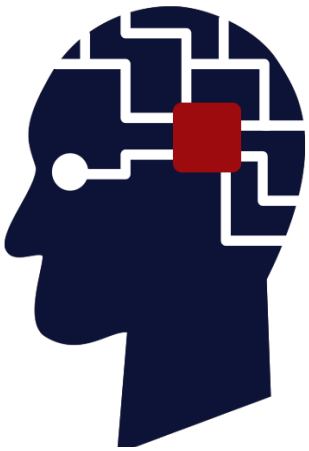

BENEFICIAL AI SOCIETY | ETHICS FOR COMPUTER SCIENTISTS

WRITTEN BY Olivia Gambelin

EDITED BY Anna Van Oosterzee
Woodley Brown
Luca McArthur
Ashe Magalhaes

DEAR READER,

Welcome to the handbook on 'Ethics for Computer Scientists'. This handbook was written for the launch of the Beneficial Artificial Intelligence Society (BAIS), based at University of Edinburgh. This society was founded to bring together a range of disciplines to discuss the implications of AI for our society. To inform our discussions we wrote two handbooks on Computer Science and Philosophy.

The handbook you are reading right now was written to explain jargon from classical ethics. 'Machine Learning for Philosophers' is meant to complement this handbook and was written for everyone who is not familiar with machine learning. These handbooks are meant to give you a brief overview of the current day debate of AI, to help you kickstart thinking about all the implications it could possibly have for our future.

1 WHAT IS ETHICS?

Have you ever found yourself next to your partner's open laptop when you hear the familiar ping of a Facebook messenger notification? Part of you says read it, as long as your partner doesn't find out there are no real consequences, whereas the other part is saying don't do it because it would be wrong. This conversation you are having with the voices in your head is what we would call a philosophical debate in ethics, as you are essentially deciding what the right and wrong actions to take in this situation would be. Ethics is the field of philosophy that is concerned with the concepts of right and wrong behavior. According to our modern understanding of ethics, this field can be divided into three general subject areas: metaethics, normative ethics, and applied ethics.

Metaethics deals with the high level questions like what do the concepts of right and wrong actually mean and where did they come from? In our Facebook message example, a meta-ethicist would be too busy questioning what is good and how did we come to use this term to describe certain actions to notice there was a message notification in the first place. Normative ethics, on the other hand, deals with more practical questions such as what actions are inherently right and which are inherently wrong. Is reading that message wrong behavior or is it the right behavior required to know what's really going on in the life of your loved one? Finally, there is applied ethics, which deals with specific questions raised by specific controversial issues such as abortion, euthanasia, and the digital privacy.

2 METAETHICS

For the sake of this handbook, and the sanity of both the authors and readers, we will keep the discussion of metaethics brief, as it is the subject area that gathers the least amount of attention outside of the philosopher's armchair. Essentially, the central theme metaethics deals with the debate over the

existence of moral facts. Moral facts are the facts by which we determine good and bad values. According to Realist philosophers, there are objective moral facts, objective meaning existing independently of the human mind, that govern what is right and wrong. Broadly speaking, Realists are divided into Naturalists and Non-Naturalists. Naturalists claim that these facts exist within the natural world, much like scientific facts, whereas Non-Naturalists believe that the origins of these facts are, you guessed it, nonnatural and instead come from something more along the lines of a spiritual world. On the other hand, there are some philosophers who argue that moral facts do not exist at all but that instead our conception of right and wrong are born solely out of cultural, personal, or social influences. These philosophers are widely characterized as Relativists, since they believe moral understanding depends on personal understanding, which is a subjective experience.

3 NORMATIVE ETHICS

Metaethics questions if moral facts exist, whereas Normative ethics in most cases assumes the existence of certain moral facts and makes use of them. Normative ethicists aim to arrive at one single criterion, whether it is a solitary principle or a single set of fundamental principles, that has been deduced from moral facts and by which we can regulate what is right and what is wrong. The Golden Rule is a great example of this. A person can decide if an action is right or wrong based off of whether they would want the same action committed to them or not. In the search for this great single criterion philosophers have pushed the development of three main theories: Virtue Theory, Duty Theory, and Consequentialist Theory.

3.1 VIRTUE THEORY

This theory is the oldest of the Western tradition, as its roots can be traced back to our favorite ancient Greeks: Plato and

Aristotle. According to virtue ethics, learning rules such as "don't kill people" and "don't read someone else's Facebook messages" are less important than forming good habits of character, or virtues, while avoiding bad habits of character, or vices. The idea is that once a virtue such as courage is learned, the person will then continue to act in a courageous manner throughout their life. It is important to note that the virtues of this theory exist on a relative scale between two extreme vices. Take again the example of courage. When a person exhibits too little of courage they begin to show signs of cowardice, which is a vice. However, when a person exhibits too much courage they become rash, which is also a vice. Thus, in order for a person to properly exhibit the virtue of courage, they must find the balance between the two extremes of vices. This is why the process of ethical education is so important in virtue theory, as a person is not learning rules but rather how to balance vices and virtues.

3.2 DUTY THEORY

Duty Theories are also known as deontological ethics, as the emphasis is on the obligation, or duty, of a person to take certain actions that are inherently right versus avoiding actions that are inherently wrong, irrespective of the consequences. For example, if you were following a duty theory, you would not read your loved one's message as you have the duty not to invade the privacy of someone else, even if the message you want to read would help you help your loved one. The main point about duty theories is that they assume that duties are inherent facts in nature.

3.3 CONSEQUENTIALIST THEORY

Inherent in the name, Consequentialist theories emphasize the consequences of actions over the value of an action itself. In this theory, the morality of the action is not determined by

the action or the decisions surrounding it, but rather by the outcome of the action taken. It works much like a point system, you add up all the positive and negative consequences of an action and then decide the moral worth of an action based on whether it has more positive points or negative points. There are different forms of consequentialist theories. For example, Ethical Egoism proposes that maximizing the outcome for yourself is best even if it is at the cost of others. Whereas Ethical Altruism proposes that maximizing the outcome for others at the cost of yourself is best, even if it ignores your desires. Then there is Utilitarianism, this is a consequentialist theory that has been growing in popularity recently in conversations surrounding technology. This is the idea that all our actions should maximize the greatest possible outcome (or utility) for all those affected by the action. Given that this theory is so popular in recent technological developments, we have provided the reader with its own section.

3.4 UTILITARIANISM

Utilitarianism rests on two primary assumptions: (i) that universal happiness for all is what human beings desire, and consequently, (ii) that the only way to work toward such an ideal is to maximize pleasure and minimize pain for all parties concerned. The utility of an action is the quantity of pain or pleasure it produces. Therefore, in this context, the concept of "happiness" is the greatest amount of pleasure for the greatest number of people. Although there remains a raging debate among philosophers as to whether this is an adequate and fair definition of happiness, due to its inclusivity, utilitarianism has become increasingly popular in today's society.

Accordingly, learned utilitarians fall into one of two camps when it comes to the method by which the maximization of utility for all is calculated. Restricted Utilitarians consider that the best way to maximize happiness overall is to follow a set of rules and procedures which are known to increase utility

whereas Extreme Utilitarians prescribe that there a no rules and instead that all right actions generate either an equal or greater amount of utility than would any other possible action.

Although Utilitarianism may seem to be the best possible answer to many ethical problems due to its unbiased nature, there are some significant potential negative consequences to this theory. Take for example Thanos in the Avengers Infinity War movie. Thanos uses Utilitarian reasoning to come to the conclusion that the only solution to solve the overpopulation crisis is to kill off half of all living beings. Crazy, right? Well, by utilitarian thinking, he's actually got a solid point. By killing off half the people, Thanos increases the happiness of the half left alive far beyond what the total amount of happiness would be if everyone was kept alive in an overpopulated world. So although Utilitarianism does seem to provide good solutions to many modern day problems while keeping things fair and equal, it is not our final and absolute solution.

4 APPLIED ETHICS

Finally, we have Applied Ethics, the form of ethics that everyone participates in whether or not they have ever had the experience of getting into an ethical debate with a philosopher. Issues covered by Applied Ethics have two criteria they must meet. First, the issue must be controversial, and second, the issue must be distinctly moral. Our example of the Facebook message is an instance of an Applied Ethical issue as it is controversial in so much as you and your loved one may be in contradiction over whether it was the right action or not, and it would be distinctly moral as you would be causing harm to your loved one by committing such an action. Other modern examples of Applied Ethical issues are abortion, animal rights, gun control laws, and euthanasia. All of these issues fit under subcategories of specific ethical fields. Currently, there are five traditional identified ethical fields: Biomedical Ethics, Business Ethics, Environmental Ethics, Sexual Ethics,

and Political Ethics. However, there is a sixth field that is quickly developing into what will be one of the most important and influential fields. This is Technological Ethics, better known as the Ethics of AI.

5 THE ETHICS OF ARTIFICIAL INTELLIGENCE

With technology developing at breakneck speeds, it has already outpaced our traditional ethical understandings. As we shall see, thought experiments that have been traditionally confined to the philosophical armchair are now becoming real life problems that we need to provide solutions to. Part of the difficulty surrounding this field comes from defining the moral status of AI, as we are still determining at what point, if at all, does an AI have the same moral status as a human, and if this is possible, then what the heck do we do with it? For example, if an AI has the same moral status as a human, it would be wrong of us to turn it on and off as we so willed, and restarting it would be on par with murder. Even if we don't assign moral status to AI, we are still faced with a plethora of ethical problems such as who has the ultimate responsibility for the AI when things go wrong, do we allow AI to make moral decisions for us, and how do we eliminate negative bias? Many of these growing ethical problems can be highlighted in new iterations of historic philosophical thought experiments. Two of the most prominent thought experiments pertaining to ethics, as well as the moral status of AI, are the Trolley Problem and the Brain in a Vat.

5.1 THE TROLLEY PROBLEM

The Trolley Problem was born in 1967 from the mind of Philippa Foot, although similar experiments can be traced all the way back to 1905. Today however, we know the Trolley Problem by a new name: the Self-Driving Car Problem. Picture

this, you are riding down the road in a self-driving car when all of a sudden a close friend and a group of six strangers try to cross the street in front of you from different directions. There is no way around it, you must either run over your friend or the group. So, which is it? What is more valuable, the life of a loved one or the combination of six lives? According to the Ethical Egoist, the life of your loved one would be more important because it is more beneficial to you than the lives of six strangers. On the other hand, the Utilitarian would argue that six lives will always be of greater value than one life, irrespective of who the people are. As you can imagine, there are countless ways in which to manipulate the circumstances of this experiment, however they all point to the same purpose of testing the different ways in which you value different lives, which in turn can highlight the contrasts in ethical understandings of individuals.

If this wasn't enough ethical turmoil for you, then consider how the addition of the self-driving aspect of the car in the revised Trolley Problem affects your moral responsibility. In the previous versions of the Trolley Problem, you would have been the active agent making the decision of who to kill and who to save by pulling a lever that would direct the car one way to kill your friend, or the other way to kill the group. This means that you would be morally responsible for whichever life you chose to take. However in the modern self-driving car version, you are no longer the active agent. Instead, it is the car that makes the decision for you of who to run over. This leaves us with the question, who is morally responsible? It seems silly to say the car, so then do we hold the programmers responsible? Or were they too far removed that they didn't even realize the code they had created would lead to such a decision being made? So even though the addition of the self-driving car may relieve the driver of having to answer the ethical debate of who to kill, the debate itself is still unanswered and has instead been even further complicated with the additional problem of assigning moral responsibility.

5.2 BRAIN IN A VAT

The Brain in a Vat is a famous thought experiment from the field of Philosophy of Mind. Imagine a brain inside a sealed container, or vat. This brain is capable of perceiving everything inside the container but it receives no sensory input from the outside. Now, suppose the brain has always been inside the container. In other words, it has grown up in there and hence has no knowledge of what anything is like beyond the container, if there is anything at all. Imagine that inside in the container is a perfect simulation of the outside world. Were this the case, would the brain believe that it is existing in the world? Many philosophers have answered yes, since it would perceive itself as existing in the world, not as existing in a container. But since we know that the brain does not have perceptual access to the actual world - since it can only perceive a simulation - we can ask whether or not it has a conscious experience of anything. Simulations are made of bits of information which compose images. Is this information capable of producing in the brain a unique experience that can in turn produce emotions? This might seem like a very strange science fiction concept, but there are currently companies working to create technology that would allow us to upload our brains into the cloud. These leads to questions such as if you were to upload your brain, would the brain in the cloud be you, a copy of you, or something we would be inclined to define as artificial intelligence? Furthermore, what would the moral status of this brain be, or in other words, how would we be able to treat this brain? If anyone has seen the episode of White Christmas from the Black Mirror series, then you have already seen an example of how this brain could be horrendously abused if its moral status is not established.

6 DO COMPUTERS THINK? THE PROBLEM OF CONSCIOUSNESS

6.1 TURING TEST

In reaction to a longstanding debate over whether or not machines are conscious in the way humans are, in the 1950's Alan Turing wrote a paper entitled Computing Machinery Intelligence. In this revolutionary paper Turing gave a conceptual account of the possibility for computers to become intelligent by way of testing the intelligence of the machine. This test has come to be known as the Turing Test, and although there are a number of versions of it, the principles underlying can be simplified. The main test involves two agents (a) and (b), and a human judge. Agent (a) is a computer and agent (b) is a human being. The identities of the agents are concealed from the judge, and the agents are tasked with trying to fool the human judge over x amount of time. Turing argued that this would be a good conversation stopper, since it would replace the standard philosophical wrangle over whether computers could be conscious with a potentially more interesting question: can computers become intelligent? This inspired the AI technological revolution. Moreover, as time passed and our ML technology advanced into the 1980's, some engineers began to hold the view that machines can not only think and be intelligent, but that actually they can one day become conscious.

6.2 CHINESE ROOM

John Searle, on the other hand, would argue that Alan Turing is simplifying intelligence too far, as seen through Searle's thought experiment of the Chinese Room. The thought experiment begins with a person inside a room with a manual for translating Chinese characters. The person themself has no understanding of Chinese, but by using the manual can take a question written in Chinese characters, match the characters with another set of characters as shown in the manual, and so give an answer to the question again in Chinese

characters. Suppose now that the manual is so intricate that the person can output answers so well that a native Chinese speaker witnessing the process from outside of the room would assume that another native Chinese speaker was inside answering the questions. It would appear from outside of the room that whoever was inside had an understanding of Chinese, even though the person inside may have no idea what the Chinese characters actually mean. This all relates to AI as we can think of the Chinese Room as a classical computer model, since it inputs a set of symbols, processes it into a new translation using formal rules, and then outputs a derived set of symbols. One might even call the manual given for the character manipulation the program. Although it seems as though the room is functioning as an intelligent mind, much like AI, Searle stresses that the person inside the room has no real understanding of the meaning of the words he is translating. Hence, there is a longstanding debate over whether or not intelligence, and thereby thinking, requires understanding, and if AI systems are capable of not only intelligence but understanding as well.

6.3 SINGULARITY

The question of whether computers can think becomes even more interesting when we start considering the possibility of a singularity. Different people have described the singularity in different ways but in general the idea is that of a super-intelligent AI. David Chalmers' argument for the likelihood of a singularity can be summarized as follows: 1) absent defeaters, humans will succeed in creating an AI that is as intelligent as we are, which is to say that they would be equally competent to humans in doing any tasks that humans typically do, 2) since this AI would most likely have an extendible method, i.e. a way of improving its capabilities, it could build a new and slightly better AI, called AI+, 3) this AI+ is in turn better also at creating better AIs, which triggers a recursive process of creating more and more AIs each

better than its predecessor versions to reach a superintelligent AI++ before long. This recursive process is also called an intelligence explosion. A lot of ethical questions arise from thinking about AI++, is this super-intelligent conscious? Would turning it off be murder? In turn, would not turning it off endanger the existence of the entire human race?

7 FURTHER LEARNING ABOUT ETHICS AND AI

If you want to dive deeper into the topics related to Ethics of AI, or just ethics and philosophy in general, head over to the Beneficial AI Society website to check out our list of resources from books to blogs to TED talks and everything in between. Additionally, if you are looking for further information on the technical side of things, check out our ML Handbook for a comprehensive overview of the popular approach to AI known as Machine Learning.

QUESTIONS?

contact: oliviagambelin@gmail.com

Beneficial Artificial Intelligence Society
5/2 Bristo Pl
Edinburgh EH8 9AL
https://bais.eusa.ed.ac.uk