Can Only Androids Be Ethical?

Introduction

In this paper, I will be analyzing Joseph Nadeau's argument in his journal article titled "Only Androids Can Be Ethical"[1]. I will lay out Nadeau's premises and investigate the accuracy of the premises. I will also be elaborating some of Nadeau's key examples which support the premises. Nadeau's argument is that responsibility is only possible through free will. Actions are only free if the actions are done for reasons and appropriate reasoning. Human reason is a retroactive memory disorder. Nadeau at the previous premise mentions the Entscheidungsproblem which I will go into more detail about the problem's implications. Android reason is done through theorem provers and neural nets, which can be empirically observed. Therefore androids can be held responsible for their actions which makes androids capable of ethical and unethical behavior, and humans cannot. Nadeau then speculates on how to make androids ethical rather than unethical giving androids "emotions" and a moral framework which I will also analyze.

I will now also draw a distinction between ethical behavior, and the capacity for ethical behavior. Nadeau is not saying only androids can behave ethically, Nadeau is saying only androids have the capacity to behave ethically.

Responsibility is linked to capacity and free will

Nadeau first mentions that for an agent to be held responsible, the agent needs to have the capacity for free will. Free will is used in the same way that Rene Descartes uses the term free will. Descarte says free will is "the ability to do or not to do something"[2]. By Descartes definition of free will, if an agent lacks an ability of choice, then the will is not free. Different

[1] Ford, K. M., Glymour, C. N., & Hayes, P. J.
[2] O'Connor, Timothy and Christopher Franklin

situations give agents different choices. I will give two examples of similar situations in which an agent has free will, and a situation where the agent does not have free will.

Imagine a spy has been caught by the enemy without having the time to eat their cyanide pill. The spy is taken to the torture chambers to be extracted of sensitive information which would compromise the spy's country. The spy has the choice between revealing the information, or being tortured by the enemy. Although the torture will permanently damage the spy physically and mentally and is not a favorable choice, the spy is held responsible by their country to not be betray the information. Moreover, the spy's will is not being constrained by variables out of the spy's control. The spy is given the choice by the enemy to either betray the sensitive information or to experience torture.

Imagine the same scenario but instead of torture, the enemy informs the spy that the spy will be injected with a truth serum. Once injected with the serum, the spy will speak the truth on any questions asked. The enemy proceeds to inject the spy, and asks the spy about the sensitive information and the spy reveals the sensitive information. The difference between these two scenarios is the truth serum. The truth serum changes the situation the spy is in by removing choice from the spy. Without choice, the spy does not have free will and therefore the spy's country would not hold the spy responsible for revealing the sensitive information. The two examples above are not to make a case for free will or against free will, but to draw a line between when actions and choices are free, and when actions and choices are not free. The notion of free actions is important for understanding Nadeau's next point.

Free action is done from reason

Nadeau's stance on free action aligns in part with arguments from the compatibilists Harry Frankfurt and Susan Wolf which Nadeau cites. I will go into the differences between

Frankfurt and Wolf's arguments and the significance of the arguments in relation to an agent's choices. In different ways, all three philosophers agree free action must be done from reason and "an appropriate reasoning process"[3].

Frankfurt's stance is that an agent acts freely when the desire on which the agent acts is one the agent desires to be effective[4]. If a person is in pursuit of health and spends the day on their couch drinking beer and smoking cigarettes, then the person is acting in accordance with their free will. Even if the person had other desires such as health, the desire which the person wanted to be effective was the desire for drunkenness and nicotine. Should the same person wake up the next day and go to the gym, the desire which the person wanted to be effective was the desire for health. In summary Frankfurts point is that reasons for actions come from desire, which is in ourselves and psychology, not something outside of ourselves[5].

Susan Wolf also argues that free action is done from reason, but the reasons Wolf argues for are external. Instead of agents acting freely from an agent's own desires, agents act freely in accordance to what is Good and True[6]. There is one reservation in Wolf's argument which is that praiseworthy actions do not need to be done through a choice, while blameworthy actions need to have been chosen. The person on the couch pursuing health and going to the gym that day would be acting freely even if there was no other choice but to go to the gym (assuming health is a part of the Good). Counterintuitively, Wolf asserts that doing Good and True actions does not require the ability to do otherwise because acting freely is not enhanced by the ability to act irrationally (against the Good and True).

---

[3] Ford, K. M., Glymour, C. N., & Hayes, P. J.
[4] O'Connor, Timothy and Christopher Franklin
[5] McKenna, Michael and D. Justin Coates
[6] McKenna, Michael and D. Justin Coates

However when choosing blameworthy actions, say drinking beer and smoking cigarettes harming one's health does require the ability to do otherwise. Because drinking and smoking is against the Good and True, there is the option of doing the Good and True, which is choosing against drinking and smoking and going to the gym instead. To summarize Wolf's argument, reasons for free actions come from the Good and True.

Free actions being the consequence of reason, whichever those reasons may be is the basis for Nadeau's notion of responsibility. However another metric by which Nadeau qualifies free actions is an appropriate reasoning process. In the next section, Nadeau explores how appropriate a process humans reason through. After laying out these definitions and perspectives on free actions and responsibility, I can begin to speak about where androids fit into the picture. We will come back to Frankfurt and Wolf's ideas about reasons when speaking about androids and programming androids to be ethical rather than unethical.

The Entscheidungsproblem Preface

Nadeau's next point is that human's logical reasoning is severely limited. To provide evidence humans cannot reason logically, Nadeau introduces Alan Turing's Entscheidungsproblem theorem. Although the Entscheidungsproblem regards computers and not humans, Nadeau compares humans and computers in terms of logical reasoning. Nadeau is not reducing humans to living and breathing computers.

The Entscheidungsproblem is known in English as the halting problem, or the decision problem. David Hilbert, mathematician of the 19th century who may also be known for the Hilbert's Hotel paradox, wondered if we could ever determine if any given statement is provable or non-provable. Hilbert's question is significant because if agents are to be held responsible for

their actions/statements, their reasoning for their actions/statements should be able to be proved as logical consequences of their situations.

Nadeau's comparison between humans and computers was closer to reality than the comparison is today, as in the time of the Entscheidungsproblem modern computers did not exist. Instead, a computer was a type of job for a human to do, which was compute data. Alan Turing found computing interesting and spent time trying to find a way to automate the computing process using machines, which gave rise to modern day computers.

Alan Turing invented a theoretical machine called the Turing Machine, which is essentially a computer at a computer's most fundamental level. The machine has a tape with cells filled with 1's and 0's, a head which moves along the tape capable of changing 1's and 0's, a set of instructions (a program) and a state register (memory). Computers are an automated process of computation, and computation sometimes involves loops. A human example of a loop would be re-reading an essay and checking spelling errors. Say a student is done reading the essay from top to bottom, finds a few mistakes but still is not satisfied, the student may loop back to the top of the essay and proofread again.

When a set of instructions given to a computer are computationally intense, the computer may take longer to complete the instructions than less computationally intense instructions. However when the instructions are unclear or not logical, the computer can begin an infinite loop. In the essay example, say a professor reads the student's essay, and finds a mistake that is in actuality not there. The professor then gives the student faulty instructions to re-read the essay and find the mistake. There is no mistake, so the student (if infinitely dogmatic and immortal) theoretically re-reads the essay infinite times trying to find a mistake which is not there.

However when instructions are clear, an agent will take X amount of time and when the instructions are fulfilled, the agent will halt.

Hilbert's question was if *any* given statement was provable or non-provable. Turing with his concept of the Turing Machine rephrased Hilbert's question in computational terms, to "can we determine if any given program will halt or not halt?" Halting is good, because halting means the instructions are fulfilled, and not halting is bad because not halting means the computer is in an infinite loop.

The Entscheidungsproblem and its implications

Turing's answer to Hilbert's question was no, we cannot determine if any given statement is provable or non-provable. Turing's answer comes from a thought experiment. Program A is a set of instructions which takes in data and returns modified data. The instructions of program A is to determine if data put into program A will halt, or loop. Program B is a set of instructions which takes the output data from program A, and returns the opposite data. If program A determines the input data will loop, the data will be inputted into program B, and program B will output that the data will halt. If program A outputs the result that the data will halt, program B will loop.

Program B takes program A's outputs, so we can simplify program A and B by calling the programs program C. Programs can be reduced to sets of instructions which can further be reduced to data. Therefore program C is just data, which means program C can take its own data as an input. When program C takes its own data as an input, what is happening inside of program C is if program A determines the data will halt, program B will return that the data will loop, which becomes the new input of program A. Program A takes the looping result and outputs the

loop to program B, which will return the result halt. Program A then takes the result halt, and program B returns the result of loop.

The behavior of program C is not a loop but a logical paradox. The purpose of program A is simply to determine if data will halt or loop. Program B arbitrarily outputs the opposite data as program A since an inverted result should not be consequential to the overall logic. The purpose of program C is the same as program A. The paradox is that program C contradicts its own results. Program C will stay in a state of looping and halting which is impossible. With the previously mentioned thought experiment, Turing showed that computers cannot determine if a program will halt or loop, and as a consequence answered Hilbert's question that we cannot determine if any statement is provable or non-provable, even with infinitely powerful computers.

Referencing the Entscheidungsproblem theorem is how Nadeau claims humans are not bound by reason. Nadeau assumes human reasoning is inferior to computer reasoning, in terms of accuracy of results and time taken to reason. If computers cannot logically reason through all data given as an input, a human would be even less capable of reasoning through any data given as input (inputs being sight, smell, touch, taste, and noise). Nadeau's reference to the Entscheidungsproblem theorem must be for skeptics of human irrationality, as figures such as Mark Twain and most European psychoanalysts take for granted that humans are not bound by reason.

In conclusion of Nadeau's point of human reason and the Entscheidungsproblem theorem, Nadeau argues that humans cannot be held responsible for their actions because human action is not reasoned through an appropriate process. Because humans cannot be held responsible for their actions, humans cannot be assessed as ethical/unethical agents. In Nadeau's next point, Nadeau argues why androids can be held responsible for their actions.

Android reason through theorem provers

In contrast to human reasons being "ex post facto confabulations"[7], android reasons would be reasoned through theorem provers, neural nets, or a blend of the two. Theorem provers are programmed algorithms designed to automate reasoning. Problems given to these theorem provers consist of two parts. The first part is a question phrased as a statement called the *problem's conclusion*, followed by statements which are the relevant information for the theorem prover to assume in order to determine the truth of the conclusion (called the *problem's assumptions*). The goal of the theorem prover is to prove the conclusion from the given assumptions by applying rules of deduction programmed into the theorem prover's algorithm.[8]

The reason a theorem prover is superior in reasoning to human reasoning according to Nadeau is because programs leave a trace. A proof of a theorem is essentially a trace of information which begins at a conclusion and logic is used to verify the truth of the conclusion given the relevant axioms. Humans also have the ability to make conclusions given a set of axioms, but outside a logical framework such as first order logic human brains have too many computations which obscure a trace of reason. A theorem prover on the other hand only has one purpose which is to prove through computation a conclusion derived from premises.

The trace of reasoning is important because the trace is what allows other agents to assess if the reasoning is done through an appropriate process. The appropriateness of a reasoning process is important for determining if an action was good or bad. A real world application of reasoning traces and appropriateness would be investigators at a crime scene. Say a woman killed her husband. The woman is not immediately sent to jail for doing something "bad" but instead sent to a questioning room to try and discover her motives. Maybe the husband was

[7] Ford, K. M., Glymour, C. N., & Hayes, P. J.
[8] Portoraro, F.

abusing the woman and their children for years, and one night she committed a crime out of passion. The reasoning for the woman's crime changes the verdict of the woman's sentence in court. On the other hand if the woman had killed her husband when she walked in on her husband cheating, the verdict of her sentence would be different because her reasoning process may have been less appropriate.

However in neither situation is the woman using first order logic or some type of reversible engineering process to decide what the best course of action is. In the sense of the woman not using a traceable logical framework to make decisions, the judge must rely on their own untraceable logical framework to deliver the woman's sentence. Human reasoning sometimes is traceable in small amounts when studied closely, and discoveries of reasoning processes are questionably appropriate. For example, judges are more lenient on sentences after eating, or taking a break.[9] Correlation does not equal causation however, and a judge's leniency is not necessarily caused by food or breaks. Which further shows the unclarity of human reasoning.

Nadeau argues that since theorem provers have a logical trace, building androids around theorem provers is the first way androids could be held accountable for their actions. The logical trace could be verified by any agent including the android itself. An android knowing its own reasons for doing things is valuable because the android can double down on its responsibility, which reinforces Nadeau's argument.

Android reason through neural networks

Neural networks as a means of reasoning differs greatly from theorem proving algorithms. Neural networks are *input nodes*, connected to *output nodes* by *hidden nodes*. The nodes are assigned activation values given by real numbers which figuratively have a "weight"

[9] Levitt, S. D., & Dubner

of importance. The activation of a neural network starts with giving an input node a value, which is mediated by a hidden node to pass the value to an output node, which passes the value through as many nodes as the network needs to finish computing.[10]

In simple terms, theorem provers try to arrive at conclusion C given some axioms A and B by using logic. Neural networks try to arrive at conclusion C given some weights A and B by arriving at some conclusion D, and readjusting weights within its neural networks in order to shift result D closer to result C. Put another way, neural networks pivot their *actual outputs* to get closer to the *desired output*. Neural networks such as the one previously mentioned are called learning algorithms.

In neural networks reasons would occur as data structures.[11] Input data through the neural network would cause the network to shift weights around its nodes once the first iteration is complete. The shift in node weights would be one data structure, and upon the second iteration the node weights would shift again adjusting the actual output closer to the desired output which would be the second data structure. The process would repeat itself until the actual output is within an acceptable range of the desired output, and each iteration through network nodes would be equivalent to each line of a logical proof. Once again there is a logical trace which actions based on these reasons would be sufficient to hold an android responsible for its actions.

Making responsible androids ethical

As addressed in the conclusion, responsibility does not equate to ethicality. Some serial killers take full responsibility for killing in cold blood through a questionable reasoning process, which of course does not make their actions ethical. Nadeau argues that to make androids behave in an ethical manner, androids must possess a moral theory through which to operate from. In

[10] Rescorla, M.
[11] Ford, K. M., Glymour, C. N., & Hayes, P. J.

addition to a moral theory, Nadeau argues androids must also possess a theory of other's minds, including emotions.[12] The reason possessing a theory of minds and emotions is relevant for behaving ethically is probably because ethics exist in relation to humans. An android would therefore need to possess a theory of human minds and emotions in order to behave in a way which humans would consider ethically acceptable.

Nadeau's argument that androids need a moral theory seems uncontroversial, but possessing a theory of minds may seem complicated or far-fetched. Some concerns may be that Nadeau is saying androids need to be conscious to be ethical. Or that verifying if the android really "knows" a moral framework/theory would be impossible. However these concerns are not important because the topic of Nadeau's argument is about androids behaving ethically, not about android consciousness or epistemology.

Possessing a theory of mind and emotion would simply be another computation which the android would be created with, or something that is learned. Nadeau refers to folk psychology when elaborating on how androids would learn a theory of mind and emotions. Nadeau claims psychologists have come to the thesis that young children are not born with a theory of mind and emotions. Instead children learn about mind and emotions by comparing their own mind and emotions with other people.

Nadeau says that there would be two ways of teaching androids these theories. The first way would be a computability theory of mind and emotions. A key part of a computability theory would be behavioral evidence for an "understanding" of mind and emotions, and another program which would enable the android to act on its computability theory of mind and emotions. Computing emotions would start with propositional attitudes which would be programmed into the android. A propositional attitude is the relation an agent has to certain

[12] Ford, K. M., Glymour, C. N., & Hayes, P. J.

propositions. For example an android with the propositional attitude that snow is frozen water, may avoid snow to avoid damaging itself due to cold or humidity. Self preservation being good is another proposition that could be programmed into the android, which is why the android would “care” about avoiding damaging itself in the snow in the first place.

The second way androids could learn a theory of mind and emotions would be similar to the way children learn mind and emotions. The android would be programmed to compare its own inner states with the states of other agents. In a similar fashion to neural networks, an android could calibrate its behavior to fit a conception of acceptable behavior. Using the snow example, perhaps the android would observe how uncommon it is that adults run around in and jump in snow. Using observation and perhaps a few errors, androids could form a theory of mind and emotions. The android would not need to understand the theory, but just have a framework so the android could behave as if it understood a theory of mind and emotions.

Another concern which may arise against Nadeau's argument for implementing a theory of mind and emotions is that androids would cease being logical and begin acting from mind and emotions rather than reason. To me it seems that Nadeau is less concerned with android behavior and more concerned with the process which drives behavior. In other words, if an android were to do something which seemed unethical, we humans would be the ones who have incorrectly assessed the “unethical” action. In the sense that humans can imagine androids doing horrible things and what the androids are doing is *the* ethical thing to do, Nadeau’s argument can be unsettling.

However I believe that the reasons we do not want androids doing seemingly unethical things, is why Nadeau argues androids need a theory of mind and emotion, not just an ethical framework. If androids had a theory of mind and emotions (human minds and emotions),

behaving in a way which is deemed ethically unacceptable by surrounding agents would discourage the android from repeating such behavior.

Choosing an ethical framework

Now that androids understand humans and emotions, androids would need a reason to behave in a way which humans consider ethical. Nadeau explores a few possible frameworks, mainly virtue ethics and utilitarianism. Virtue ethics is a rigid ethical framework in which axioms must be followed absolutely. Nadeau rejects virtue ethics as a viable framework for androids as the rigidity of the framework could create conflicts within the android, paralyzing the android. A human example of the problems of conflicting virtues are medieval knights. If the knight takes vows to serve his king, and also takes a vow to protect the innocent, when the king on a whim orders some peasants to be executed, the knight must break a vow. An android however may get stuck in a computational loop about how to move forward if given such a dilemma with virtue ethics.

Nadeau also contemplates utilitarianism. Utilitarianism is a moral framework which aims to do the most good for the most people. A utilitarian agent would assign a utility value to all given actions and consequences then act upon whichever set of actions results in the maximum utility. The main issue Nadeau brings up regarding utilitarianism is the impossibility of computing the utility of all possible action and consequences.

However there is one branch of utilitarianism which Nadeau finds fitting since artificial intelligence works with heuristics, and there is a branch of utilitarianism which works in heuristics. [13] Rule utilitarianism is a branch of utilitarianism which operates based on prior information. Patterns and behaviors which cause good things and bad things are taken into memory, and repeated or stopped depending on the outcome of the situation.

---

[13] Ford, K. M., Glymour, C. N., & Hayes, P. J.

For example if an android were driving a car and arrived at a stop sign, from prior experience the android may have a memory of almost getting into a car accident because cars were driving very fast on the road. So the android must stop and wait until there are no more cars before proceeding. However in rule utilitarianism, rules can also be overridden if the situation's outcome is worse if the rule is not overridden. If the android is driving the car to get a woman who has just begun to give birth and is ill, the android may drive faster than the speed limit, and forgo stopping at a stop sign if it sees there are no cars anywhere close. Overriding the rules in the previous example is how an android, understanding mind and emotions, can calibrate itself to act appropriately to a given situation.

Conclusion

Bibliography

Ford, K. M., Glymour, C. N., & Hayes, P. J. (2006). *Thinking about Android epistemology*. AAAI Press (American Association for Artificial Intelligence).

McKenna, Michael and D. Justin Coates, "Compatibilism", *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.)

Portoraro, F. (2019, March 2). *Automated reasoning*. Stanford Encyclopedia of Philosophy. Retrieved November 8, 2022, from https://plato.stanford.edu/entries/reasoning-automated/

Levitt, S. D., & Dubner, S. J. (2007). *Freakonomics*. Denoël.

Rescorla, M. (2020, February 21). *The computational theory of mind*. Stanford Encyclopedia of Philosophy. Retrieved November 8, 2022, from https://plato.stanford.edu/entries/computational-mind/