8)
8. It is now common to consider some publicity in the use of language. The most influential argument is Wittgenstein’s public language argument interpreted by Kripke. [↑](#footnote-ref-9)
9. See Huntington’s disease. But what I have in mind here is a much less severe condition. [↑](#footnote-ref-10)
10. See <https://www.youtube.com/watch?v=Y1-s37zrm1M> [accessed on 27/04/2023] [↑](#footnote-ref-11)
11. See <https://github.com/giuven95/chatgpt-failures>. [accessed on 27/04/2023] [↑](#footnote-ref-12)
12. Kantianism/Contractualism will likely not accept the forward-looking account in the first place. They argue from the assumption that moral agents are reason-responsive. [↑](#footnote-ref-13)
13. Similarly, rewards can be given for reinforcement and enhancement. [↑](#footnote-ref-14)
14. The issue is tricky. As we have the tendency to personify non-human agents, we could wrongly blame personified AI. Nevertheless, such blame is a projection. As I will discuss in the argument of the responsibility gap, this blame is best understood as blame against the AI institution. [↑](#footnote-ref-15)
15. For example, X(human) can be manual driving while X(AI) is autonomous driving [↑](#footnote-ref-16)
16. This is another counter-example to conceive sentience as a necessary condition for moral agents; an institution does not seem to be sentient as a whole, so it cannot be a moral agent on its own. [↑](#footnote-ref-17)
17. Collaborative agency overgenerates too easily to be a significant kind of agency. For example, when someone hurts others with a knife, this knife can be said to have some co-operate moral agency as it is part of the institution that bears the responsibility. [↑](#footnote-ref-18)