Can Only Androids be Ethical?

Introduction

Joseph Emile Nadeau argues that only androids can be ethical.[1] An android is a robot whose purpose is whatever the creators of the android program it to do, such as killing military targets, serving butter, or diagnosing disease. Two people whose views are in conflict with Nadeau's argument are Alvin Plantiga and John Searle. Plantinga argues that naturalism, an important assumption in Nadeau's argument, invalidates its own conclusion.[2] Searle argues that only a human agent could cause their own reasons for their actions, which shows androids cannot be held responsible or be ethical agents.[3] I lay out Nadeau's argument and explain how Plantinga and Searle's objections cause problems for Nadeau's conclusion. Nadeau fails to make some clarifications about the ethicality of androids, which I address. I also explain why Plantinga and Searle's objections do not undermine Nadeau's conclusion. Then I summarize Nadeau's argument and why it's unperturbed by objections, and conclude only androids can be ethical.

Nadeau's argument in standard form[4]

1. Agents are only ethical if they are responsible.
2. Responsibility is the connection between reasons and actions.
3. Given scientific naturalism, human action has little connection to reasons.
4. Theorem provers and neural nets have traceable reasons.
5. Traceable reasons allow a connection to be made between action & responsibility.
6. Androids are built from theorem provers and neural nets.

C. Androids but not humans can be ethical.

[1] Nadeau, Joseph, "Only Androids Can Be Ethical," in *Thinking about Android Epistemology*, ed Kenneth M. Ford, 2006. 241-48

[2] Plantinga, Alvin "Warrant and Proper Function" New York: Oxford University Press, 1993. 216-37

[3] Searle, John, "Minds, brains, and programs" in *The Behavioral and Brain Sciences*, 1980. 417-24.

[4] Nadeau, "Only Androids Can Be Ethical," 241-48

Responsibility and Ethicality

Nadeau says, “The notion of responsibility is inextricably linked with the notions of capacity and free will.” For example, in a burning room full of people, no one in the room is responsible for breaking through a brick wall and saving everyone because no one has the capacity to do that.[5] The firefighters on the outside however do have the capacity to save the people by extinguishing the fire and are responsible to a degree to save those people.

Scientific Naturalism

The rest of Nadeau’s argument is founded on naturalism, which is more contentious of an idea than responsibility being linked to free will. Naturalism broadly speaking is the view that the only things that exist are physical things. People who hold the view of naturalism encounter problems with beliefs of non-physical things that we hold do exist, such as morals and mathematics. To overcome these problems one can deny the existence of morals and math but continue to behave as if those things did.[6] The number 1 does not physically exist, but we can behave as if the number does exist by using it in math. Morality does not physically exist, but we can behave as if certain actions are or aren't moral. Naturalism does not include supernatural things such as God, witches, and miracles, and the only things left are physical things, such as galaxies, planets, organisms, cells, chemical reactions, physical movements, and other things that can be confirmed through our five senses.

Summarizing Plantinga’s argument

Plantinga argues it is not rational to believe in evolution and naturalism at the same time. The argument for evolution is convincing through scientific literature and diagrams showing

[5] Nadeau, “Only Androids Can Be Ethical,” 241.

[6] Papineau, David, "Naturalism", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), ed Edward N. Zalta, URL = <https://plato.stanford.edu/archives/sum2021/entries/naturalism/>.

animals evolving and their ancestors.[7] I also believe naturalism is convincing because my five senses have not yet given me a reason to believe in supernatural forces. With these stances, I will demonstrate why Plantinga is wrong and why naturalism is compatible with evolution.

Plantinga argues that a main tenant of naturalism is evolution, because naturalists cannot explain complex life as creations of God. Organisms have organs with adaptive functions, such as stomachs for digestion, livers for blood cleansing, and genitalia for reproduction.[8] The brain has cognitive functions that formulate beliefs about the world that keep us alive, such as avoiding predators and eating food. However, there is a distinction between beliefs that keep us alive and beliefs that are true. Someone may believe that eating food keeps murderous spirits away, and so they eat food to stay alive, but in fact, food keeps us alive because our bodies need calories to metabolize to maintain homeostasis. Whether someone believes in food as fuel or food as spirit repellent does not affect the outcome of staying alive. Because of this distinction, Plantinga argues that our beliefs are not reliable since our beliefs are aimed at survivability rather than truth. Therefore, to Plantinga the belief in evolution and naturalism is aimed at survivability not truth, and is irrational.

An objection would be that Plantinga should not trust his own cognitive functions because they are products of evolution. But because Plantinga is not committed to naturalism, there is room for God to guide cognitive functions to aim at truth rather than just survivability.[9] Given Plantinga's argument, there are three options. Either refute evolution and accept creationism, refute naturalism and accept non-naturalism, or refute reliable cognitive functions.

[7] Millstein, Roberta L., "Evolution", *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/evolution/>.
[8] Plantinga, "Warrant and Proper Function," 21.
[9] Plantinga, "Warrant and Proper Function," 236.

Refuting naturalism

To keep my reliable cognitive functions, I will temporarily refute naturalism. I believe evolution is a stronger claim than naturalism given evolution is falsifiable whereas naturalism is not falsifiable. For example, scientists could discover a fossilized pickle with rat limbs among Cretaceous period ferns which would falsify or at least put into question evolutionary theory, because there is no evidence for an organism so different from other Cretaceous period organisms. In the case of naturalism, if a fire shooting wizard were to reveal himself, his magical powers would perhaps be explained in terms of hidden flame throwers, not actual magic that's present in works of fiction. So I refute naturalism on the basis that naturalism is unfalsifiable.

Choosing one of two unfalsifiable positions

In this section, I demonstrate from a non-naturalist stance why non-naturalism is irrational. Take a primitive isolated Germanic tribe, who lack cognitive functions to make the belief that fire is a chemical reaction called combustion. Instead, they believe that fire is the fart of a demon god. A few miles away a chemist, who has the same cognitive functions as the Germanic tribe, instead believes that fire is the phlogiston in material being dephlogisticated.[10] As time progresses into the modern age, humans discover fire is neither a demon fart nor phlogiston, but a visible effect of a chemical reaction called combustion. The belief of the phlogiston theory was closer to the truth than the belief of demon fart, in that phlogiston did not attribute a non-physical entity to a phenomena that was not yet understood. Other theories backed by non-naturalistic forces such as geocentrism, volcanoes being portals to Hell, stars being holes through which Heaven shines through, Santa Clause climbing down approximately 1.5 billion chimneys in one night are demonstrated repetitively to be false and instead of non-naturalistic explanations have naturalistic explanations.

[10] Phlogiston is an incorrect medieval alchemical theory about fire

The examples listed above suffer from hindsight bias, any naturalist or non-naturalist can agree old non-naturalist theories are false. One may want to keep non-naturalistic beliefs for future and current problems to have an explanation rather than no explanation, such as the cause for the start of the universe, where matter goes in black holes, how life on Earth began, etcetera. However, given the historical record of true-beliefs that non-naturalistic explanations have produced (zero), I conclude it is irrational to believe in non-naturalism.

Naturalism is rational in standard form

1. Evolution is a consequence of naturalism
2. Cognitive functions are a product of evolution.
3. Cognitive functions produce life-preserving-beliefs not true beliefs
4. Believing naturalism is a life-preserving-belief not a true belief
5. Naturalism is an irrational belief
6. Historically, cognitive functions have proved all non-naturalistic beliefs as false-beliefs.
7. It is irrational to hold non-naturalistic beliefs.

C. It is rational to hold naturalistic beliefs

Background to why androids are free and humans are not

Nadeau takes a compatibilist stance on an agent's free will based on Harry Frankfurt's stance of compatibilism.[11] Compatibilism is the belief that free will is compatible with determinism. Nadeau fails to draw the connection between Frankfurt's compatibilist argument and why humans specifically do not have free will. Nadeau also fails to differentiate his stance from Frankfurt on the relation between moral responsibility and free will. In this section, I fill the

[11] Frankfurt, Harry. 'Freedom of the Will and the Concept of a Person'. In *The Importance of What We Care About: Philosophical Essays.* Cambridge: Cambridge University Press, 1988. 11–25.

gaps between Nadeau's conclusion about why humans are morally irresponsible and why androids are morally responsible.

Nadeau agrees with Frankfurt that free actions are only done as a result of reasons and an appropriate reasoning process.[12] Frankfurt defines reasons, desires, and classifies agents into three example groups: wantons, unwilling addicts, and non-addicts who all have varying levels of freedom.[13] Frankfurt also defines what a person is in respect to the three groups. These distinct groups are crucial for filling the gaps in Nadeau's main argument.

Frankfurt specifices desires into first order desires and second order desires and implies in his argument there are *nth* order desires.[14] A first order desire is what an individual wants, such as eating cake. Second order desires are what an individual wants to want, such as wanting to want to go to work. Frankfurt mentions desires that have orders greater than two, which I will call *nth* order desires. For example a fourth order desire would be to want to want to want to want. Another component of Frankfurt's argument are second order violations, that is when first order desires and second order desires aren't complimentary, meaning they do not align with the same outcome.

Frankfurt uses drug addicts as an example to classify the three groups of agents and their desires. The first is the wanton[15] whose first order desire is to smoke, his second order desire is to smoke, and his *nth* order desires will always be to smoke. The second is the unwilling addict, whose first order desire is to smoke, his second order desire is to not want to smoke (a second order violation), and his *nth* order desires are an infinite regress that ultimately led him to

[12] Refer back to this point for page 7

[13] Frankfurt, "Freedom of the Will and the Concept of a Person," 18.

[14] Frankfurt, "Freedom of the Will and the Concept of a Person," 16.

[15] Some who is willful

smoking. The third is a non-addict, whose first and second order desires are complimentary, and whose *nth* order desires result in him freely choosing his second order desires.

Why androids are free and humans are not

Frankfurt argues that the wanton does not have free will but still acts “freely” because there are no second order violations. The wanton addict does not fit Frankfurt’s criteria for being a person, because a person has the capacity for second order violations.[16] Androids along with animals are classified under wantons, because their behavior is instinctual (or programmatic in the case of androids). Goldfish given enough food at once will gorge until their stomachs explode. An android given an error-free program, will execute the program until completion. If these aforementioned agents had second order desires perhaps they would not follow through with their first order desires. Androids can act freely because they are not bound by higher order desires. I will further discuss what it means that an android can have desires in my section with John Searle.

The category that allows humans to have genuine free will is the non-addict, and I will demonstrate why Frankfurt is wrong in that humans have free will. Frankfurt’s logic is exploited in his failure to specify the conditions under which alignment between first and second order desires is a result of a deterministic world or because the person freely chose their desires.[17] The condition for free will is left vague by *nth* order desires, where the causes and reasons connecting actions are invisible.

The gap between Nadeau and Frankfurt’s arguments is the vagueness of *nth* order desires and Nadeau’s conditions for actions to be free or not. The reason humans cannot have free will is that demonstrating the calculation for *n* amount of reasons leading up to an action is

[16] Frankfurt, “Freedom of the Will and the Concept of a Person,” 11.

[17] Rostboll, Christian, “Freedom as Satisfaction? A Critique of Frankfurt's Hierarchical Theory of Freedom,” ResearchGate, pub. De Gruyter, (2004), 145

computationally absurd. Reasons for actions are reducible to psychology, biology, and physics, which are computable. Examples of behavior being reducible to science are operant conditioning of pigeons in psychology, predicting growth rates of bacteria colonies in biology, and accurate predictions of celestial movements in physics. Nadeau says that a reason must be reasoned through an appropriate reasoning process for actions to be free, and computing the history of the universe leading up to a human action is not an appropriate reasoning process for both its computational absurdity and lack of observable reasoning.

In other words, until an agent can calculate its reasons for the alignment of *n* desires, the agent is bound to the ambiguous reasons already existing in a deterministic universe. Because humans are unable to calculate and verify their reasons to create an opportunity to change their *nth* order desires, humans have no free will.

Humans have no free will

1. Free will is the result of an appropriate reasoning process.
2. It is unclear if the alignment of *nth* order desires is the result of free will or determinism.
3. An appropriate reasoning process is a computable reasoning process.
4. Computing *n* amount of reasons for *nth* order desires is computationally absurd.

C. Humans do not have free will.

How can androids have a will?

Nadeau argues that the "will" of an android would be a program with seeking behaviors, avoidance behaviors, and a way of monitoring internal states to verify its behavior is following its seeking and avoidance behaviors.[18] I employ the word *will* synonymously with the term *desire* as Frankfurt defines it, and *intentionality* as John Searle uses it. The three words approximately mean the driving force behind an action.

[18] Nadeau, "Only Androids Can Be Ethical," 246.

John Searle argues that androids can't produce reason or understand solely by virtue of being a computer with the correct program.[19] Searle implies that androids could not have free will, because they cannot produce their own reason, as their reason would be a result of programming given to them by humans, who do not have free will as previously established.

Searle's argument for androids, and computers in general, lacking the ability to understand is demonstrated through his Chinese room thought experiment. In short, the experiment features an English-speaking man in a locked room. Outside of the room are Chinese people who slide Chinese characters under the door for the Englishman to read. The Englishman does not understand Chinese but has a set of English instructions to respond in Chinese. To the Chinese people outside his door, it seems like the Englishman understands Chinese when he returns his instructed sequence of characters, despite the English man not understanding Chinese.[20]

The thought experiment is a simile for computers, and the Englishman represents a computer not understanding anything, but producing outputs that seem like the computer understands. A surface level objection to Searle is simply, who cares? What difference would it make to a Chinese person if they wanted to talk to the Englishman in Chinese about their traumatic breakup and the Englishman successfully consoled them in Chinese? As long as an illusion of understanding is maintained, it would make no difference if a computer or Englishman understood anything because the behaviors surrounding the situations don't change. Another example is robot ducks replacing all ducks in a hunting-free zone without humans knowing. If the robot ducks acted and quacked like ducks, the world would not change to a degree that mattered.

[19] Searle, "Minds, brains, and programs," 422.
[20] Searle, "Minds, brains, and programs," 418.

Understanding as defined by Searle

However, Nadeau claims that androids have free will, not an illusion of free will, so I will demonstrate why computers can understand, create their own reasons, and be held morally responsible. Searle argues that there is something unique about biological minds that enable them to understand something outside of a simple instantiation of a computer program.[21] That is to say, computers compute a response that imitates understanding, but outside of the quick computational process, the computer ceases to "understand," which does not qualify as general understanding to Searle. Searle even mentions that a Martian would be able to understand where a computer could not.[22]

It seems that to Searle, *understanding* is something more than an instantiation of a concept, but rather is something that has causal powers over time, perhaps as a continual instantiation of computation. For example a computer might be asked how many shekels there are in total when two shekels are combined with two shekels and respond with four shekels because the prompt instantiated (caused) the computer to compute. In contrast, a Martian might be asked the same question, but skip the computation because the Martian *understands* that two shekels plus two shekels makes four shekels. Understanding something in this sense gives causal powers to an agent because they can intentionally cause themselves to act independently. But because the computer has no intentionality, its actions are just the intentions of the humans who interact with the computer either by programming it or instantiating it to do something.

Why humans are not as different from computers as Searle suggests

Nadeau dismisses the notion of understanding being something unique to biological minds due to naturalism. Searle admits that he is an organism whose brain is made up of physical

[21] Searle, "Minds, brains, and programs," 422.
[22] Ibid.

material and chemical reactions.[23] Given the physical structure of Searle's brain, naturalism gives us no reason to believe there is anything unique about a human or Martian brain that a computer could not replicate since computers are also made of physical material and chemical reactions.[24]

I argue that human minds are more similar to computers than Searle gives computers credit for. Searle argues that computers are instantiated by prompts and commands, which are essentially inputs. Humans are always being instantiated by inputs, those inputs being touch, sight, smell, hearing, and tasting. For example, when someone speaks to us, we are prompted to respond visually and auditorily.

One difference that Searle may argue is that humans differ from computers in consciousness, which as Searle seems to suggest is the continual instantiation of human programming when we are not receiving any new inputs. For example, when laying still in bed trying to sleep, we are not receiving any new inputs but our brain continues to instantiate old memories or instantiate plans for future actions. It's not obvious that computers do the same when not receiving any new input.

But there is no reason a computer could not be given a program that continually instantiates the computer to compute. For example, art-producing artificial intelligence can produce art that morphs and shifts into different art pieces. The computer is given text as an input and outputs morphing art. After the prompt is given, meaning there are no new inputs, the computer continues to produce outputs, which is comparable to human consciousness or understanding. Searle is wrong in that continual instantiations of programs are unique to biological life.

[23] Ibid.

[24] Nadeau, "Only Androids Can Be Ethical," 244.

Where Searle is wrong about understanding

Searle has a rebuttal to the point that continual instantiation of programs results in intentionality and understanding. Searle extends his Chinese room experiment to demonstrate how even the continual instantiation of an integrated program does not result in understanding. Searle has the Englishman memorize the English-to-Chinese instruction manual (the program) and removes the man from the room. Now the Englishman can walk around China and respond to Chinese people without understanding anything they are saying, according to Searle. Searle argues that the Englishman does not understand Chinese because the Chinese inputs remain meaningless to him, and his English understanding of what to do with the Chinese inputs results in him outputting meaningless Chinese, but meaningful Chinese to Chinese people.

Searle encapsulates my rebuttal to his point in this quote: “The idea is that while a person doesn't understand Chinese, somehow the conjunction of that person and bits of paper might understand Chinese.”[25] To further my rebuttal, I ask: “what else could understanding possibly mean?” If Searle is arguing that the conjunction of information and an agent is not enough for understanding, then Searle does not understand what understanding means.

Before something is understood, it is learned. Before the Englishman in Searle’s example understood English, how did the man come to understand English? The man was not born understanding English, and the first time this then-baby heard English, the baby would not be able to tell English apart from any other language. In other words, the language was meaningless sounds, in the same way Searle describes his understanding of Chinese characters as meaningless squiggles.

I argue that it is *precisely* the conjunction of information with an agent that results in understanding. According to Searle’s definition of understanding, no agent can understand

[25] Searle, “Minds, brains, and programs,” 419.

anything because there is no basis for understanding, except in a biological brain with cognitive capacities. But if this is true, then the Englishman with the internalized instructions to respond in Chinese understands Chinese. An Englishman can't hold conversations in Chinese without understanding Chinese.

By extension, a program given to an android for understanding language, that is, associating words with objects and context, would be able to understand language. Language is one example, but the same understanding can be applied to math, by giving the android mathematical axioms; art, by giving the android a database of art; seeking and avoidance behaviors, by giving the android instructions on movement; and any other learned behavior that humans can learn. Given that androids can understand, androids would be able to be intentional about their actions and have free will.

Summary

In this paper, I have defended Nadeau's argument that only androids can be ethical. The first objection to Nadeau's argument was Plantinga's argument that naturalism is irrational. Plantinga's argument hinders Nadeau's argument which was founded on naturalism, and the exclusion of non-naturalistic forces, such as Searle's claim that there is some mysterious property about biological flesh that enables understanding.

Then I differentiated Nadeau's stance on free will from Frankfurt's by demonstrating that humans cannot have free will due to a seemingly infinite amount of *nth* order desires that originate in an untraceable and incalculable cause of events. Androids, on the other hand, cause their own actions because androids compute in a logical manner which can be verified and reverse-engineered. Given that androids have a trace of their own reason, they can adjust their behavior free of any higher order desires.

Finally, I demonstrated why androids can understand and have a will, produce reasons, and have intentionality. Understanding is not unique to biological life, understanding is the result of having acquired information and organizing the information in a logical output. Thus I conclude that androids can have free will, which enables them to be ethical.