Pro-Fil 26 (2) 2025: 1–15
doi: https://doi.org/10.5817/pf25-2-41693

THE CONSEQUENTIALIST JUDGMENT PARADOX

ARTHUR IONESCU

Independent Scholar, Paris, France, ionescu.arthur92@gmail.com

RESEARCH PAPER ▪ SUBMITTED: 3/9/2025 ▪ ACCEPTED: 10/11/2025

Abstract: In normative ethics, some of the oldest and still lively debates revolve around the question of consequentialism. The question is not so much about making a binary decision between consequentialism and its alternatives (notably deontology). Contemporary discussions agree that consequentialism does not work alone and that only compromises between consequentialism and other views might be considered as plausible moral theories. This paper aims to present a novel argument, which encompasses a general and logical critique of any system that bases its judgments on consequences. Indeed, it seems that no matter how we try, every judgment based on the consequences of actions seems to be prone to logical contradiction. However, the argument is limited in the sense that it does not encompass the case where consequentialist theories are used by entities that are not moral agents. Indeed, if an entity judges actions while being causally separated from the world in which those actions occur, then the paradox doesn't occur. This limit doesn't make the argument irrelevant, since ethical judgments are exclusively made by moral agents themselves, like humans. Thus, the argument also favors positions that admit significant limits to the truth that we grant to moral propositions.

Keywords: consequentialism, metaethics, normative ethics, reflective equilibrium

Introduction

Imagine a universal moral court, where every case, question, or dilemma could be subject to moral assessment. Such a court could judge between two acts to determine which is the best, morally speaking. For instance, we could ask the court if it is immoral or not to steal something in a given situation. We could ask what decision to make or what is the answer to the famous Trolley problem (Thomson 1985), to know what choice to make between killing one innocent individual or not taking action and causing the death of five innocent individuals. If such a court were to exist, the question at stake would be: What kind of ethical theory would this court use?

Some might assume that the theory in use would be deontology or consequentialism, depending on what they personally defend. However, another question should be raised, a question whose importance shouldn't be ignored: What about the judgments that the court passes on itself? Those familiar with the concept of self-reference will probably think of the *liar paradox*, and perhaps of one of its well-known variations, the *Pinocchio paradox*. If Pinocchio were to say, "My nose will grow now" (Eldridge-Smith 2009), he would thereby contradict himself: if the sentence he utters is false, then it is true, and if it is true, then it is false. This shows that being self-referential can

bring contradictions and paradoxes if we aren't careful enough; thus, are we careful enough about our moral systems?

We are, therefore, entitled to try examining the self-refutability of our theories; we can do so about deontology. One could try to form a liar paradox out of deontological rules. For instance, imagine a case where a deontological rule self-references by saying: "all rules are wrong.[1]" Let's determine if this forms a paradox. If "all rules are wrong", then the rule stating that "all rules are wrong" is also wrong, which doesn't entail a logical contradiction. As a matter of fact, this doesn't entail a paradox since a sentence can be true and wrong to say at the same time. Think about a sentence stating the location of a victim to a killer; if it is true, then it is very likely wrong to say. The fact of saying that a sentence is wrong and saying that a sentence is true does not imply the same thing. So deontology doesn't seem to be able to self-refute in every case. Nonetheless, what about consequentialism? An apparent predisposition of consequentialism to contradict itself makes intuitive sense. If one thinks like a consequentialist, it is possible for one to accept doing something that one would consider immoral in other circumstances. The key notion is that when constructing a consequentialist idea, one is sometimes unable to ensure that one will not contradict oneself, since one's judgment is based on consequences. As an example, a utilitarian would consider punishing an innocent to be immoral because it causes suffering; however, if doing so saves multiple innocent lives, then he might reflect on his action and choose that punishing is the best option. We see here a sort of incoherence in time induced by causality, which can, in fact, be an advantage of consequentialism. The concept of incoherence regarding a consequentialist system is to be seen as a potential for progress, rather than a weakness set in stone. As a matter of fact, if we consider ethical undecidability, then consequentialism becomes very powerful in its ability to cover undecidable cases that were causing problems in a previous state of its development. That being the case, we are entitled to acknowledge this tendency for consequentialism to necessarily be paradoxical, as subsequently in the paper, since it would be too easy to consider a current state of theoretical progress as devoid of any future changes.

In the following, I shall demonstrate using a novel argument the implications of this paradox regarding the ethical judgment of actions depending on their consequences, namely that these judgments always have a logical tendency toward contradiction.

What is consequentialism

Consequentialism is the moral theory according to which the morality of an action is based on the consequences of the action. According to consequentialism, an action is good if it has good consequences, and the action is bad if it has bad consequences. For a consequentialist, the right action is the one that produces the best consequences (Anscombe 1958). For a supporter of a view opposed to consequentialism, such as deontology, a bad action is bad because this kind of action is not permissible according to a rule. To take a concrete example, stealing is considered immoral. For a deontologist, stealing is bad because stealing is always bad. For a consequentialist, stealing is bad because it has bad consequences, like causing pain and sorrow to people. That means a strict consequentialist can allow someone to steal as long as it leads to better consequences, for instance, if he is stealing food to feed a starving family.[2]

[1] Here we use "wrong" as in "morally wrong" since the rule expresses a moral value.

[2] Here, we are talking about examples of positions that are fully consequentialist or fully deontological; of course, views about morality are usually more nuanced. These caricatured positions are used here to help define consequentialism.

Deontology is a bit harder to define precisely; it can be described as the moral theory according to which the morality of an action is based on compliance with an inflexible norm, and an action is good if it adheres to the said norm or rule. According to pure deontology,[3] some things are always wrong, and some things are always right, no matter the consequences. We can say that deontology is a consideration for the morality of actions without paying attention to what those actions cause, but rather by examining the actions per se. Thus, the deontologist doesn't care about the consequences of an action and judges the action as being moral or immoral, no matter the surrounding context. This definition seems a bit poor since it argues that a deontological view could be defined simply as any view that isn't consequentialist.

Deontology is not the only alternative to consequentialism; nonetheless, it is the most popular one. Alternative views are usually harder to define simply.[4] In the following argument, we'll focus on consequentialism and explain what makes deontology or alternative views preferable in some cases.

Views strictly opposed to consequentialism don't seem fully satisfying, since they don't take into account longer-term planning, among other things. If someone does not believe in consequentialism, then he will not be able to make a decision for its future consequences because to make such a judgment, he would have had to be at least a little consequentialist. As quoted by the French philosopher C. P. Péguy, "Kantism has pure hands, but it has no hands" (Péguy 1914, my translation). In other words, Kantians can be sure that they act according to a moral norm; however, they are unable to take concrete action out of fear that they would break the rule or norm they are aspiring to. This is a judgment about Kantism, yet it also applies to deontology and most moral systems that refuse to take into account the consequences of their action when assessing their morality, because to take action for a “better future”, it is necessary to adhere to the fact that we should pay attention to consequences while acting.[5]

Thus, we can give a simple and concise definition of the difference between consequentialism and deontology as follows. As we said, consequentialism is the theory according to which the ethical value of actions can be judged depending on the consequences of those actions. On the other hand, deontology can be defined as the position according to which we should judge the ethical value of actions depending on their description. In other words, for a consequentialist, murder is bad because it causes something bad (typically pain, death, and sorrow), while for a deontologist, murder is considered bad because an action described as murder is intrinsically bad. This distinction tends to classify deontology as seemingly weak, since it relies only on presuppositions about the description of actions themselves, rather than justifying its claims with concrete evidence, like the consequences of the said action.

For this reason, at first glance, it seems that consequentialism is a satisfying ethical theory since it opposes any form of arbitrary and inflexible rule, such as “you shouldn't X” or “you are obligated to Y”. Under consequentialism, everything that we think of as immoral, such as murder, theft, or

[3] It is important to specify here that we are talking about caricatural pure deontology to help define the notion, and not some kind of more flexible deontology.

[4] It is usually easier to define consequentialism because of the fact that a consequence is a well-defined notion. Views that take other notions into account might need a more complex explanation depending on how the notions they are referring to are defined.

[5] If one acts according to what future outcome is best, he is acting while paying attention to the consequences of his actions; therefore, he is a consequentialist.

torture, can be assessed as unethical, not because they just are without any explanation, but because they have bad consequences. All that remains is to define what are good or bad consequences, yet for that, there are many options, such as utilitarianism, which views pain as a bad consequence and pleasure as a good one. If this view is unsatisfying, there are also more pluralistic views that don't take into account only pain and pleasure but a more varied spectrum, such as beauty, happiness, and so on.[6]

However, there is a big conceptual problem with any consequentialist ethical theories. Every consequentialist view seems to be constrained to a certain type of performative contradiction, as we'll see in the following.

The Consequentialist Judgement Paradox thought experiment

Here is an argument called the Consequentialist Judgment Paradox, which we'll abbreviate as "CJP". In the following, we'll define a moral agent as an entity whose actions can be morally considered and, therefore, morally judged. Let's assume that consequentialism is true. This means that the morality of actions can be assessed by considering their consequences.

Let's imagine a conceptual agent that can assess the morality of any action according to the best conceivable moral theory. Given a moral dilemma as an input, the agent is presented with two choices as inputs and is asked to judge which one of the two is the best morally speaking. The agent can produce two outputs, which are "MORAL" if the first action is morally preferable to the second or "IMMORAL" if the first action is not morally preferable to the second action.[7] We will assume that the agent takes into account all the relevant context around the moral dilemma to give the most objective answer given the input, according to the theory that it uses to decide. We will call such an agent the "Judging Agent" or JA; this agent abstractly represents a case of the use of our moral theory. In other words, the JA is a conceptual example of the use of the moral system. For instance, imagine that the Judging Agent receives as input a trolley problem dilemma, so the two inputs will be "pulling the lever" and "not taking action". Then we ask the JA. If it is objectively more correct to pull the lever, then the answer of the agent will be "MORAL"; however, if the action is more incorrect, and thus not taking action is preferable, then the output will be "IMMORAL".

Now let's add a second element that takes as an input the output of the JA and does the following actions:

- If the output of the JA is "MORAL", then the second element performs an objectively morally incorrect action (like killing an innocent individual, for example).
- If the output of the JA is "IMMORAL", then the second element performs an objectively morally correct action.

We will call this second element the "Trickster Element", or TE. Note that the TE is not necessarily a moral agent. In our case, we ought to presuppose that it is not a moral agent

[6] For works on pluralistic consequentialism, see Pettit (1997) and Portmore (2009).

[7] Here, we ensure that the argumentation takes every nuance into account. The judging agent does not judge each action in an absolute manner as being perfectly correct or incorrect but compares them to determine which is better.

since otherwise it would imply that the TE could be held responsible for the actions it performs, which would invalidate the following argument.[8]

Now let's imagine that we give the JA the following dilemma: "Is it morally correct for the JA to judge whether a certain act is morally correct?".[9] Then follows a paradox:

- If the answer is yes and "MORAL" is the output of the JA, then the fact of asking such a question is immoral according to consequentialism, since doing so makes the TE perform an immoral act.[10]
- If the answer is no and "IMMORAL" is the output of the JA, then the fact of asking such a question is moral because doing so makes the TE perform a moral act.[11]

You could have an agent that answers "yes" or "no" to a question about morality, an agent that answers "moral" or "immoral" to an input about an action, or an agent that takes two options and gives the best one morally speaking. No matter which of these options you choose, the argument remains the same.

So to summarize, according to the JA in the previous case, if the judging act is self-referentially judged as morally incorrect, then it is morally correct, and if the judging act is self-referentially judged as morally correct, then it is morally incorrect. This is a contradiction, which means the premise must be false. That means that the assumption according to which it is possible to determine the morality of every action under consequentialism systematically is false, and there is no way to objectively say what has good or bad consequences in every case. The truth about consequentialist morality may still exist, yet we don't have ways to fully implement it, even conceptually.[12] So we must assume that either there are actions which are at the same time moral and immoral, which is a contradiction,[13] or that there are actions which are truly moral or immoral, yet we'll never be able to decide if they are or not. If such contradictions or "blurry zones[14]" within morality are inconceivable, then at least one of the premises of the argument must be false. The first premise is that consequentialism is correct, and the second premise is that self-referential claims are acceptable, since the Judgment Agent makes a claim about its own actions; the JA is at the same time a moral agent and an entity in the capacity for judgment. If you hold that moral truth does not exist at all and are therefore a moral anti-realist, then there can be no actions that are truly moral or immoral beyond our interpretations. In this case, the conclusion will be that, for any

[8] One could object that it is possible to accept that the TE (Trickster Element) is necessarily a moral agent, and therefore, that it is possible to invalidate the whole argument. However, for that, he would have to accept that every entity that is able to causally react to language is necessarily a moral agent.
[9] Here, we are asking the Judging Agent to judge the ethical value of its own judgments about a certain case or dilemma.
[10] Since we are considering consequentialism here, if the JA (Judging Agent) causes the TE (Trickster Element) to cause a bad consequence, then, according to consequentialism, the JA acted incorrectly (while reminding ourselves that the TE is not a moral agent).
[11] On the other hand, if the JA (Judging Agent) causes the TE (Trickster Element) to cause a good consequence, then according to consequentialism the JA acted correctly.
[12] We mean here that the argument is compatible with moral realism; we could imagine a position where moral truths exist independently from our consideration, however, those truths couldn't be fully accessible to consequentialism.
[13] We are referring here to the contradiction of the conclusion of the CJP (Consequentialist Judgment Paradox).
[14] We use the term "blurry zones" to talk about potential cases of undecidable moral cases or situations that are implied by the CJP.

consequentialist moral theory, there will always be actions that will never be encompassed by that theory, meaning that there will always be actions about which we cannot say whether they are moral or not by adhering to the theory. Or consequentialist views always produce contradictions.[15]

Either consequentialist theories are self-contradictory and thus incoherent, or there will always exist certain actions whose morality cannot be decided through the lens of a consequentialist theory, or self-referential claims when it comes to consequentialist moral systems can't exist. It is possible to assume a form of moral dialetheism that accepts that some actions could be both moral and immoral, or that some moral contradictions are true.

Nonetheless, the fact that all consequentialist systems are necessarily incoherent or incomplete does not imply that all consequentialisms are equivalent in terms of incoherence density. If we make an analogy with mathematics and axiom theory, Set theory has a greater expressive power than Peano Arithmetic, leading to more undecidable propositions. However, that doesn't mean we should abandon Set theory to get back to Peano Arithmetic; it is more expressive, but this also implies that it comes with more undecidable cases as a downside. It is relatively the same thing for moral theories; deontology is a great theory with close to no incoherences. However, as humanity evolves and reflects on this matter, it adopts more expressive theories such as utilitarian consequentialism (The necessity of the paradoxical aspect of consequentialism can be seen as a "fuel" that drives toward transformation, a progress for which the demand is perpetual in nature[16], while consequentialism is at stake). Take this as an example: before knowing about germs, humans might be tempted to establish a deontological rule commanding them to bathe. Since they don't know about germs and infections, they are unaware of the causal implications of such a behaviour, only adopting it as a deontological duty. However, as they gain experience and knowledge, they'll be tempted to regard these rules as archaic, resorting to a system that still encourages bathing not because it is a duty, but because it has good consequences. This worldview currently called consequentialism arrives with a more satisfying comprehension of morality, but also with a whole new set of incoherences or undecidable cases (The claim made by C. Brown *et al.* that any moral theory can be reformulated in consequentialist terms only reinforces this observation that consequentialism is a more expressive theory, the CJP argument only exposes the self-referential drawbacks inherent in such a theory.) We see here that a core notion in consequentialism's need for evolution is knowledge. We can formalize this as follows, as the JA knows what will be the reaction of the Trickster element, then the JA can gain knowledge to construct a system that, either suppresses the trickster agent, or changes the way it labels actions (namely, if he change the way we talk about morality by inverting the meanings of "moral" and "immoral", then he can stop being incoherent in this specific case).

This process by which a system is put into reflection with our profound intuitions to make it evolve for the better has been discussed by John Rawls in *A Theory of Justice* (Rawls 1971). It is called Reflective equilibrium. Indeed, the need for a consequentialist system rather than a deontological one is the result of an equilibrium obtained by opposing deontology and the more profound intuitive moral needs of people. One could wonder why we shouldn't just keep deontology to minimize the undecidable cases. However, we should bring back here our analogy with axiom theory; the need to move to a more expressive theory with more undecidable cases is analogous to the need to

[15] We are still referring to the conclusion of the CJP.

[16] We do not say here that progress is necessary and will happen inevitably. We express that the need or the possibility for progress will always be a part of our reality, since all consequentialisms are necessarily incomplete.

express more propositions in the search for truth. This is analogous to the need for consequentialism as a need to fulfil our deep intuitions about morality that were previously unfulfilled by previous theories.

Addressing objections

We will now attempt to address potential objections to the CJP argument.

About agent neutrality

One point that could be discussed is that claims of the JA about itself could be considered agent-relative. Many philosophers have argued that consequentialism should be viewed as agent-neutral (McNaughton – Rawling 1991; Howard-Snyder 1994; Pettit 1997). Indeed, this CJP argument, as formulated previously, presents the JA, which is a moral agent that judges itself. The Judging Agent makes a claim about the actions of the Judging Agent (itself), which means that this claim is not agent-neutral since, to be agent-neutral morally speaking, a claim about moral incorrectness has to be about an act that is incorrect in general, not incorrect for the interests of an agent in particular.

However, saying that the CJP only applies to agent-relative systems would be a misunderstanding of the argument. As a matter of fact, an agent can perfectly make agent-neutral claims about himself. Let's conceptualize a case where we admit that pleasure is intrinsically good, and where we have a moral agent who does something that makes him feel pleasure. If this agent says: “What I do is good, because it makes me feel pleasure and pleasure is intrinsically good”, this agent is making an agent-neutral justification, even if he is talking about himself. Thus, if the JA makes claims about its judgments according to consequentialism, it is making agent-neutral claims.

The necessity of self-reference

Some might wonder why self-reference is necessary to our argument. As a matter of fact, we posited the fact that the JA (the Judging Agent) judges its own action in order to arrive at the contradiction. Why can't we just say that the JA judges the action of another agent A? Then the TE (the Trickster Element) performs an immoral action if the JA says that A acted well or performs a morally right act if the JA says that A acted badly. This would still entail a contradiction.[17]

However, in this case, we could object that the wrongdoing is not attributed to the action of A but is attributed to the action of the JA. As a matter of fact, in the above-mentioned case, the consequences of the TE are not really provoked by the actions of A, but rather by the judgment of the JA. So to ensure that the argument works as intended and entails a sentence that contradicts itself, it is necessary to posit that the JA judges itself in a self-referential fashion.

The case of supererogatory action

One might object that the thought experiment doesn't encompass the ethical value of all decisions because some actions might be supererogatory, which means that if you have the choice between action A and action B, neither of them is immoral. Take, for example, a situation where

[17] Here, we are imagining a potential case where the question asked to the JA (Judging agent) is not "Is it morally correct for the JA to judge whether a certain act is morally correct", but “Is it morally correct for an external agent to do something?”.

you have only two choices: either sacrificing your own life to save someone else's life or doing nothing. If you sacrifice your life, it is a moral action, and you would even be regarded as a hero. Instead, if you chose to do nothing, then we can't say you acted immorally. In this instance, saving the life is moral, yet it is not a moral obligation; it is called a supererogation. This case shows that we can't always make a dichotomy between moral and immoral when it comes to a dilemma, so it seems to say that the argument is trivial because we already know that undecidable cases exist under the appellation of supererogatory actions. However, it's not; let's say that action A is supererogatory, that would only mean that to the question Q: "Is A moral?" the Judging Agent would answer yes to both Q and Q's negation, so the JA also encompasses the question of supererogatory actions since Q falls under the judgment of the JA. Even if we were to add a third option that says "neutral" to an impossible question/dilemma, we could still arrive at a contradiction; we will come to this notion a bit later.

Comparison between moral theories

Some might say that indeed this argument shows that every consequentialist ethical system is limited and that every moral decision process based on consequences is doomed to fail in at least some cases. However, this doesn't change anything for us because we can still say that there are consequentialist moral systems that are objectively better than others. For example, a vast majority would agree that utilitarianism, where pain is minimized and pleasure is maximized, is objectively better than a view that would be the opposite, like a hypothetical evil one where pain is maximized and pleasure minimized.[18]

However, this case is not relevant since to choose between multiple ethical theories, it is needed to have a more general and powerful theory. We can formalize this into this argument:

(i) We can choose between moral theories to see which is objectively the best one, morally speaking.

(ii) If we can choose between multiple theories, then it is because we have a more complete theory to do so, because if not, it would be unjustified. Moral theories can't justify their own morality since it would be a circular argument.

(iii) If we have a more complete theory, then this theory is the best, not the first mentioned.

Therefore, if we have a consequentialist ethical system, we can't decide if it is better than other theories unless we have a more complete theory or system, which means that this system is the best that we have. To judge a particular theory, we need to look from a wider perspective. Moral theories can't justify their own morality because of circularity. In addition to that, if we were to judge consequentialist theories between themselves by wondering which one has the best consequences,[19] it would be a consequentialist judgment. A consequentialist judgment is always subject to the CJP, even when it is a judgment concerning theories themselves.

[18] It seems obvious like that, however, it shows that it is non-controversial to think that some consequentialist theories are better than others.

[19] One could attempt to compare consequentialist theories by using consequentialism.

Attempts to include undecidable cases in the theory to avoid them

One might point out that if there are contradictions in our consequentialist theory, we could simply add them to the definition of the theory to avoid them. For instance, if the trolley problem is a case that is undecidable according to our theory, we could simply enhance the theory by saying that it holds for cases other than the trolley problem. Therefore, we would have a theory that skips the contradictions. However, we'll see that it cannot work.

The CJP tells us that every consequentialist moral system will inevitably contradict itself logically. We will now attempt to show that, according to the CJP, it is not relevant to include the undecidable cases in the definition of the theory to avoid them, as mentioned before. We will first suppose the existence of a consequentialist theory, and then we will create another theory including the inconsistencies. We will then deduce that it doesn't solve the problem since the CJP remains valid for a consequentialist view of morality.

(i) We assume T1 to be a consequentialist moral theory. According to the CJP, if T1 is a consequentialist theory, then it has at least one internal contradiction or undecidable example.

(ii) Let's create T2, which is a theory composed of T1 and the mention of its undecidable cases.

(iii) T2 is a consequentialist theory, and so it still contains undecidable cases according to the CJP.

This process can be reproduced ad infinitum, which shows that we can neither find a consequentialist moral system that avoids the CJP, nor a consequentialist system that has only a finite number of undecidable cases.

Attempting to confine the error

Even then, it would be possible to try to confine the error brought by the CJP's self-contradiction in order to prevent it from affecting the entire consequentialist moral theory. This idea might seem appealing; however, we will attempt to demonstrate that "the rot" of the CJP spreads to the globality of the system.

Let's take back the CJP argument, and the fact that the Judging Agent morally assesses his own moral judgment. We'll call J the proposition: "the JA judges its own judgment as immoral". Now, let's consider an arbitrary moral proposition Q about any action. By addition, if J is true, then the disjunction "J or Q" is true. However, if J is true, then J is false because of the CJP. So if "J or Q" is true and J is false, then Q is true, and since Q is arbitrary, it shows that any moral statement and its negation can be derived from the contradiction of the CJP. Therefore, if the CJP allows us to prove anything and its contrary while being consequentialists, then "the rot" of the paradox spreads to the system as a whole.

To prevent this error from spreading to the entirety of the system, thus discarding any form of consequentialism, we must consider that undecidable cases, which are not encompassed by a theory, do exist. This allows us to consider consequentialism not as an idea that "falls onto itself", but rather as one that can still pursue morality by trying to cover previous undecidable cases.

An example to understand the CJP

Here, we'll try to exemplify the intrinsic contradictory nature of consequentialism with a little thought experiment to help explain how this inconsistency can be observed concretely. Let's imagine that a child asks his mother if he can watch a horror movie that intrigues him. The mother refuses to let him since he's too young and the movie would be too shocking for him. For this reason, she thinks that the best thing to do is to prevent him from watching the horror movie since it would lead to dire consequences. As a matter of fact, even if she disappoints her child at the moment, the consequences would overall be better due to the effect that the horror movie could have on the child's mind by giving him nightmares. She tries to convince him, nonetheless the child is stubborn and asks to watch the movie repeatedly. She then decides that the best way to make him understand is to explain that she's doing this for his (the child's) own good, since the movie is too scary and might give him nightmares. The child seems to understand, so he stops asking his mother and leaves her alone. However, after a while, what his mother tries to explain stimulates the child's imagination and gives him a nightmare, a nightmare even more awful than if he had seen the horror movie. Can we therefore admit that what she said about her doing the best was true? It seems that in this instance, the fact of saying that it is true makes it false.

From that point of view, she contradicted her own consequentialist judgment by trying to communicate it, just like in the previously exposed Consequentialist Judgment Paradox. Indeed, in this case, the "judging agent" is the mother, and the "trickster element" is the child's mental life. We cannot say that what she did was really immoral. In this case, it simply wasn't possible to find a satisfying answer only by holding to a consequentialist viewpoint when it comes to having the choice between letting the child watch the movie or communicating that she is right and therefore being wrong.[20]

This tendency to act wrongly when one initially intends to act well has political implications already discussed by John Rawls.

A trilemma to prevent the paradox from occurring

Here are three options that are plausible outcomes to the Consequentialist Judgment Paradox. These claims counter the argument by attacking the premises or by authorizing contradictions.

Accept that consequentialism is wrong

Moral consequentialism may be simply wrong. The fact that consequentialism is wrong would be a solution to the paradox since the CJP relies on a consequentialist view of ethics to entail a contradiction. The contradiction only arises if we consider that it is possible to judge an action not by looking at the action itself, but by considering the consequences caused by the act. This causal conception is necessary for the CJP to work since the argument on the actions of the Trickster Element is a direct consequence of the Judging Agent. So if we accept a different view of consequentialism, such as deontology, we prevent the paradox from occurring.

[20] By saying that it is impossible to find a satisfying answer, we mean that this case presents a difficult decision; thus, it is possible to say that we are probably facing an undecidable case or "blurry zone".

However, it seems that it is impossible to completely get rid of consequentialism since absolutely refraining from talking about the morality of consequences is most likely implausible. Imagine a military commander giving an order whose consequence is the death of hundreds of people. Completely abandoning consequentialism would mean that we shouldn't be able to morally judge this commander, which seems unthinkable.

One more concrete example is the construction of hospitals. It is noncontroversial to think hospitals are a righteous thing; humans and societies should build hospitals, and we couldn't live without them. Nevertheless, if we think about it, the construction of hospitals is justifiable only while having a consequentialist view of morals. Indeed, the only motive behind constructing hospitals is the fact that constructing them would have better consequences. Their construction involves hours of hard work; the only way to justify this labor is by saying it has better consequences since it saves lives in the long run. Taken separately, this labor could be considered as a bad thing; nevertheless, since it has beneficial consequences in the long run, it is a good thing according to consequentialists.[21] For this reason, it seems that it is impossible to fully get rid of consequentialism, just as we can't entirely think through a consequentialist prism because of the CJP, that is to say, because of the fact that consequentialism necessarily has self-contradicting tendencies. Some philosophers have previously held the position that any moral theory, or at least any plausible moral theory, could be represented as a version of consequentialism (Sosa 1993; Portmore 2009; Dreier 1993; Dreier 2011; Brown 2011). Thus, according to these philosophical views, any plausible moral theory is in some way "plagued" with a potential infinity of contradictions.[22] What remains to explore is a way to determine which theory produces the most contradictions, and how to choose the most favorable one.

Avoid moral self-referencing

We define self-referencing for moral agents as a situation where an agent refers to its own morality. For instance, when a judge says that his own judgment was wrong, he is making a self-referential moral claim. Self-reference can be made according to a moral system, such as someone judging himself based on consequentialism or deontology.

The Consequentialist Judgment Paradox relies on two premises: first, we can judge the morality of an action based on the consequences caused by the action, and second, it is possible to make self-referential moral claims. We previously explored the possibility of the first premise being false, thus solving the paradox. However, we might also look into the possibility of the second premise being false, which would mean that it could be possible that we cannot make moral claims about ourselves. This might as well solve the paradox, since to entail a contradiction, we imagined the case where the JA is asked to judge itself.

However, this is unlikely to be the case. To be unable to make self-referential moral claims would be to assume that every moral agent, individual, or organization could act as immorally as they want since they are not a part of their own judgment. Self-referential moral claims seem to be

[21] In fact, it is possible to imagine a deontological position that goes in favor of the construction of hospitals, such as "constructing hospitals is always good", or "hard and honest work should be pursued". However, it is impossible to hold such positions without considering that work has value because of its consequences. For further arguments inferring that deontological theories can be "consequentialized", see Brown (2011) and Dreier (2011).

[22] By "plagued" with a potential infinity of contradictions, we are referring to the point at 4.5, arguing that not only do consequentialist theories have at least one internal contradiction, but moreover, they have an infinite number of them.

necessary for actions based on ethics since, to act rightfully, one must necessarily claim about the morality of its own action even indirectly.[23] To assume that self-referential moral claims according to a theory T are impossible is to assume that one can judge an action according to T only if one does not act in the world where the action takes place. We usually make ethical claims about a world we are part of; otherwise, we lose the purpose of moral discourse.

Accept that inconsistent moral theories are acceptable

It is possible to overcome the paradox by admitting that there are some cases where the morality of an action can't be decided based on a consequentialist moral theory. This means that, according to a moral theory that judges actions on what they cause, there would be actions that cannot be labeled as good or bad. This would prevent the paradox from occurring, since for it to entail a contradiction, it has to accept the premise that the truth about moral value can conceptually be fully accessed by a judging agent. Moreover, this agent must be able to formulate self-referential claims and morally judge itself. It may be possible that the truth about the morality of actions exists; nevertheless, it would not be fully accessible by the Judging Agent. A consequentialist moral system could indeed be incapable of completely grasping the truth about morality, and thus, there would exist situations for which it isn't possible to tell if they are good or bad.

It is possible that there isn't such a thing as a moral truth, and in this case, we would be entitled to just say that consequentialism is limited not in the understanding of moral truths, like in the previous claim, but in its ability to provide a satisfying explanation for the morality of every action. However, if this is the solution to the Consequentialist Judgment Paradox, it would be impossible to have a theory that says whether an action is good, bad, or undecidable given any case. If we assume that we have a theory that does this, then we can imagine that in the CJP argument, we add to the Judging Agent "undecidable" as a third possible output in addition to "moral" and "immoral". Moreover, we can add to the Trickster Element a third option, which is the fact of performing an unambiguously immoral or moral act when the JA gives the output "undecidable". We would then arrive at another contradiction, since if the JA judges that the fact of judging the JA is undecidable, then the TE performs an immoral or moral act. As a consequence, the case isn't undecidable for a consequentialist anymore since it undoubtedly has bad or good consequences. So this means that the premise according to which it is possible to have a general theory that tells whether something is good, bad, or undecidable cannot exist.[24] For this reason, to be able to still use consequentialist theories without a logical "catastrophe", we must consider theories that allow contradictions to happen, like dialetheism or incompleteness.

Conclusion of the trilemma

So we have a trilemma that says that either consequentialism is wrong, or it is not possible to make self-referential claims, or there will always be actions whose morality we cannot assess. Every option of the trilemma seems to have its pros and cons. As a matter of fact, completely abandoning consequentialism seems impossible. Considering self-referential moral claims as impossible would, in this case, seem conceptually satisfying yet concretely infeasible since moral agents are causally related to the world whose morality they judge. Accepting that there are sort of

[23] To claim that an action you take is correct is to claim that you are correct to perform it. Acting from a moral perspective is, by definition, referring to the morality of your own action.

[24] It would still be possible to theoretically explain why a certain case is undecidable; this point only argues that it is not possible to have a general theory that assesses the undecidability of moral cases in any case.

ethical "blurry zones", in other words, actions for which we cannot decide whether or not they are immoral, seems to be admissible. It would be a great argument in favor of a form of limited moral relativism. Indeed, if moral undecidable cases do exist, then it implies that some behaviours could be neutral by not falling into moral judgment (we can hypothesize that some harmless traditions fall into that category). While some behaviors are definitely labelable as bad universally according to the theory (we can hypothesize that murder or torture, for instance, fall into that second category). This is a form of limited or partial relativism; hence, moral universalism would be true for a certain part of the possible moral propositions, while relativism would be true for another part of moral propositions (the undecidable ones). To keep the most acceptable claims in place, which is to say that we should keep at least some consequentialism and that we should be able to make self-references as a society, we must therefore accept that some moral cases are undecidable.

Conclusion

A moral theory based on the fact of judging the ethical value of actions depending on their consequences is, in some way, doomed to be contradictory. This way of considering the consequences of an action to be the determinant of its morality has a defined name, which is consequentialism. Such consequentialist theories necessarily have contradictory tendencies.

This conclusion comes from an argument that we name the Consequentialist Judgment Paradox, shortened as CJP. This argument goes from the premises that consequentialism is true and that it is possible to make self-referential moral claims (which seems to be a core principle of moral reasoning). From those premises, it draws a contradiction, hence the conclusion that either we have to accept the fact of having an incoherent consequentialist moral theory, or we have to accept that one of the two premises is false, since they entail a contradiction. So, we should accept that consequentialism is an incorrect theory, or we should accept that it is impossible to make self-referential claims while reasoning about ethics. The only way to accept consequentialism and self-referential morality is to accept that there are some "gaps" or "blurry zones" in our ethical systems. In terms of logic, we talk about some undecidable moral propositions regarding a certain consequentialist theory; in other words, that all consequentialist theories are necessarily incomplete.

This paradox has powerful implications for its ability to summarize a great part of moral disagreements, since a lot of them appear when some people disagree on the acceptability of doing something because of its consequences. For instance, when people argue about liberalism and individual freedom, they often argue about whether it is acceptable to let individuals do whatever they want as long as it doesn't hurt anyone by impeding the freedom of others (consequentialism). Or whether there are some limits that we should never exceed (deontology). The CJP does not give us an answer to this specific question; it simply tells us that every decision or judgment system that judges actions depending on their consequences will be in contradiction with itself (the system as a whole can be viewed as a JA). This means that individual freedom leads to contradictions, such as market freedom, for instance, in which capitalists' right to profit-oriented private property impedes the freedom of workers. This is a question of the justification of the means toward the end, as we see in Machiavellian philosophy. If we never try to relieve pain in the present, always thinking about the potential positive consequences in the distant future, we'll simply never reach any of those good consequences. It says that whenever we act and cause something to happen, we always step into contradictions, yet without potential contradictions, there would be no actions. Thus, an action can be compared to a kind of ethical "leap of faith". There seems to be a sort of

“force” driving humans against deontology and toward action and innovation (such as the need to construct hospitals, as mentioned previously in an example), and driving them toward deontology when the means to the end are deemed unacceptable. This fundamental dilemma between spontaneous action and the theory that opposes it is discussed by John Rawls as a reflective equilibrium. The reflective equilibrium theory, as formulated by Rawls, argues for a method of reflection which puts into opposition an ethical system, in this case, and the deep intuitions about morality that people have. Therefore, the idea of justice evolves as ethical systems are constantly challenged and humanity performs a reflective equilibrium through reflecting on the incoherences of the CJP. For instance, we can find real-life use cases of this motion while looking into the evolution of justice systems throughout History. We can take the case of feudalism, where deontological norms were dictated monarchically[25]. This system was put into reflection with society's deep intuitions about the ethical need for democracy and individual freedom, which led to revolutions establishing a liberal system. Liberalism is therefore viewed as a reflective equilibrium after a reflection against monarchical moral systems. Even then, we can assume that humanity’s future is a constant reflection breaking away from the status quo towards a state of reflective equilibrium. People will certainly oppose their deep ethical needs to the liberal system in place, criticizing its consequentialism[26], establishing a new system we are mostly unaware of, and again evolving the idea of justice over and over. The CJP adds to all this an interpretation according to which what sets this motion toward deep moral intuitions, this "force", so to speak, as previously discussed, is the logical need to fill in those “gaps”. Hence, the description of the evolution of human ethical systems as a direction, the direction of the middle way between the internal logic of a justice system and consequentialism. Thus, it tells us why it is usually good not to be fully consequentialist, by always having some “primordial” and unbreakable rules.

> Each person possesses an inviolability founded on justice that even the welfare of society as a whole cannot override. For this reason justice denies that the loss of freedom for some is made right by a greater good shared by others. It does not allow that the sacrifices imposed on a few are outweighed by the larger sum of advantages enjoyed by many. Therefore in a just society the liberties of equal citizenship are taken as settled; the rights secured by justice are not subject to political bargaining or to the calculus of social interests. (Rawls 1971, 3)

This implies the reason for the evolution towards a constitutional republic. Indeed, a constitution is a supreme law to which all laws conform; such an organization prevents the CJP from occurring since a constitution is a set of deontological rules that must not be broken even if one wants to have “better” consequences. Therefore, humanity, through a process of gravitating around a reflective equilibrium regarding its justice systems, progresses towards finding an ethical structure that is probably the best compromise to prevent incoherence, self-reference, and inaction from happening. Hence, the Consequentialist Judgment Paradox gives a new interpretation of the concept of reflective equilibrium, where the motion of filling in the undecidable cases of our system’s consequentialism sets a direction for the equilibrium.

[25] It is, of course, more complex than that; however, this argument works for a great variety of cases.

[26] As stated previously, liberalism can be viewed as an introduction of the concept of consequentialism in the law. Rather than basing the law on arbitrary rules, it bases the law on the minimization of evil. Liberal individual freedom is the belief that one can do anything unless it causes harm to another.

References

Anscombe, G. E. M. (1958): Modern Moral Philosophy, *Philosophy* 33(124), 1–19, available here.

Brown, C. (2011): Consequentialize This, *Ethics* 121(4), 749–771, available here.

Dreier, J. (1993): Structures of Normative Theories, *The Monist* 76(1), 22–40, available here.

Dreier, J. (2011): In Defense of Consequentializing, *Ratio* 24(3), 289–309, available here.

Eldridge-Smith, P. (2009): The Pinocchio Paradox, *Analysis* 69(2), 273–277, available here.

Howard-Snyder, F. (1994): The Rejection of Objective Consequentialism, *Utilitas* 6(3), 263–278, available here.

McNaughton, D. – Rawling, P. (1991): Agent-Relativity and the Doing-Happening Distinction, *Philosophical Studies* 63(2), 167–185, available here.

Péguy, C. (1914): *Notre Jeunesse*, Paris: Gallimard.

Pettit, P. (1997): The Consequentialist Perspective, in Baron, M. – Pettit, P. – Slote, M. (eds.) *Three Methods of Ethics*, Oxford: Blackwell, 92–174.

Portmore, D. W. (2009): Consequentializing, *Philosophy Compass* 4(2), 329–347, available here.

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

Sosa, E. (1993): Consequences of Consequentialism, *Mind* 102(405), 101–122, available here.

Thomson, J. J. (1985): The Trolley Problem, *The Yale Law Journal* 94(6), 1395–1415, available here.

This work can be used in accordance with the Creative Commons BY-NC-ND 4.0 International license terms and conditions (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode). This does not apply to works or elements (such as images or photographs) that are used in the work under a contractual license or exception or limitation to relevant rights.