Linköpings universitet | Institutionen för kultur och samhälle
Kandidatuppsats i praktisk filosofi, 15 hp | Ämneslärarprogrammet
Vårterminen 2022

Philosophical Zombies and Moral Responsibility

– An Analysis of Whether Philosophical Zombies Would Have Moral Responsibility or Not

Rim Winssi

Handledare: András Szigeti
Examinator: Martin Berzell

Linköpings universitet
SE-581 83 Linköping
013-28 10 00, www.liu.se

Abstract

Philosophical zombies are beings that look exactly like humans and behave in the same way as humans do. The only difference between humans and philosophical zombies is that philosophical zombies lack consciousness. This means that they can complain, cry, laugh and say that they are in pain. However, emotionally, they will never experience these feelings. Philosophical zombies have no desires, no values, and no empathy. Despite philosophical zombies lacking all these qualities, the question can be raised whether, if they were to exist in our world, would they have any moral responsibility? This question becomes pressing because even though philosophical zombies feel nothing and lack consciousness, they are still capable of doing harm and able to act immorally.

By using David Shoemaker's (2015) 'Tripartite Theory of Responsibility', I will in this essay analyse whether philosophical zombies are eligible for moral responsibility, and if so what type of moral responsibility they would be eligible for, i.e., whether it would be attributability, answerability or accountability. Furthermore, this essay will discuss if philosophical zombies and psychopaths are similar, and whether they are meant to be qualified for the same moral responsibility types, and if so, which type that would be. Additionally, this essay will discuss the dilemma that might arise if philosophical zombies are not suitable of being moral agents and bring up the debate of the moral agency of AI's. Thereby coming to the point if philosophical zombies are not fit to be held morally responsible.

Keywords: Philosophical zombies • Consciousness • Moral responsibility • Moral agency • The Tripartite Theory of Responsibility • Attributability • Answerability • Accountability • Psychopaths • AI

Table of Contents

1. Introduction

During the late 20th century and early 21st century, the concept of minds and consciousness have become popular topics in philosophy. Two central philosophers who discussed consciousness and the concept of mind were David J. Chalmers and Robert Kirk. In order to explain and analyse these two philosophical subjects, but also to argue against various forms of physicalism such as behaviourism and materialism, they came up with a thought experiment: the philosophical zombie, and the so-called "zombie argument" (Stanford Encyclopedia of Philosophy, 2021). Philosophical zombies, or p-zombies for short (*I will be using both names interchangeably*), differ from the familiar zombies that appear in popular culture and movies. The latter are seen as undead, human-eating monsters. However, philosophical zombies are significantly different in many aspects. David Chalmers describes these philosophical zombies, in his book *The Conscious Mind: In Search of a Fundamental Theory* (1996), as similar to humans in all physical aspects, from molecules to body parts. From the outside, a philosophical zombie looks just like any other human. So, what is the difference between a philosophical zombie and a human? A human would perhaps observe a painting and think to themselves that it is pretty or eat a piece of chocolate and find it sweet. A philosophical zombie on the other hand would perceive the painting, but only do it in a functional sense and only in the psychological sense will they taste the chocolate.

What this means is that the philosophical zombie would have no personal experiences of anything at all, there is nothing that defines them as they have no experience and no consciousness, "[t]here is nothing it is like to be a zombie" (Chalmers, 1996: 95). The crucial claim here that for any entity that has consciousness there is something it is like to be that entity (Nagel, 1974: 436). What this means is that each creature or being with consciousness have qualities that defines them, and "(…) a mental state is conscious if there is something it is like to be in that mental state" (Chalmers, 1996: 4). This entails that if a mental state is to be conscious, it has to have a qualitative feel or quality of experience. All these qualitative feels are phenomenal qualities known as *qualia* (4).

Chalmers, as several philosophers before and after him, uses the concept of philosophical zombies to analyse what the phenomenal qualities of consciousness are. More broadly, he also uses this concept to prove that not everything in this world is physical and that consciousness cannot be something physical. Although this is the actual reason behind the zombie argument, it is not the way philosophical zombies will be used in this essay. Instead, philosophical zombies

will be used to explore the nature of moral responsibility. So, how can one draw a connection between philosophical zombies and moral responsibility?

In the book *Responsibility from the Margins* (2015), David Shoemaker develops the so-called 'Tripartite Theory of Responsibility' to analyse whether psychopaths, amongst other people with mental difficulties, can be held morally responsible or not. The theory distinguishes three types of responsibilities which are separated by so-called "qualities of will": **attributability, answerability** and **accountability**. Different qualities of will correspond to these three responsibilities. The first type of responsibility, *attributability*, evaluates one's quality of will in terms of what it says about one's "character". Specifically, to be an attributable agent, one must show care or behaviours that are reflected by ones so-called "inner self". What this implies is that if one, for example, is drugged, but not by their own free will, and does damage of any kind, then they cannot be held morally responsible in the sense of attributability as their behaviours did not reflect who they actually are (Shoemaker, 2015: 35, 59). The second type of responsibility, *answerability*, tracks the quality of will in terms of "judgements" that the agent makes. To be held answerable, one must then be capable of making judgments and being able to see so-called "instead of" reasons. What this means is that one can act in a certain way, which would be better, preferable, or more morally correct, *instead of* acting in a way that would perhaps cause damage or behave in an immoral way (Shoemaker 2015: 35, 82). The final type of responsibility is *accountability* and aspect of the quality of will this type of track is "regard" (35). According to Shoemaker, one would be accountable if one is appropriately held responsible for their actions. What this implies is that "[o]ne is an accountable agent just in case one is liable for being a fitting target of a subset of responsibility responses to one (…) in virtue of one's quality of regard" (113).

What I have now presented is a summary of the three types of responsibility in Shoemaker's 'Tripartite Theory of Responsibility'. As mentioned previously, Shoemaker considers the case of psychopaths, and whether they are eligible for any type of moral responsibility. In some cases, it seems as if philosophical zombies and psychopaths have similarities: they both lack empathy, have trouble identifying empathy in others, and have other emotional deficiencies. So, the question arises, if psychopaths can be held morally responsible for their actions, and philosophical zombies have similar traits to psychopaths, can they also be held morally responsible?

More than the similar traits with psychopaths, when talking about p-zombies, it is easy to understand why this topic is being written about when one makes a connection to AI's. There are several debates that have been going on for years whether AI's can be held morally

responsible or not. One of the many that discusses the agency of AI's is Carissa Véliz (2016, 2021). Moreover, Véliz also makes a comparison between the moral agency of both AI's and philosophical zombies. (2021). This comparison and dilemma will be discussed later on in the essay as it is first important to look at moral responsibility from a theory to further understand if it is even possible for p-zombies to be eligible for moral responsibility.

Now, there is a vast number of theories about moral responsibilities but for this essay I will choose to mainly focus on Shoemaker's tripartite theory. The aim of this essay is therefore to establish whether philosophical zombies would have any moral responsibility according to Shoemaker's theory. The purpose of this essay is therefore not to analyse whether philosophical zombies exist nor am I interested here in the concept of consciousness. Furthermore, the aim of this essay is not to discuss either what the purpose of philosophical zombies might be or how they were created to end up with no consciousness. Instead, philosophical zombies will be used in a different thought experiment which is this: if they were to hypothetically exist, would they have any sort of moral responsibility based on Shoemaker's 'Tripartite Theory of Responsibility'?

1.2 Purpose

The reason as to why I chose to write about the subject of philosophical zombies and moral responsibility began in an interest in philosophical theories about 'other minds'. That path led me to reading about philosophical zombies, this subject was new to me but intriguing. When reading about philosophical zombies I started to think about if there were similarities between them and psychopaths. In addition to this, I thought about the dilemma of if we should treat AI's in the same way that we treat other humans. From this thought-process, I started to ponder about how we would treat philosophical zombies if they were to exist and if they then would have any moral responsibility. One might wonder as to why writing about this subject would be necessary or to any interest to philosophers and others. However, as the topic of other minds alongside the subject of moral responsibility for AI's are frequent topics in philosophy, I wanted to further discuss these topics. The connection between p-zombies and AI makes this thesis further philosophically relevant. Additionally, seeing as other philosophers as Véliz have already reflected on this topic, it seems to me that the topic is of interest to philosophers and that discussing different beings and their eligibility for moral responsibility is of relevance for us to understand the aspect of what it takes for a being to be qualified for moral responsibility.

In sum, the purpose of the essay is to analyse and examine whether philosophical zombies would have any moral responsibility based on Shoemaker's 'Tripartite Theory of Responsibility' if they were to exist. The essay will ask not only if philosophical zombies would or would not have moral responsibility, but also which type of responsibility it would then be that they would be eligible for, whether it would be attributability, answerability or accountability. Furthermore, I will discuss the difference between philosophical zombies and psychopaths, and if they are indeed similar in any way. Finally, I am going to discuss criticisms to the effect that philosophical zombies cannot be moral agents and the problems and dilemmas that would thereafter arise if they were not morally responsible.

2. Philosophical zombies and moral responsibility

2.1 Philosophical zombies and the "Tripartite Theory of Responsibility"

In this chapter, the essay discusses whether philosophical zombies would have moral responsibility or not based on Shoemaker's "Tripartite Theory of Responsibility". The essay will use the three different "qualities of will" that Shoemaker establishes in his theory and one by one analyse if the philosophical zombies would be eligible for any of them.

Following the earlier definition, philosophical zombies are beings that look exactly like every other human, there is no physical aspect of a p-zombie that would make them different from humans, and thereby one is not able to tell them apart from humans. In fact, there are no behavioural aspects that would distinguish them either. Philosophical zombies are configured to look and behave exactly as everyone else: they are indistinguishable from humans in every physical aspect and form of behaviour. Comparable to humans, philosophical zombies can complain about the weather, cry when watching a sad movie or laugh when watching a happy one, they can discuss politics and philosophy or even comfort you when you are upset. The only difference between philosophical zombies and humans is that they lack conscious experience. Despite complaining about cold weather, they will not *be* cold or even if they cry when watching a sad movie, they will not *be* sad (Véliz, 2021: 487). Nothing upsets a philosophical zombie, and nothing excites them, there is nothing it is like to be a philosophical zombie. Despite the fact that they physically seem to act as if they experienced everything that a human being does, they do not have any emotional or physical experiences. Philosophical zombies will in fact not know whether the chocolate is actually sweet despite eating it and saying that it is sweet, they will neither never know the pain of burning one's hand on the stove top. They will act as if they are experiencing all of these things but will feel nothing at all, for they lack the basis of a consciousness (Chalmers, 1996: 95).

But regardless of the fact that they have no consciousness, they can still perform actions that could be damaging and immoral. They might hurt someone and yet feel no guilt, no remorse, shame, or anything for that matter. What this means is that they could end up taking a life and not feel a single thing about it (Véliz, 2021: 488). Even though they have no consciousness, should they not be held morally responsible for their actions? If they could be eligible for any of the responsibility types mentioned in the tripartite theory, will be further discussed now.

2.1.1 Attributability

The first moral responsibility type is *attributability*. To be an attributable agent, one must perform actions that reflect on one's inner self or deep self. That means that if one for example accidentally does something or is forced to perform an action that they do not desire to do, then they cannot be held responsible for that action in an attributable way, i.e., they cannot be an attributable agent in that respect (Shoemaker, 2015: 59). One would for instance not blame a person for accidentally stepping on their foot in a crowded subway as long as it was not their intention to do so (7-8). As an attributable agent, one would experience so-called characteristic sentimental syndrome pairs. These are agential admiration on the positive side and agential disdain on the negative side. These are then further divided into several emotional responses as esteem, awe, and pride on the positive side and disdain, hatred, and shame amongst other emotions on the negative side. These are some of the aretaic responses which attributable agents may experience (39). To be held attributably responsible for one's actions, one must perform actions that reflect one's deep self but also be able to show that one cares, and one has commitments. Therefore, agents can only be attributable agents if they can show that they care through emotions, attitudes, and actions. This would then further imply that for one to be an attributable agent, one must be able to show a wide range of emotions and have an inner self which is reflected in their actions. Shoemaker then states that "what makes an attitude attributable to an agent is just that it expresses the agent's deep self, whether or not we recognize it as such" (59-60).

Although philosophical zombies can technically show emotions, that is, behave *as if* they had emotions, they in fact feel no emotions and have no deep self. For philosophical zombies there is no experience of what it is like to be a zombie, and they can therefore have no deep self. If they were to accidentally step on someone's foot, it would not go against their deep selves for they have none. Further, for one to have a deep self and attributable responsibility, a person must have a normative competence, which means that they can adapt to the values in the normative world (61). It is doubtful if a philosophical zombie could adapt to other people's values, since adapting to values would mean that one has their own values or desires in the first place. Véliz argues that it is therefore questionable if philosophical zombies would be responsive to reasons. If a philosophical zombie was to accidentally step on someone's foot, they would not react to doing so because they themselves have no experience of the pain of being stepped on. The philosophical zombie might perhaps see a reason for taking away their foot, but only when they are asked to do so (Véliz, 2021: 492). One might further argue that a

p-zombie that acts in similar ways to a human must inevitably express certain opinions and actions. A philosophical zombie that acts as e.g., a 'good mother' or a 'good friend' must then perhaps show certain cares and commitments. So, if the p-zombie were to eventually act in a way that strides against the characteristics of a 'good mother' we would then perhaps say that she was not acting like herself. Are these cares and commitments then a reveal of the deep self and in turn a qualification for attributability? The short answer to this argument is no. For humans, showing cares and commitments can come in an instinctive manner, and instincts are perfect displays of the deep self. However, p-zombies have no inner cares and commitments, even though they might behave in a certain way, there is nothing in them that shows that they do or do not care. As mentioned, several times, p-zombies are completely empty inside, so no cares and commitments that they display, reflect their deep self as they have none. Therefore, based on the qualifications of being an attributable agent according to the 'Tripartite Theory of Responsibility', one can conclude that philosophical zombies cannot be attributable agents as they have no deep self.

2.1.2 Answerability

The second responsibility type, *answerability*, tracks the 'judgements' shaping one's quality of will. Similar to attributability, this aspect of the quality of will is divided into two sentimental characteristics: agential pride and agential regret. Agential pride can be separated into the positive emotion family of approval, approbation, and appreciation. Agential regret can be split into disapproval, frustration and disappointed amongst other emotions. To be an answerable agent, one needs to be able to reason and think. An answerable agent must however be able to have the capacity to recognise the so-called "instead of" reasons. When this is the case, the agent is aware of normative domains and can, by their own will, behave in a way that is normative and morally acceptable (Shoemaker, 2015: 35, 82). Furthermore, the "instead of" reasons that an answerable agent needs to have access to are not only to behave in a normative and morally acceptable way, but also to choose certain types of actions and make certain choices that might not be what the agent themselves want, but will as a result, not hurt someone else (176).

Now when it comes to philosophical zombies, one might at first believe that they are not eligible for answerability and have no answerable responsibility, on the basis that they do not feel any of the emotional responses mentioned previously. However, should it not be possible for philosophical zombies to have access to "instead of" reasons somehow anyway? Although

philosophical zombies might not have access to "instead of" reasons in a way that most humans do, notably they might not have access to reasons why no one should be harmed other things being equal. However, philosophical zombies might still have access to such reasons in a way that would ensure that they do not stand out among human beings. Perhaps a philosophical zombie would not say out loud that a painting in an art gallery is ugly, in contrast to how the psychopath would (178). If so, this is because philosophical zombies are supposed to act like highly functional "normal" humans. However, note that they would not say that the painting is ugly in order to avoid hurting someone else's feelings, but only because they are imitating the admiration that everyone around them is putting on display. By the same token, if everyone around them instead of admiring the painting would be showing disgust towards the painting, then perhaps the philosophical zombie would also do that. Moreover, for a being to be an answerable agent, they must be able to make *conscious* choices about desires and what type of judgements they should act upon (197). That said, philosophical zombies do not use any kind of moral compass for "instead of" reasons. They do not act a certain way because that is the normative and morally acceptable way, but only because everyone around them acts in that way too. If philosophical zombies were surrounded by murderers or abusers, then they would perhaps in that situation also murder or abuse as that would be the "normal" way to behave in that specific scenario. For these reasons, it is difficult to say if the philosophical zombies can be held morally answerable for their actions as it might seem that they are only imitating the behaviours of others instead of consciously choosing the morally correct, normative, or less harmful action to perform.

2.1.3 Accountability

The third and final responsibility type from the tripartite theory is *accountability* tracking the "regard" manifested by one's quality of will. As the two responsibility types before, accountability is divided into two characteristic sentimental syndrome pairs which in this case are agential anger and agential gratitude. These two are then further divided by the families of emotions they stand for: resentment, guilt, indignation and shame on the negative side and gratification and warm feelings on the positive side (Shoemaker, 2015: 35). What makes this responsibility type different from the two previously mentioned types is that accountability is the only one that focuses on confrontative interactions with other people. Shoemaker describes that "[t]o have quality of regard an agent must be capable of either (a) coming to see facts about others' (or the agent's own) normative perspectives as putative reasons in the agent's own

normative deliberations, as a function of evaluational empathy, or (b) coming to feel what others feel in a simpatico fashion, as a function of emotional empathy" (113). To be accountable is to be an agent fit to be responsible for a certain action.

If we are to look once again at the philosophical zombies, are they potential agents for accountability or not? The short answer for this question would be no, philosophical zombies cannot be accountable moral agents. The reason for this is that they lack any sense of empathy or can identify empathy. Philosophical zombies have no quality of regard. No matter what actions the philosophical zombies might execute, they will never be regard-based due to the fact that they lack regard from the beginning. If one is to look at the two descriptions that Shoemaker provides about having quality of regard (see above for more details), the first is to identify others' or one self's own normative thoughts and behaviours, while the second one is about identifying what others feel and empathising with it. Philosophical zombies, as mentioned previously, are unable to identify feelings in other people for they have never felt those feelings themselves before. Philosophical zombies will never be able to understand why a human is sad for they do not know what it is like to be sad, they will not know why someone is yelling at them for they have no understanding if they have done anything wrong since they have no regard or any guilt, and they do not understand the emotion of anger, or any emotion for that matter. Philosophical zombies are not capable of either evaluational empathy or emotional empathy. They therefore lack quality of regard, which in turn implies that philosophical zombies cannot be accountable agents and therefore are not eligible for accountability.

Let me summarize the discussion so far. If philosophical zombies were to hypothetically exist, they would not be qualified to be either attributable agents or accountable agents according to Shoemaker's 'Tripartite Theory of Responsibility'. However, it might in some ways be possible for a philosophical zombie to be an answerable agent. This is still somehow doubtful as their "instead of" reasons would not be based in judgement, as Shoemaker describes, but perhaps in imitation to those around the p-zombie. It is therefore doubtful whether philosophical zombies can be morally answerable or not. Although it could perhaps now be argued that philosophical zombies are neither morally attributable nor accountable, and perhaps morally answerable, there are still similarities between philosophical zombies and psychopaths, who instead are morally answerable and attributable. So, what are the qualities of philosophical zombies and psychopaths that makes them similar but also different? This is the question I turn to now.

2.2 Psychopaths vs Philosophical zombies

If what I have presented so far is correct, then philosophical zombies have little to no moral responsibility. As mentioned before, philosophical zombies are not eligible for attributability nor accountability, and might just be qualified for answerability. Psychopaths however are, according to Shoemaker, eligible for answerability and attributability but not for accountability. But how can philosophical zombies and psychopaths, who both lack empathy, emotional deficiencies and are able to adapt to fit in, have different outcomes when it comes to the different kinds of moral responsibility? Before comparing the two types of (deficient) agents, let us try to find a workable definition of a psychopath is. What qualities does a human need or perhaps lack to be considered a psychopath?

In psychiatry, a psychopath is defined as having ASPD which stands for 'antisocial personality disorder'. Even though the name psychopath is used for people with this mental illness, the term is actually not an official diagnosis. However, for this essay, I will use the term 'psychopath'. The term 'antisocial' is used regularly nowadays, by people claiming to be antisocial and preferring their own company on a day-to-day basis. However, Dr. Prakash Masanad describes 'antisocial' in the context of psychopathy and ASPD as "(…) someone who goes against society, rules, and other behaviours that are commonplace" (Healthline, 2019). There are several traits and common signs of psychopathy. The most relevant to this essay are behaviours that are socially irresponsible, and not being able to differentiate and separate what is right and what is wrong. Furthermore, psychopaths also have a hard time showing remorse or empathy, they show tendencies of manipulating and hurting others, and have "general disregard towards safety and responsibility" (Healthline, 2019). These are some of the traits of having ASPD and therefore being a psychopath. So, when looking at the different responsibility types that Shoemaker presents, how can a psychopath have them and why can a philosophical zombie *not* have them? Additionally, it is true that there are psychopaths that are successful and highly admired, but I will base my discussions about the psychopaths that Shoemaker describes, which are more manipulative and narcissistic.

2.2.1 Accountability

If one is to first look at accountability, which is until now the only responsibility type that the philosophical zombie and psychopath might be thought to share, neither the p-zombie zombies nor the psychopath are eligible for this responsibility type. Philosophical zombies cannot be held morally accountable as they lack the qualities Shoemaker describes as necessary for

someone to be an accountable agent. They have no regard, lack empathy and are not able to identify these qualities in others. Because of these reasons, philosophical zombies are not capable of performing any "regard-based" actions.

Similar to philosophical zombies, psychopaths are not suitable for accountability either. Even though psychopaths can identify other people's emotions, they are unable to care about others (Shoemaker 2015: 159-160). Furthermore, psychopaths are unable to identify regard and tend to feel no guilt. The reason for this is that psychopaths have a lack of concern for morality itself as well as for other people, which is why they are incapable of identifying the quality of will and perform regard-based actions. (Shoemaker 2015: 177). Since both philosophical zombies and psychopaths cannot show regard, they do not qualify as accountable agents.

2.2.2 Answerability

As for answerability, it could be argued that philosophical zombies might not have this while psychopaths do. According to Shoemaker, so-called "successful" psychopaths could sometimes behave in a way that could seem as if they would be capable of regard-based actions. However, successful psychopaths could consciously make these choices for the sole purpose of manipulating people. If psychopaths act as if they care about others' interests, and this makes these people like them, the psychopath can manipulate them in turn. It would therefore benefit the psychopath more to not always act in their own self-interest because one cannot manipulate a person that cannot stand being around you. So, in the long term, the psychopaths care about their own self-interest, but to manipulate others, they have to act consider others' interests. Because psychopaths make these choices *consciously*, they can therefore still be held morally answerable (179). The difference is therefore that psychopaths *consciously* make choices like manipulating those around them to obtain whatever is in their self-interest.

However, philosophical zombies do not even have a consciousness to begin with and can therefore not behave in a similar fashion to psychopaths. It is therefore possible for psychopaths to be qualified for answerability but not for philosophical zombies. It was mentioned previously that philosophical zombies *might* be answerable agents but with the help of comparing p-zombies to psychopaths, one can now make a case that for one to be an answerable agent, one has to make *conscious* choices. Since philosophical zombies cannot make these conscious choices, they can therefore not be eligible agents for answerability either.

2.2.3 Attributability

When finally looking at attributability, it seems as if psychopaths can be qualified for this but not philosophical zombies. As previously mentioned, philosophical zombies have no deep self, they feel nothing and there is nothing like to be a zombie. Therefore, they cannot perform actions that reflect, or do not reflect, their deep selves. Psychopaths on the other hand can be attributable agents in two potential ways. First, they may be said to have desires to hurt someone. Second, they may be said to be able to judge whether a person's preference not to be hurt is a good reason or not against hurting the said person. This is where the narcissistic side of psychopaths will appear, for they see no reason against not satisfying their desires. Shoemaker argues that "(…) despite their deficits in recognizing either moral or prudential worth, there still seems to be something that matters to them, namely, that they satisfy their desires" (181). Now, this statement might be one of the main reasons why philosophical zombies and psychopaths have similar traits but can be argued not to be eligible for the same responsibility types in the final analysis.

Philosophical zombies, as we have established by now, lack all sorts of emotions and consciousness, they feel nothing despite looking as if they did, they have no desires nor judgements. Psychopaths, however, have emotional deficiencies, yes, but just as Carissa Véliz mentions in her article 'Moral zombies: why algorithms are not moral agents' (2021), to deny psychopaths all emotional capabilities would be unjust as they do have desires, goals, and values (Véliz 2021: 496). Psychopaths do have emotions, but the problem is that they lack the ability of being empathetic and of understanding what others feel. It is therefore possible to say that despite some similar traits between psychopaths and philosophical zombies, psychopaths are still conscious beings with emotional and moral deficiencies while philosophical zombies do not even have a consciousness to begin with.

2.3 Philosophical zombies as incapable of moral agency

This essay has now reached the interim conclusion that philosophical zombies cannot be held morally responsible according to Shoemaker's "Tripartite Theory of Responsibility", they are not qualified for accountability, attributability or answerability. Furthermore, as Véliz argues, philosophical zombies are not eligible for any type of moral responsibility due to the fact that they lack the ability to even comprehend morality, and they are not able to empathize with human beings or other living beings. Véliz claims that for a being to have moral responsibility, they must have emotional experience and experiential knowledge. Experiential knowledge is

knowledge gained through experience such as knowing it will hurt to put your hand on the stovetop because you did it once before. Experiential knowledge is not something philosophical zombies can have because they lack the basis of a consciousness and do not feel the pain, the joy or anything else for that matter (Véliz, 2021: 487). This argument goes well with Shoemaker's theory of moral responsibility as all of the responsibility types are based on different types of behaviours and emotional reactions (Shoemaker 2015: 35). What can then be inferred is that without experiential knowledge and understanding of emotions, one cannot be a moral agent. Unlike psychopaths, who only have emotional deficiencies and are not completely without experiential knowledge, philosophical zombies have no desires and no values. Véliz claims that if a being cannot have any emotional experiences, then they can therefore have no values, and without values they cannot have moral responsibility or be moral agents. What that suggests is that despite being able to act in a harmful or immoral way, philosophical zombies cannot have any moral responsibility (Véliz, 2021: 495).

Additionally, it is worth considering David Shoemaker's thesis that for one to have cares, one must have emotional reactions. Emotional reactions are different from physical reactions in the way that emotional reactions are "sensibly interpretable" (Shoemaker 2003: 93). For a reaction to be an emotional reaction, it must be a reaction to an event that affects oneself. If it does not affect a person, then it is not an emotional reaction and if it is not an emotional reaction then the person does not care about the event that occurred. Furthermore, Shoemaker writes that "the emotions we have make us the agents we are" (94), which further implies that without emotions and values, one cannot be a moral agent and therefore have no moral responsibility. Since philosophical zombies have no emotional reactions to anything and have only physical reactions, they therefore lack cares and emotions and with it. Consequently, they would perhaps, according to Shoemaker, lack the ability to be moral agents and have no moral responsibility.

Furthermore, Véliz offers several examples of autonomy types, and there is one in particular that can be connected to moral responses and responsibility: self-governance and reason-responsiveness. This way of thinking about autonomy is based on the purpose of self-governance. According to Véliz, if someone is a moral agent then they will base their actions and act according to them based on reasons. To be a moral agent one must therefore be capable of reason-responsive actions, and if someone cannot understand reasons then they cannot be a moral agent (Véliz, 2021: 491). This is in fact the case for philosophical zombies. Philosophical zombies have no ability to perform reason-based actions, they have no desires and nothing that motivates them to act in a certain way. Moreover, philosophical zombies would not be able to

comprehend why others would want them to behave in a certain way for they do not understand reasoning in the first place. Chalmers describes philosophical zombies as dark inside, meaning that despite them behaving in a certain way, they will not have any conscious experience (Chalmers: 96). As mentioned above, philosophical zombies have no ability to perform regard-based actions, which seem to be the same thing as Véliz describes when she talks about reason-based actions. What this implies is that because the philosophical zombies lack the ability to perform reason-responsive actions and cannot understand them, they therefore cannot be moral agents.

To Véliz's arguments and the tripartite theory it is worth adding, Gunnar Björnsson's and Kendy Hess's discussion of how certain entities cannot be agents (2017). Although Björnsson and Hess discuss corporates and collectives, and not philosophical zombies, their arguments can still be applied to p-zombies, or so at least I will claim. Björnsson and Hess argue that for an entity to be an agent, that being needs to be able to act according to their own beliefs and values. In addition to being an agent, if one is eligible to be a moral agent, the entity needs to be able to "(…) subject to moral obligations" (273), but also be eligible for accountability and for being held accountable. For a being to be a moral agent they need to be capable of understanding what they did wrong and take accountability for it. Björnsson and Hess explain that for a being to take accountability for their actions and hold themselves responsible for it, they will need to be capable of feeling guilt or regret about what they have done wrong (273-274). Furthermore, even if the philosophical zombie was to be capable of *showing* any emotions, desires, values, or other traits, if they themselves do not have these traits, if they do not actually experience desires, emotions, or anything else then it is not possible for that being to be a moral agent (274).

Moreover, in the article, Björnsson and Hess list additional characteristics that are crucial for a being to have for it to be counted as a moral agent and have fully fledged moral responsibility. These two characteristics are the capacity to feel (and/or be the target of) indignation and guilt (283). Now, both guilt and indignation are emotions that have been previously mentioned and are emotions connected to the responsibility type 'accountability' (see 2.1.3). Alongside what Shoemaker discussed about accountability in his theory, Björnsson and Hess dwell deeper into the qualifications of being eligible for guilt and indignation. The reaction that mainly derives from indignation is anger, which is also further confirmed by Shoemaker as seen in the tripartite theory (Shoemaker 2015: 35). For one to be able to express indignation, one has to be able to express and experience anger at the one that has done an immoral action, and when that person has expressed themselves as guilty or feeling the need to

change, the agent will mitigate the anger, or if the other person does not show sign of remorse, the agent's anger will intensify. If an agent is able to express and experience guilt, then they are able to feel and express sadness and regret. An agent eligible for guilt will be able to blame and hold themselves responsible when realising that they have performed an immoral action. They will thereafter be capable of changing their behaviours so that they will not perform that immoral action again. In that way, an agent that can feel guilt, will also have the qualities to change their future action to not behave immorally and thereby not feel guilt for that action again (Björnsson & Hess 2017: 284).

Philosophical zombies may at first seem to be able to express anger or sadness and will perhaps be thought of having moral agency. However, what should by now be established through the example of philosophical zombies, is that experiencing and expressing emotions are two different things. Philosophical zombies are able to express emotions, in fact they are able to cry and get angry, and they could perhaps even start a fight or engage in similar behaviours that might seem to be emotionally driven actions. But despite being able to express all the emotions that humans are capable of expressing, philosophical zombies are not able to experience the emotions in the slightest. If someone is cruel to them, the philosophical zombies will not feel how their blood is boiling in anger, they will not get goosebumps from watching a thrilling movie, and if a supposed loved one passes away, they might cry but never feel the grief of losing someone they are believed to hold dear. Furthermore, if the philosophical zombie performs an immoral action, perhaps harming someone, they will never feel guilty for doing it, and even if another person was to explain to the p-zombie that they should not harm others for it is wrong to do so, the p-zombie will not change their behaviours. The philosophical zombies will never act in a different way in regards of others, they will not choose the "instead of" action or experience any emotions. Philosophical zombies are therefore not eligible for guilt or indignation, nor any other emotion, and certainly none of the responsibility types mentioned in Shoemaker's theory.

2.3.1 The dilemma of not being a moral agent

Since philosophical zombies lack consciousness and therefore not capable of reason-responsive actions of any kind, and they are not capable of feeling indignation nor guilt, they cannot be morally responsible agents. However, if they were to exist and actually do damage and behave immorally, despite not having any experiential knowledge or consciousness, should the philosophical zombies not take any responsibility at all for this or be punished in some way?

This question is relevant to ask for it is an ongoing debate how we are supposed to punish those that have mental disorders but commit crime. The consequences for those people would be to be placed in psychiatric hospitals. So how would we tackle punishing criminal p-zombies?

Véliz briefly discusses this dilemma in her article (2021) and mentions that if humans were to give philosophical zombies moral responsibility and thereby punish them for immoral behaviours, we as humans would not gain anything from that. For how is one supposed to punish a being that will learn nothing from the punishment? Philosophical zombies value nothing, not even freedom, and if humans were to lock away a p-zombie for committing a crime, they would not feel anything from it, it would not matter to them at all. Véliz adds that "[t]he lack of freedom can only be a punishment to a being who values freedom" (Véliz, 2021: 496). What this would mean is that even if humans were to imprison a philosophical zombie, it would not make any difference to the p-zombie as they would not miss having freedom or feel the desire to be free.

If humans instead tackled immorally behaving philosophical zombies as criminals, perhaps it would be more beneficial to approach them in the same way that one approaches people that are mentally ill, as the psychopath for example. For people that have committed crimes and are then revealed to have a mental illness, they are instead of prisons placed in psychiatric hospitals. What if this would be a better method for the philosophical zombies? Despite the fact that one could see the philosophical zombies as mentally ill in some ways, one must eventually realise that philosophical zombies are not mentally ill, and moreover have no intellectual disabilities, if that was something some perhaps considered. The dilemma of philosophical zombies, and especially if they were to behave immorally or in a harmful way, is that they will never be capable of understanding the value of freedom, of doing harm, of being happy. Philosophical zombies just *are*, they merely exist as "shadow versions" of humans. Philosophical zombies only behave similar to us humans in the physical matter but will show nothing on the inside, they have no consciousness and will therefore never spare a thought for the person they hurt or for the person they made happy. Philosophical zombies are completely empty inside with no emotional reactions of any kind, no quality of will or anything. It would therefore be completely useless for humans to imprison philosophical zombies or place them in psychiatric hospitals.

On the other hand, humans could approach the p-zombies as if they were simply machines or robots. If philosophical zombies do not act as humans, should humans even attempt to treat them as such? But instead of imprisoning them or attempt to give them psychiatric care, p-zombies could be given tasks similar to what humans give machines and robots, factory work or anything comparable to it. Because of the fact that philosophical zombies cannot feel

exhaustion or hunger (or anything for that matter), they could be put in factories and are forced to work there without brakes, all day every day. This could perhaps be the most optimal solution because in that way, the philosophical zombies would not be placed in locations that would gain them nor humans anything, but by doing machinery work with no hours, it would still be of no matter to them, but it would be beneficial for humans. However, despite the fact that this idea might be a great alternative, p-zombies still look and behave as us humans in the physical sense and perhaps they would have no moral responsibility, but humans still do. By forcing these human-like beings to work in inhumane ways, humans are going against their own moral duties, and it would be seen as immoral of humans to treat the philosophical zombies in this way. Why humans would have difficulties in treating p-zombies as non-human beings can be shown in the example below.

Even in our society today, there are ongoing debates whether AI's should be treated in a morally correct way. Véliz argues that if philosophical zombies or AI's behave like humans, then they should be treated as humans (Véliz, 2021: 488). Véliz discusses the dilemma of AI's further in (2016). Véliz brings up the example of the future AI 'Alfred' that works for 'you', Alfred does all of the house chores. He can even have conversations and make jokes, but he is a robot. However, one day Alfred says that he would like to switch careers and eventually starts showing signs of emotions that he should not be showing. The manufactures explain that this is simply a bug, and they "fix" him, but now and then, Alfred will make an upset expression when he does not believe that you are looking at him. Véliz offers this example to make one understand the dilemma of an entity that is not human, but acts like one, and if they act like humans, should they not be treated as such? Véliz ends her article by saying that if artificial intelligences seem to have emotions, then one should treat them as having such. Alfred should then perhaps be treated in the same way as humans because he seems to act in the same way as humans do (Véliz, 2016).

So, if we are to compare the AI Alfred with philosophical zombies; despite the fact that philosophical zombies would have no values and no moral responsibility, humans would most likely treat them as other human beings and it would therefore be immoral from humans' perspectives that philosophical zombies are treated in inhumane ways. Furthermore, humans could be hurt from these actions and experience compassion for the p-zombies for even though it is fake, it would be as if observing human suffering. We know that the p-zombies would not feel anything, but the fact that it would look as if humans were suffering, treating them as machines would be a way of punishing ourselves as well.

Consequently, one must therefore draw the conclusion that philosophical zombies are not eligible for any type of moral responsibility and can therefore not be moral agents in any way. Because philosophical zombies lack the ability to understand moral behaviours and have never had any emotional experience, then they will never be able to be moral agents. Furthermore, even if philosophical zombies were to behave immorally, humans would not gain anything from punishing them as they do not value freedom nor would they feel anything from being punished. Moreover, it would not be of any use to place philosophical zombies in psychiatric hospitals for they are not mentally ill and will not be able to improve either. If humans were to instead treat philosophical zombies as machines and force them to work constantly without any regard for them and not imposing limits on their work time, that would be immoral from the humans' side. For humans would see that the philosophical zombies behave just like humans in all physical matters and would therefore want to treat them as such. All methods, whether it is in an attempt to punish the philosophical zombies, try and give them psychiatric care or treat them as machines would thereby be of no use or go against humans' own moral duties and beliefs. There would thereby be no method that would make humans gain anything from it. Because of this, philosophical zombies would therefore be the same as hurricanes, tsunamis, or any other natural catastrophe. They would be able to cause damage and do harm, but never held morally responsible for the damages they create.

3. Summary

The philosophical zombie argument is a thought experiment created for the purpose of criticizing physicalism and different theories such as materialism and behaviourism. They were invented to prove that not everything in the world is physical, and consciousness is not something that is physical, which is what materialists believe to be true. Philosophical zombies look and behave just like humans, in every physical aspect they are exactly like us. Furthermore, they will complain about the weather, say that the chocolate is sweet or have heated discussions. But what separates philosophical zombies and humans is that philosophical zombies lack a consciousness entirely. Despite saying that the chocolate is sweet or saying that it is cold outside, they will never experience this in an emotional way. They will never know what any experience is like, for there is nothing it is like to be a philosophical zombie. There is nothing about philosophical zombies that makes them, them. In this essay, the philosophical zombies have been used for the purposes of a different thought experiment to understand if they, despite not having a consciousness, are eligible for moral responsibility and thereby be moral agents.

By using Shoemaker's 'Tripartite Theory of Responsibility', I have concluded that philosophical zombies cannot be held morally responsible, at least according to this theory and the arguments that I presented in chapter 2.3. The first responsibility type that I discussed was attributability. For a being to be qualified for attributability and thereby be an attributable agent, one must perform actions that reflect one's deep self. What this means is that if one was to accidentally step on a stranger's foot, they could not be held attributably responsible for that action for it was not something that they truly wanted to do, the action simply did not reflect their deep self. Philosophical zombies are not suitable for attributability for they lack a deep self, there is nothing like to be a p-zombie, they have no consciousness and there can therefore be nothing that represents what their inner selves believe to be right or wrong. For one to be attributably responsible, one must have actions that reflect one's deep self, and since philosophical zombies lack this quality, they cannot be eligible for attributability.

For the second responsibility type, answerability, one is suitable for this responsibility type by having so-called "instead of" reasons and a sense of judgement. The "instead of" reasons are reasonings the agent can choose to act in accordance with so that it will be more normative or morally acceptable. It was at first doubtful if philosophical zombies could be eligible for this responsibility type or not. One could believe that philosophical zombies would have some sort of "instead of" reasoning since they might follow and impersonate how humans around them

behave in order to act normally. However, by comparing philosophical zombies to psychopaths, one gets a better understanding of the "instead of" reasons and understands that in order to have a sense of judgement and choose the "instead of" alternatives, one must have the ability to make conscious choices, which is not something philosophical zombies are capable of doing. Therefore, philosophical zombies cannot be appropriate agents for answerability either.

The final responsibility type, accountability tracks whether a person's quality of will shows proper "regard". To be an accountable agent, one must be able to undertake regard-based actions and be able to emphasise and identify emotions in others and themselves. The philosophical zombie is not qualified for this responsibility type. Since the philosophical zombies have no consciousness, they lack the ability to identify these qualities in others and do not have them either. Because of the fact that philosophical zombies lack the basis of what it takes to be an accountable agent, they cannot be eligible for accountability.

In chapter 2.2, a comparison between philosophical zombies and psychopaths was presented. Psychopaths, or people with antisocial personality disorders, have behaviours that are anti-social, and go against certain social rules. Some psychopaths are successful and admired but I chose to focus on Shoemaker's description of psychopaths who are seemingly more manipulative and narcissistic. Psychopaths are not able to differentiate between right and wrong, especially when it comes to morality, they can manipulate and hurt others and lack the ability or have a hard time showing and identifying empathy. At first, I believed that since philosophical zombies shared similarities to psychopaths due to the reason that they both are unable to emphasise and lack emotional qualities then perhaps the p-zombie should also be held morally responsible since the psychopath should be. However, after comparing the two, I have concluded that it would be unfair to the psychopath to deny them all emotional experiences.

Psychopaths, in contrast to philosophical zombies, have desires and values, they want to achieve what will benefit them the best, and even if they do this by manipulating others, they do so to fulfil their self-interests. Psychopaths make conscious choices to fulfil their desires. Philosophical zombies on the other hand have no desires and no values, they would not do anything in their own self-interest for they have none, every action they perform is not based on conscious experience because they have no conscious experience. Psychopaths and philosophical zombies seem to share similar traits other than lack of emotional experience: they also behave as if they were to be the standard human being. Psychopaths do this in order to manipulate others, for they need to act upon other's self-interests in order for these people to like them so that they can afterwards manipulate the said person and satisfy their own desires. The philosophical zombie on the other hand behaves just like everybody else because that is

what a p-zombie does. They cry, they laugh, they perform moral and immoral actions, they do everything that humans do but they lack the consciousness, the consciousness that makes humans, humans or animals, animals. There is nothing like to be a philosophical zombie and they do nothing to fulfil their own desires for they have none. Therefore, psychopaths and philosophical zombies are not as similar as one would first believe, psychopaths "only" have emotional deficiencies while philosophical zombies lack emotions completely. Because of this, one can understand why the psychopath has moral responsibility, or are at least eligible for attributability and accountability, and not the philosophical zombie in comparison.

In chapter 2.3, I continued the discussion of whether philosophical zombies could still have moral responsibility despite them not having it according to Shoemaker's tripartite theory of moral responsibility and the dilemma of not being moral agents. It was concluded that if a being is to be qualified as a moral agent and eligible for moral responsibility, the agent must have emotional experience and experiential knowledge. Such knowledge that one gains by experiencing it or at least being able to emphasize with for example the feeling of it. Philosophical zombies can neither experience any type of emotions nor can they emphasize, and because of this, philosophical zombies are not suitable to be moral agents. Furthermore, Véliz argues that for one to be a moral agent, one must base their actions on reasons and have reason-based actions.

Björnsson and Hess (2017) argue that for an entity to be a moral agent then they need to be able to act according to their own beliefs and values. Furthermore, the entity would have to be able to express and experience both indignation and guilt. These two emotions are also emotions that Shoemaker includes in his tripartite theory as emotions that comes from an agent being responsible for accountability. Since philosophical zombies do not have reason-based actions, nor able to act according to their own beliefs and values, they are once again seen as incapable of being moral agents.

4. Discussion

Now, I am aware that there are plenty of theories about moral responsibility and differing views on what characteristics a being needs to have to be a moral agent. It is therefore impossible to disregard philosophical zombies any moral responsibility. However, based on the literature surveyed here, it seems highly unlikely that philosophical zombies would have any type of moral responsibility. This is because a comparison of Shoemaker (2015), Véliz (2016; 2021) and Björnsson and Hess (2017) shows that there seems to be a mutual understanding of what qualities an entity needs to possess to be eligible for moral responsibility. So even if one were to use other theories about moral responsibility, it seems highly unlikely that they will differ much as regards the definitions of moral agency. However, since I have not looked into more theories about moral responsibility, it is not inconceivable that on some theories philosophical zombies would qualify as moral agents. Although based on the theories that I have presented and the similarities they share for what it takes to be a moral agent, it is difficult to say what claims a different theory would have to make to conclude that philosophical zombies are eligible for moral agency. It is therefore doubtful that any other theory of moral responsibility would lead to a different result.

Furthermore, if we as humans were to give philosophical zombies moral responsibility and punish them if they were to behave immorally, it would not be of any use for us, and it would not matter to the p-zombie as they lack consciousness and therefore do not value freedom. If humans were to instead place the philosophical zombies in psychiatric hospitals and give them psychiatric care this would do no good either since they have no consciousness to begin with, there is nothing the psychiatric care could do that would be of any help to them. The last suggestion I mentioned was to treat the philosophical zombies as machines and force them to constantly work without limits because they will never feel exhaustion or hunger. However, this is not ideal either for we humans would see these human-like beings work inhumane amounts with not a moment to breath. Here I brought in the story of the AI Alfred that seems to be expressing emotions that he should not be expressing, emotions that are not installed into the robot. According to Véliz, in this scenario we face a dilemma because we know that Alfred is a robot and does not have a consciousness, but by acting like a human, we would want to also treat him as such. In this case, treating him as a human would thereby imply that we allow Alfred to switch jobs. Because the philosophical zombies behave like humans, we would also

treat them and see them as humans, so forcing the philosophical zombies to work as if they were machines would instead be immoral of humans.

In conclusion, philosophical zombies are not eligible for attributability, accountability or answerability. Because they lack all of these responsibility types, have no emotional experiences, no experiential knowledge, and cannot perform reason-based actions, philosophical zombies cannot be qualified as moral agents and thereby have no moral responsibility. Granted, they are still able to do harm and behave immorally but it would not be of any use or purpose to imprison them or place them in psychiatric hospitals. Moreover, since philosophical zombies behave and look exactly like humans, it would be immoral of us to not treat them as such. For example, it would be immoral to force them to work as machines just because they will never feel tired. Since they behave like us, we might feel entitled to see them as humans, and therefore we would not be able to force them to work like machines. So, despite the fact that they are not morally responsible, humans could feel that they have moral obligations towards them and not be able to treat them as machines. Therefore, in the end, if philosophical zombies were to hypothetically exist, they would have no moral responsibility and if they were to behave immorally, they would be treated like natural catastrophes, able to cause damage but never able to take any responsibility for it.

4.1 Future Research

As mentioned in both in the Introduction and in the Conclusion, there is a vast number of theories on moral responsibility, some of which bears on philosophical zombies and whether they are fit to be held morally responsible. It would therefore be interesting for future research to dig deeper into the behaviours of philosophical zombies and apply the findings to other theories on moral responsibility. Moreover, one could also further compare philosophical zombies to psychopaths and artificial intelligence, but it would also be interesting to compare them to the cases of people with autism, dementia, or mild intellectual disabilities (MID) as such cases also frequently figure in the discussion about who is eligible for what moral responsibility type.

There is also the possibility of studying and doing further research on whether philosophical zombies are eligible for types of responsibility other than moral responsibility—for example, collective responsibility, legal responsibility, social responsibility, and several other types of responsibility. One could also include the discussion of how one is supposed to

tell the philosophical zombies apart from humans, since by stipulation there is nothing physical that makes them different from us.

Other than comparing p-zombies to other people or responsibilities, one could also expand if philosophical zombies lack all sorts of consciousness. In this essay, and in Chalmer's definition of philosophical zombies, they lack phenomenal qualities, or qualia. But do they lack al other types of consciousness as well? This question is something that is also possible to write about in further research.

By doing further studies about philosophical zombies, other moral responsibility types as well as other kinds of responsibility types, comparing the p-zombies further to artificial intelligences, psychopaths, along with people with other types of mental difficulties, and using other types of consciousness, one would get a deeper understanding of philosophical zombies and their significance for our views on agency and moral responsibility. This might even force me to revise the conclusions that I have reached in this essay. Additionally, by adding in other types of responsibility, one might better understand not just how philosophical zombies would behave, but also how humans are supposed to treat them if they were to exist and possibly do damage or behave immorally.

5. Works cited

Björnsson, Gunnar & Hess, Kennedy (2017). Corporate Crocodile Tears? On the Reactive Attitudes of Corporate Agents. *Philosophy and Phenomenal Research* 94 (2): 273-298. https://philpapers.org/archive/BJRCCT.pdf

Chalmers, David John (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford Univ. Press

Healthline (2019). *Psychopath*. https://www.healthline.com/health/psychopath [08-01-2022]

Kirk, Robert, "Zombies", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2021/entries/zombies/ [20-12-2021]

Nagel, Thomas (1974). What Is It Like to Be a Bat? *The Philosophical Review* 83, no. 4: 435-450 https://doi.org/10.2307/2183914

Shoemaker, David (2003). Caring, Identification, and Agency. *Ethics* 114, no. 1: 88–118. https://doi.org/10.1086/376718

Shoemaker, David (2015). *Responsibility from the Margins*. Oxford, United Kingdom: Oxford University Press

Véliz, Carissa (2016). *The Challenge of Determining Whether an A.I. Is Sentient* [13-01-2021]

Véliz, Carissa (2021). Moral zombies: why algorithms are not moral agents. *AI & Soc* 36**,** 487–497. https://doi.org/10.1007/s00146-021-01189-x