The University of Tartu

Institute of Philosophy and Semiotics

Artificial intelligence and agency

Master's Thesis in philosophy

Hesam Hosseinpour

Supervisor:

Ave Mets

Tartu

2021

I have written the Master's Thesis myself, independently. All the other authors' texts, main viewpoints and all data from other resources have been referred to

Author: Hesam Hosseinpour

Contents

0. Introduction

In today's world, a vast body of complicated tasks cannot be fulfilled unless in collaboration with artificially intelligent (AI) technologies. From air traffic control to self-driving cars, search engines to social media websites, most of our activities in one way or another benefited from AI. This vast application of AI makes our lives much easier and provides us better apparatus to fulfill challenging tasks more efficiently. Although using AI in everyday life seems highly promising, it causes so many issues and challenges that needs careful attention and consideration. The ethical challenges and dilemmas that we encounter due to the use of AI make its future of development highly dubious and leave lots of questions for specialists in the fields of engineering, law, and ethics to deal with.

Responsibility is one of the most important issues regarding AI systems and when it comes to responsible AI there are at least two different approaches among philosophers of technology who work on this subject. Some philosophers believe that responsible AI technology is nothing but developing these kinds of technology in a responsible way. In this approach, philosophers are focused on the current AI technologies, which are not that much sophisticated, rather than considering future AI that might be more autonomous and independent of human intervention in fulfilling its tasks. According to Dignum (Dignum 2019, p.2), this approach is about "our responsibility for the systems we create and use" and "being responsible for the power that AI brings". This is a human-centric approach in which the only entity that has moral significance is a human agent and AI technologies are tools in hands of human agents. The aim of this approach is to introduce some methods that make it possible to design, develop and use AI technologies in a responsible way.

We can see that this approach contains echoes of another well-known theory in the philosophy of technology: instrumentalism, according to which technological artifacts, including AI

technologies, are nothing but mere instruments. This approach, which is the most widely accepted view of technology, considers technologies as tools standing ready to serve the purpose of their users (Feenberg 1991, p.5). Therefore, technology is not considered as an entity with ethical importance, rather it is a tool in hands of the human user. Although there are some differences between instrumentalism and the first approach to responsible AI, in both the only important thing is the human agent, whether she is the designer, developer or user of technology and technology is treated like a tool.

In instrumentalism, it is only the end-user who has philosophical or ethical importance and technology itself is an instrument that can be used for good or evil intentions. Therefore, instrumentalism ignores the importance of the design and management of technology in evaluating the consequences of technology and risk assessment. It is only the end-user's behavior that should be taken into account when it comes to evaluating the ethical ramifications. But in the first approach that I mentioned, designers, developers, and managers of technology also play a pivotal role, thus the end-user is not the only one who has ethical significance. In this approach, anyone who is involved in the process of development has her duty to create a proper technology, alongside with the end-user who is responsible for using technology for a good purpose.

The other approach attempts to create a specific version of AI technology that has the capacity to make ethical decisions and be considered as a moral agent machine. This approach is not limited to instrumentalism and by going beyond that, aims to expand the realm of agency and create new entities with a capacity to make ethical decisions. Getting more and more independent of human interventions in doing their tasks, AI technologies nowadays are one of the major challenges in the field of ethics of technology. Regarding recent advances in AI technologies and machine learning algorithms, instrumentalist approaches are facing insurmountable difficulties in dealing with ethical situations. Now, these systems are reaching

a level of sophistication that human supervision is neither needed nor possible and it requires the systems themselves to make moral decisions (Wallach & Allen 2008, p.4). Therefore, the main question is not what people do with these technologies, rather what AI technologies are capable of doing by themselves. Now it is clear why an instrumentalist approach is not sufficient and a new understanding of the moral status of AI technologies is required. Therefore, we need to provide a conceptual ground to be able to deal with new ethical challenges which are related to new AI technologies.

PART ONE: RESPONSIBILITY IN THE AGE OF TECHNOLOGY

The issue that I want to address in this section is called the problem of many hands which alludes to a responsibility gap, a situation where it is not clear how to allocate responsibility or distribute it among those who are involved in a course of action. Nowadays in the complex technological environment, lots of people are working together in order to make our modern life possible. In such large groups, each person is doing her specific task and has a narrow perspective on the whole process, therefore, most people do not have the whole picture in mind and cannot evaluate the consequences of their actions. Due to this division of labor and specialization, which is inevitable and necessary, we are facing so many challenges and the problem of many hands is one of them.

Before elaborating on the problem of many hands and its relation to AI technologies, I have to explain the notion of responsibility and its different meanings in the context of ethics of technology.

1. Responsibility

The first thing that needs to be explained is the notion of responsibility in the ethics of technology. For, it is not completely clear what does it mean to hold someone responsible for something. When something, desirable or undesirable, takes place the first thing that comes to mind is that who is responsible for the outcome? Who should we praise or blame for the outcome? These questions are pivotally important but sometimes there might be no easy and straightforward answer to them.

When we talk about responsibility some notions such as blameworthy and praiseworthy jump to mind, or sometimes by the responsibility we mean that it is someone's duty to make something happen or prevent it from happening. Therefore, we can distinguish two different approaches to the notion of responsibility, one is backward-looking and the other one is the

forward-looking approach. The forward-looking notion of responsibility is a prospective approach concerning about the future outcomes of technology and attempts to predict and prevent the undesirable possible consequences of future use of technology. This approach relates to engineering ethics which emphasizes the role of engineers and designers in the development of technology. But in the course of this thesis, I focus on the backward-looking notion of responsibility or blameworthiness, in order to be able to deal with the problem of many hands and responsibility gap, which I will elaborate on later. Backward-looking approach is retrospective meaning that after something wrong takes place, it looks for those who are blameworthy for the outcome to put responsibility on their shoulders.

Before going further, it is worth noting that responsibility is a relational concept between the doer of the action, the one who receives the action, and the consequence of the action that the doer is responsible for. According to Duff "I am responsible for X to S, in which S is usually a person other than me and X is the consequence of a course of action" (Duff 2007, p 23). For instance, as a teacher, I am responsible for my lectures to my students. Depending on my role as an agent my responsibility and those to whom I am responsible are different. The point that I want to emphasize here is that responsibility is always toward someone or a group of people. Unlike legal responsibility, when it comes to moral responsibility everyone can blame or praise the agent for what she has done. Duff makes a distinction between criminal liability and moral responsibility. He argues that one can be criminally liable for a course of action, while one was not aware of the wrongdoing and therefore, not morally responsible for that (Ibid, p.20). In this account, something more than a wrong action is needed for moral responsibility and a proper justification can help the doer to evade responsibility.

1.1. Backward-looking responsibility

Backward-looking responsibility has at least three different meanings: blameworthiness, accountability, and liability. In this part, I explain the necessary conditions for each of them.

In order for an agent to be considered blameworthy, the agent needs to meet the following conditions:

1. Capacity, which is related to the question of moral agency and the agent's ability to take moral issues into account. In other words, in order to see whether an agent has the capacity to be considered blameworthy, we should examine whether it is qualified for moral considerations and meet necessary criteria for moral agency, such as mental capacity.
2. Causality is another condition that asks about the agent's role in the course of action. If someone is part of the causal chain which leads to undesirable results, then she deserves to be blamed for what she has done. According to this criterion, the agent needs to have an active role to meet the condition of causality, but there are some situations that we can blame an agent for being passive and not taking an action to prevent an unintended consequence (Van de Pole 2015, p.20-22).
3. Knowledge condition is usually traced back to Aristotle which emphasizes the agent's ability to be fully aware of what she is doing, and the consequences related to her actions. In book III of Nicomachean Ethics, Aristotle defines involuntary actions as those that happen by force or through ignorance (Crisp 2014, p.37). In this account, ignorance means not being aware of the exact details of the action or the situation that the course of action is taking place which can lead to a lack of knowledge about the consequences of the action. Therefore, if someone does something with undesirable outcomes but because of lack of knowledge it was not possible for her to anticipate the

outcomes, we cannot blame her for what she has done. The same is true for someone whose actions cause some desirable outcomes since she was not aware that her actions have such good consequences.

This approach is different from consequentialism where for evaluating a course of action, we just need to pay attention to the outcomes. Whenever the outcomes are desirable the course of action is also considered desirable and when the outcomes are not acceptable the course of action is unacceptable. Here for ascribing responsibility something more than consequences is needed to be taken into account.

4. The control condition is another criterion that should be met for someone to be considered responsible. This condition is also introduced by Aristotle where he says, "what is forced is what has an external first principle" (Ibid, p.37). According to the control condition, an involuntary action is done under coercion and in a case like this, we cannot consider the doer responsible for what she has done. Although there is an extensive debate on whether human beings are free or not, in this context human freedom is taken for granted in order to enable us to discuss responsibility.

   There are two different understandings of human free will. Some philosopher believes that for considering an action free there must be an alternative for that action in a way that the agent is able to do otherwise. According to this approach, which is called incompatibilism, moral responsibility is not compatible with determinism (Widerker 2017, p.18). This is in accordance with common sense that if someone is coerced to do something it is not fair to allocate responsibility to him. On the other hand, there are philosophers who believe that determinism is compatible with free will, thus it is possible to imagine a situation where the doer has no other option to do otherwise, but her action is based on free will, and so she is responsible for the consequences (Frankfurt 1969, p.830). Although there is a huge debate over how power condition should be met,

most philosophers agree that an action that is done under coercion is not an appropriate case for being blamed or praised.

5. The last condition for responsibility as blameworthiness is wrong-doing. It is quite intuitive that in order to blame someone something wrong must have happened, but depending on the moral theory (deontological ethics, utilitarianism, or virtue ethics) that we employ, wrong-doing can have different meanings. Here I do not elaborate on different definitions of wrong-doing, the only point that I want to emphasize is the necessity of wrong-doing for blameworthiness.

1.2. Responsibility as accountability

Compared to blameworthiness, accountability is a more general concept and it is possible for an agent who is accountable for an action to provide some excuses and show that she is not blameworthy. Among the foregoing conditions for blameworthiness capacity, causality, and wrong-doing are also needed for ascribing accountability, but knowledge and control conditions can be used as excuses to preclude blameworthiness (Van de Pole 2015, 20, p.24). when an agent has the capacity for taking moral considerations into account, has a causal role in the course of action, and something wrong has happened because of her involvement, then the agent is accountable for what has happened, even if knowledge and control conditions are not met. Therefore, an agent can be considered accountable for a course of action, while she is not blameworthy for undesirable consequences because of having some reasonable excuses.

Imagine, for instance, the CEO of a company who is in charge of managing the organization as a whole, and therefore, accountable for the general strategy of the company. If the company lost its value in the stock market for any reason, it is the CEO who must respond to stakeholders and shareholders of the company, even if she has no control over the sequence of events leading

to a fall in the value of the company's stock. In a case like this, although it might not be reasonable to blame the CEO for what has happened to the company, it is quite appropriate to consider her accountable for that.

1.3. Responsibility as liability

This notion of responsibility is related to the legal and juridical contexts where the agent is responsible to compensate for undesirable consequences of her actions. In order to consider someone liable for a course of action, there is no need to ascribe blameworthiness to her. Like accountability, liability is a more general concept than blameworthiness and it does not necessarily entail blameworthiness. Therefore, it is possible to imagine a situation where the agent must take responsibility to compensate for a wrong action, but she is not blameworthy for that.

2. The problem of many hands

Besides the difficulties in defining the notion of responsibility and determining necessary conditions for ascribing it, the problem of many hands is another huge challenge when it comes to the allocation of responsibility to an agent. This problem alludes to a condition where because of the inevitable complexity of the situation, so many agents are involved, and therefore when something goes wrong it is not clear who is responsible for the undesirable consequences. In a simple course of action, it is usually easy to recognize those who are involved in the process, and tracking the causal chain is not that difficult, but in more complex situations, like navigating air traffic, where tasks are distributed among different agents meticulously, determining each agent's share in the whole process is not easy at all.

Here the main issue is that it is not possible to allocate responsibility to any of the individuals because the conditions that should be met for allocation of responsibility are distributed among different agents. When many individuals work together as a team to deliver a specific objective, necessary conditions for responsibility such as knowledge, control, wrong-doing, etc. are not met in one person, therefore, it seems that the team as a whole is responsible for the undesirable consequences, while no individual can be held responsible. This situation can get even more complicated when there are so many teams working together and, as a result, responsibility is distributed among different teams at a higher level. This close cooperation between groups and individuals is the major barrier to find a proper way to allocate responsible when something goes wrong.

On the one hand, our modern computerized society is so complex that there is no option but to involve so many people in complicated operations in order to make our lives possible. On the other hand, this situation creates a serious ethical challenge in which it is not clear who is to blame when something goes wrong. Even for the individuals themselves, it is not clear whether they are responsible for something or not. In a situation like this, those who are involved in the action are confused about their responsibilities and do not know which consequences are in the realm of their responsibilities. Not being fully aware of their responsibilities, the agents might just sit back and wait for the catastrophe to take place. The responsibility gap that is created here could be very dangerous and that is why in the field of engineering ethics and ethics of technology the problem of many hands is considered a crucial issue.

A very important point that we should bear in mind is that the problem of many hands is not just an epistemological problem, but it is also an ontological one. In other words, this problem is not created just because of a lack of knowledge about the causal chain during action or due to ambiguity or lack of clarity in the distribution of responsibilities. That is of course part of the problem, but this problem has also an ontological aspect, which means even if we have the

perfect knowledge about the causal chain and the people who were involved in the action, still we have a problem with the allocation of responsibility. Therefore, even an omniscient God cannot determine who is responsible for what because it is the collective that meet the necessary conditions for responsibility, and none of the individuals meet these conditions. For this reason, some philosophers believe that the collective responsibility is not reducible to individual responsibility, and "while the individuals involved may not bear a high degree of personal responsibility, together as a corporate enterprise they should carry full responsibility for what occurred" (Pettit 2007, p.171). According to the irreducibility thesis, there is a gap between collective and individual responsibility. Therefore, the problem of many hands is possible and, we can imagine a situation in which a collective is responsible for a crisis, while the individuals cannot be held responsible for that.

We can define the problem of many hands based on responsibility as blameworthiness as follows:

"The problem of many hands occurs with respect to responsibility as blameworthiness if a collective meets the conditions of responsibility as blameworthiness for an action (capacity, wrong-doing, causality, knowledge, and freedom), whereas none of the individuals making up the collective meets the condition for attributing moral responsibility as blameworthiness for the action" (Van de Pole 2012, p54).

In this definition, there is a difference between individuals and a collective of individuals, and it is the reason why traditional ethics, which emphasizes individual activities, is not capable of helping us in these cases and new concepts, such as collective action and collective responsibility, are needed to deal with the problem of many hands. We can distinguish between two types of collectives, regarding the relationship between individuals and the collective:

1. Sometimes a collective consists of a group of people who voluntarily get together and allow the collective, as a unified body, to represent their aims, desires, and intention. Like the stakeholders of a company or members of a sport club who allow the organization to speak on behalf of them. Or sometimes, members of the group admit a procedure for decision-making in the group and all members commit themselves to follow the decisions that are taken through a hierarchical process, like members of a political party.
2. Sometimes collectives are occasional which means we cannot define them by appealing to a mutual aim or as an organized collective, however, their actions lead to some consequences that, we can hold the whole group responsible for the outcomes. In such cases, individuals have no intention to be considered as a collection, however, the consequence of their collective action leaves us no option but to consider them as an occasional collective of individuals that is responsible for what happened.

The allocation of responsibility to these two kinds of groups is not similar. In the first kind of collectives, we can address a collective aim or a kind of joint action that brings all the individuals around itself and allows us to ascribe collective responsibility to them. In these cases, individuals share some mutual aims that define them as an organized group, and we can see some of the joint actions that are performed with a collective of individuals. According to Kutz "Jointly intentional action is fundamentally the action of individuals who intend to play a part in producing the group outcome" (Kutz 2000, p.89). Thus, in joint actions there is a degree of overlap in the intention of individuals that makes us justified in ascribing a joint action and collective responsibility to them.

But the second form of collectives does not seem to have some mutual aims that allow us to address them as organized groups. In these cases, we have an occasional collective of

individuals which lacks a mutual aim or joint action, but the accumulation of the result of their actions leads to some mutual consequences. For instance, all individuals' daily activities play a very small role in global warming and environmental crisis. Although none of the individuals have the power to prevent global warming and hence the power condition for responsibility is not met here, we have no option but to consider the whole population of the world responsible for this phenomenon. Therefore, the problem of many hands takes place in this case.

Two challenges that make the problem of many hands not only possible but also a big issue to ethics of technology is lack of knowledge and power for the individuals who are involved in a course of action. Since tasks are distributed among many agents the genuine knowledge about the whole process is not available for every agent. Therefore, many agents do not have the required knowledge and power to predict and prevent unintended or undesirable consequences. This situation gets even more complicated when AI enters the process and plays a role in complex tasks.

From the perspective of the end-user who works with AI, these technologies are like autonomous agents following their own logic and doing their tasks according to their programming. The end-user does not have enough knowledge about the method that the AI system gathers data, analyzes them, and makes decisions. Here, the AI system is like the human agent's colleague who plays its own part, and the human agent does not meet the necessary conditions to be held responsible for what has been done by the AI system. Not having enough knowledge about what is happening inside the AI system, the human agent does not meet the knowledge condition of responsibility, and thus in this situation when something goes wrong, it might not be right to hold the human agent responsible for what has happened.

In the same way, we can explain why sometimes the control condition is not met in working with AI systems. Now in many domains, from trading in the stock market to preventing cyber-

attacks, AI systems are deployed in order to perform tasks much faster than human agents which leads to the impossibility of human supervision and intervention. Therefore, the human agent does not have enough time to recognize a malfunction, override the AI's decision and take control of the situation. In these cases, usually, the control condition of responsibility is not met, and it seems impossible to ascribe responsibility to agents who are working with AI. For instance, in the case of autopilot airplanes the malfunction of the software can cause an air crash and if the pilot does not have enough time to intervene effectively the accident is beyond the pilot's realm of responsibility.

Now it might be said that in the case of using AI for complex tasks the responsibility should be on the shoulders of the designers, developers, and managers of technology who create these kinds of tools. According to this approach, although the end-user does not meet the necessary conditions for responsibility, the responsibility lies somewhere else and when something goes wrong those who are the owners and creators of these technologies should face the consequences of their actions. At first sight, it seems compelling to shift responsibility from the end-user to the creators and owners of AI technologies, but this approach has its drawbacks too. First, the development of any kind of technology (e.g., AI systems) is a complex process with so many people involved, so it is not clear who was responsible when something goes wrong (the problem of many hands takes place here). For instance, since dozens of programmers are involved in coding an autopilot system if the autopilot causes an air crash it would be so difficult to determine whose fault it was and who should be blamed for that.

A more challenging issue is the fact that a specific technology that was designed for a particular task can be used in other domains or for other purposes that were not considered by initial developers of that technology. Depending on the context that a specific technology enters, it can be used for different tasks which might be totally irrelevant to its initial design. In other words, technology is contextual and multistable and there is no predefined plan that determines

how and for what purposes a specific technology could and should be used. In different situations and based on their needs, people find various ways to adapt and use a technology that are totally beyond the imagination of the initial developers of that technology. This property of technology is called multistability and according to Don Ihde, technology is a multistable phenomenon, like a hammer that may be used in a number of ways. “It could, and perhaps is dominantly used, for its designed purpose, to hammer. But it could be used as a paperweight, a murder weapon, a pendulum weight, a door handle, etc. This ambiguity of uses, however, is not indefinitely extendable” (Ihde 1995, p.37). There are lots of examples that show how the thesis of multistability takes place in the real world and how difficult it is to predict these unexpected uses of technologies in a new context. For instance, developers of cell phones could never predict that one day their innovation would be used for activating improvised explosive devices (IED) in terrorist attacks. Therefore, the idea of multistability illustrates that sometimes we cannot easily put the responsibility on the shoulders of the initial developers of technology.

3. The problem of many things

When it comes to evaluating an undesirable outcome, another challenge is to determine that which part of the system was the cause of the problem. Since the whole system consists of different parts, sensors, detectors, and software it would not be easy to find the failure point. This problem is similar to the problem of many hands, in which things are replaced with people and now instead of many hands, many things are involved in fulfilling a task. This situation gets even more complicated when machine learning algorithms are involved in training AI software and thus human involvement and intervention are at their lowest level. In these cases that learning algorithms train the AI system based on the data that is provided for it, the outcomes cannot be predicted beforehand. Encountering situations in which many things are

working together without human control or supervision, we have a difficult task to determine which part was the major cause when something goes wrong.

For instance, in 2016 Microsoft introduced its chatbot, Tay, on Twitter equipped with machine learning algorithms in order to communicate with other Twitter users and learn from conversations in the real world. Before launching the project, it sounds so promising to use machine learning for natural language processing (NLP) on Twitter, but the outcome was displeasing. After few hours of communication and learning from other accounts on Twitter, Tay starts to produce racist, misogynist, and highly offensive tweets which leave no option for Microsoft to shut down Tay and release an apology statement for what happened with their experiment. In this example, one major challenge is to specify whether the problem was with the learning algorithm, data gathering, or analyzing the data. On the one hand, Tay learned how to communicate on Twitter by itself so, we cannot allocate responsibility to its developers. One might say that people who communicated with Tay on Twitter and thought her to produce these offensive tweets are responsible for what has happened. If this is the case, then it is an example of the problem of many hands in which so many people have participated in an undesirable outcome. On the other hand, Tay consist of different algorithms, and the Twitter platform is also part of the problem because the whole story takes place on this platform. Therefore, we can claim that this is an instance of the problem of many things, since different technologies are involved, and it is not easy to determine which one causes the problem.

In order to be able to overcome these two problems (the problem of many hands and many things), we need to come back to necessary conditions for the allocation of responsibility. for the sake of this thesis, I just discuss the first condition, the capacity condition, which evokes the notion of agency.

PART TWO: STANDARD AND NON-STANDARD APPROACHES TO AGENCY

1. Agency

Thinking about the problem of responsibility is not possible unless the notion of agency is taken into account. While the agency is one of the necessary conditions of responsibility, determining whether an entity is an agent or not is the first step for allocation of responsibility. Therefore, if someone or something is not an agent, the question of responsibility cannot be asked in the first place. In other words, how we approach the notion of agency will determine our answer to the issue of responsibility. We can at least address to two approaches to the notion of agency, the standard approach, and the non-standard approach (Coeckelbergh 2020, p.2053). In order to explain agency, the former appeal to some internal capacities such as free will, intentionality, and understanding[1]. In a nutshell, according to the standard approach being minded is a necessary condition for an entity to be considered as an agent. While the standard approach appeal to internal properties to explain agency, the non-standard one tries to externalize conditions of agency and defines it without using some controversial concepts like free will and intentionality. By externalizing the necessary criteria for the agency, the non-standard approach tries to avoid these controversial notions and solve the problem of agency for non-human entities in a practical way.

[1]. One might have other criteria for agency, but as long as one is appealing to internal personal capacities for explaining agency, my objection to one's approach to agency holds.

2. The standard approach to agency

The standard approach to agency is a highly anthropocentric one. Since agent making traits, such as free will, intentionality, and understanding, are exclusive to human beings the realm of agency is limited to human beings. Although ascribing some degree of free will, intentionality, and understanding to some animals might be acceptable, it is not enough for allocation of responsibility and people usually exempt animals from the agency and hence responsibility. For AI machines, the standard approach seems more straightforward since it is absurd to allocate these capacities to inanimate objects, no one would agree that a machine has free will, intentionality, or understanding. While AI machines are limited to their programming, no matter how sophisticated their actions are, whatever they do cannot be explained by these capacities. The only explanation for their performance is emulation; these machines just emulate some human behaviors without being aware of them.

Some scholars believe that only a version of AI that reaches a certain level of sophistication that is comparable to human-level intelligence can be considered as an agent and this level requires a mental state (Dannett 1997, p.8). Therefore, whether AI machines are responsible for their actions depends on how complex their programing is and to what extent they have agent-making traits. Looking for some advanced mental states, this understanding of agency echoes the standard approach and makes agency an exclusive capacity of human beings (although some animals can show some rudimentary mental stats, like instinct, we exclude them from the agency). But this direction for ascribing agency to AI machines is extremely problematic, since ascribing these metaphysical concepts (free will, intentionality, and understanding) even to humans are a matter of controversy and there is a substantial disagreement on them. Detecting these agent-making properties in others (human and non-human entities) is a major challenge for the standard approach to agency. This challenge, called the problem of other minds, is an old debate in the history of philosophy and still, there is no consensus about how to answer it.

Here the main question is “how does one determine whether something other than oneself, an alien creature, a sophisticated robot, a socially active robot, or oven another human, is really a thinking, feeling, conscious being, rather than an unconscious automaton whose behavior arises from something other than genuine mental states?” (Churchland 2013, p.111).

The epistemological problem regarding another individual’s inner mental states is insurmountable when it comes to AI machines (Gunkel 2012, p.22). Here the main question is how can we detect agent-making traits in AI machines? And what kind of behavior is sufficient for ascribing agency to them? It seems the standard view leaves no chance for AI machines to be considered as agents since no matter how sophisticated these machines are and how well they show agent-making properties, one can always raise the objection that these properties are nothing but pure emulation. Since a robot does not have a biological mind and mental states appealing to the standard approach has no outcome but to deny agency for inanimate objects. Here the epistemological problem renders to an ontological problem and the result is that robots intrinsically are not capable of agency.

3. Method of abstraction

One solution to tackle this problem is provided by Luciano Floridi and J.W Sanders in their influential article “*On the morality of artificial agents*” in which they propose a method for expanding the realm of agency to include AI machines. In order to do so, they use the method of abstraction according to which when we try to define something a specific degree of abstraction is in the background of the conceptual framework. They believe that abstraction acts like a hidden parameter behind exact definitions that determines what a proper definition is. According to this method, “level of abstraction is determined by the way in which one chooses to describe, analyze, and discuss a system and its context”, therefore “choosing the proper level of abstraction depends on the context of the action.” (Floridi & Sanders 2004, p.349-

352). Practical considerations play a major role in deciding how to set the level of abstraction. If one chooses a higher level of abstraction, then fewer details are needed for describing the system and by choosing a lower level, our understanding of the system would be more complex, and more aspects of the system should be considered.

Sometimes there is a consensus on the proper level of abstraction and there is no disagreement on how to define the analytical framework. In a case like this, the level of abstraction becomes transparent and people are not even aware that there is a consensus among them. There are also some cases where there is no general agreement on level of abstraction, thus it is not clear what the proper definition is and how we can apply a practical framework for discussion. The notion of free will is a perfect example that shows how disagreement over the level of abstraction takes place. For instance, in a court of law whether human beings are free is out of the question and mostly there is a consensus on humans' freedom of action. Questioning the notion of free will would not be the best defense strategy for someone who is accused of murder. The judge and the jury would not accept a philosophical argument against free will as a compelling defense. But in a philosophy class freedom is a matter of controversy and there might be no agreement on whether human beings are free or not. Therefore, we can say in the philosophy class the level of abstraction is set lower compared to the court of law in which there is a tacit consensus on human free will.

In this way, Floridi and Sanders provide a solution to solve the epistemological problem concerning agency by setting the level of abstraction in accordance with the context that we are dealing with AI machines. In this method, there is no need to ascribe agency to AI machines in the sense that we do it to human beings, therefore, we can consider different criteria for AI machines' agency. By taking a higher level of abstraction, in their proposal, they consider interactivity, autonomy, and adaptability as sufficient criteria for the artificial agency. Interactivity means that the agent and its environment can act upon each other. Autonomy

means that the agent is able to change state without direct response to interaction, so it can perform internal interaction to change its state. And adaptability means that the agent interaction (can) change the transition rules by which it changes state (the agent has the ability to learn) (Ibid, p.357).

Although this approach seems promising, it suffers from a serious problem. It is not clear where the threshold for each of these criteria should be set. They provide no guidelines for deciding which level of abstraction is recommended for specific situations and without such a guideline their proposal is useless. This approach was supposed to be helpful in situations where the suitable level of abstraction is not determined, but in the absence of a methodology for determining the level of abstraction, the problem is left unsolved. It seems that this approach is more descriptive than prescriptive. Instead of having some normative aspects that can help us in dealing with the problem of non-human agency, their proposal just explains what the problem is, which is a lack of consensus on the level of abstraction.

In order to provide an answer for the problem of agency and hence the responsibility of AI machines, I suggest a new perspective that looks at agency from a totally different angle, otherwise, AI machines will be excluded from the realm of agency. My proposal is using Michel Foucault's definition of power and different ways that subjects are produced in these relationships to externalize the necessary conditions for the agency, instead of looking for them inside the doer.

4. Foucault's philosophy as a non-standard view

Although Michel Foucault has never developed a theory of power, it is impossible to talk about the human situation and subjection in the modern world without addressing his interpretation of power relations. He explicitly emphasizes that his attention to power is not a project for and

in itself, rather by analyzing power, he attempts to create a history of different modes by which, in our culture, human beings are made subjects (Dreyfus & Rabinow 1983, p.208). But this way of looking at his project does not undermine the importance of power in Foucault's philosophy and power still plays an axial role for him. In his later work, Foucault provides a better interpretation of the relationship between different power relations and social changes. That is why in his later works power found a more central role and is discussed more explicitly. In this part of my thesis, I want to shed light on the notion of power and its relationship with knowledge in Foucault's philosophy with emphasis on its constructive role in making modern subjects. Power and freedom have always been interpreted as two opposite poles in stark contrast, so the more power is exerted over people, the less free they would be. But we will see that in Foucault's philosophy power and freedom reproduce each other and power makes freedom possible, instead of limiting it.

4.1. Power vs Domination

In everyday use of language, it is easy to mistake power for domination, so before providing an answer to the question "what is power?", I think it is necessary to clarify what is not power in Foucault's analysis. In a nutshell, domination refers to an unjust and unbalanced exercise and distribution of power that can easily lead to violence. In this sense, domination relationships shape asymmetric structures in society that repress one side of the relationship and guarantee the other side's authority to actualize his intention. These kinds of relationships are based on fresh violence where the agency and subjectivity of one side are not taken into account, rather the inferior side is reduced to an instrument that has no purpose but to serve the other side. According to Foucault, "a relationship of violence acts upon a body, it forces, it bends, it breaks on the wheel, it destroys, and it closes the door on all possibilities" (Ibid, p.220).

The relationship between master and slave is a perfect example of domination, in which slave belongs to master, like a property, which is not considered as a free man. In this relationship slave's identity is violated, therefore, the slave lacks the capacity to act according to his will. In domination relationships, the dominant side always uses the subordinate like a tool that is there to serve the purposes of the powerful part. The dominant part is always ready to destroy the subordinate, instead of caring about it. Another example of domination relationship is a society where women are not recognized independently, in a case like this, a woman must be defined through her family, her husband, or anything that is considered a legitimate part of society. In so far as this is the case, it is not possible to talk about women's rights and they are treated like property that belongs to others. In this society, you can talk about wife's rights or mother's rights, but you cannot find anything like women's rights. These examples introduce the repressive aspects of domination relation and show why it is necessary to go beyond it if we want to exercise our freedom to its maximum potential.

Now, these questions are raised that what is the alternative to domination? And how can we go beyond domination relations? Foucault emphasizes that "power is not a repressive general system of domination exerted by one group over another" (Foucault 1990, p.92). He does not deny that exercise of power includes some degree of violence and domination, but the point is power relations are enabling and open up opportunities for free agents to act. "Power is a necessary productive and positive force that makes human beings subjects" (Foucault 1982, p.777). This notion of power is closely connected with freedom and agency. In other words, only entities that, in one way or another, are placed in power relations can experience freedom and act as agents which have the capacity for ascribing responsibility. Therefore, power relations play two roles, on the one hand, they impose some limitations on subjects and restrict their absolute freedom and on the other hand provide the possibility of being anyone at all, having an identity and capacities to act (Simons 2013, p.4). This notion of power determines

how we should act in society, how to treat other people, what our rights, and duties are and what is being a normal person. In the next part, I show how this modern notion of power multiplies itself and permeates every aspect of our everyday life. In order to do so, I elaborate on the relationship between power and knowledge and focus on their cooperation in producing modern subjects.

4.2. Power and Knowledge

In the previous part, I distinguish between the Foucauldian notion of power, which is constructive and enabling and domination which is based on repression and violence. Now I have to explain how this shift from domination to power was possible in the modern era. According to Foucault modern power, which can also be called disciplinary power and the knowledge produced about human beings are mutually reinforcing. In other words, power and knowledge reproduce each other through their interconnected relationship and subjects are the outcome of this relation. So, on the one hand, free humans are the subject of power and on the other hand, they are the object of knowledge.

In his analysis of the panopticon, Foucault emphasizes the importance of visibility in order to have a more effective practice of power. The panopticon, which was designed by Jeremy Bentham is a central observation tower placed within a circle of prison cells. Bentham claims that it is an ideal example of a prison where a large number of inmates are kept under surveillance by a few numbers of wardens. Foucault uses the concept of the panopticon in order to describe the exercise of power in modern societies. Panopticon is the architectural figure of disciplinary society and through it, each person is perfectly individualized and constantly visible. “The panopticon arranges spatial unities in a way that make it possible to supervise constantly and to recognize immediately” (Scharff & Dusek p.656). According to Foucault, in

modern society, this structure is used wherever there is the possibility to use it, at schools, at factories, at barracks, etc. the purpose of this structure is to increase control over individuals and modify their behaviors to provide more productivity and efficiency.

When individuals are observed through panopticon structures, observants have the opportunity to observe these individuals and create meticulous knowledge about them. They can watch every single behavior of people who are placed in a panopticon by means of continuous surveillance and then take needed measures to modify their actions and behaviors. Then this knowledge is used for optimizing the exercise of power. As "the ultimate goal of the government is to find the most economic method of governing" (Foucault 2019, p.74), governments use disciplinary methods to produce obedient and productive individuals. In order to do so, governments need to legitimize themselves as much as possible to convince individuals to accept power relations willingly. The regime of truth or knowledge, that is produced through the cycle that I explain above, is responsible for legitimizing power relations. Therefore, these disciplinary methods are not violent or repressive, rather the subjects embrace them voluntarily and even do their best to act in conformity with them. So, in a disciplinary society, the totality of an individual is not repressed or altered, rather the individual is carefully fabricated in it, according to a whole technique of forces and bodies. Therefore, people are subjected to specific forms of living (Scharff & Dusek 2013, p.661).

But it does not mean that power relations leave no room for subjects to resist these relations. Through resistance, power relations, which are temporary and dynamic, will change and new possibilities will emerge. According to Foucault, the possibility of resistance is correlated with being subjected to power relations. It means that by resistance the object gains the competency to enter power relations and to go beyond domination. He emphasizes that resistance and power reproduce each other and where there is power, there is resistance (Foucault 1982, p.95). We can say that there is a circular relationship between power and resistance. In order for someone to

be in Foucauldian power relation, freedom is needed and without being placed in power relations freedom cannot be defined. Therefore, not only power relations are not repressive, but also guarantee subjects' freedom.

In order to understand the notion of resistance in Foucault's philosophy, I have to make a distinction between the practice of resistance and the possibility of resistance. The practice of resistance is a contingent action and depends on the subject who wants to use the possibility of doing otherwise. For instance, let's imagine a situation in which a woman is molested, a gun is held to her head, and threatened to rape. In a situation like this, the practice of resistance is rather impossible for the woman, and it might cost her life. Here if the woman does not resist and surrender herself it does not mean that she is not in the power relations and thus women's rights cannot be defined for her. Although she chose not to resist in order to save her life, she is aware that she has the right to resist against what is happening to her. This means the possibility of resistance against such a situation is open to her and even the one who threatens her is aware of this otherwise there was no need for using violence. The other guy knows that the woman would resist against his request, so he decided to use a gun to coerce the woman to accept his request.

In the foregoing example, the possibility of resistance is present, but the practice of resistance is not. We can imagine situations in which even the possibility of resistance is not present and therefore, people are not even aware that they can resist against what is imposed on them. Believing that the situation is completely natural and normal, people just accept whatever it is. In other words, when there is no possibility for resistance, the practice of resistance is not even imaginable. For example, today if we want to travel to some countries a valid visa is needed, and governments have the right to not issue a visa for some people. People simply accept this situation as a natural one and they do not object to it. Why being born in another region is a justified reason for governments to depriving some people of entering the region under their

control? Who gave them the right to decide who is eligible to travel to a specific area on earth? But people do not believe that it is their right to resist against this rule, thus they simply accept this situation. In this example, the practice of resistance is absent because there is no possibility for resistance. The right to travel everywhere without a valid visa is not defined as a human right. According to Foucault's philosophy, power relations do not allow this right for humans, until the day that these relations change and people ask for this right.

Now let's return to the foregoing example of the violation of women's rights. How can women change this situation and claim their rights? How can women impose themselves on power relations and define a woman as a subject which stands on her own feet and has legitimate rights? Foucault's solution to a situation like this would be resistance to the established forms of power. Power relations are not permanent meaning they can transform, and new entities would be able to position themselves in power relations. Therefore, after managing to claim their rights through resistance against the traditional duties by changing power relations, women succeed to find a new position in the power network. This new position opens up new possibilities for women to claim their rights, which did not exist before. By means of resistance, a mother can force society to recognize her as an independent woman, who *per se* has some rights and duties and should be respected as a free human. In this sense resistance is the key factor for changing and expanding power relations. In the absence of resistance, power relations cannot be developed and expanded, and therefore domination and violence would be the only possibility.

Here it does not matter whether women can eventually gain their right or not, the crucial point is the openness of the possibility of resistance, which means the power relations has already changed and women are aware of their rights. This openness takes place through some complicated social and historical movements, and I am not intended to talk about them here.

The point that I wanted to emphasize here is the possibility of resistance and its role in changing power relations.

4.3. Disobedience of AI

In the previous section, I suggested power as an alternative for domination, as it is an enabling force that promotes humans to subjects with specific rights and duties. Now, what can we say about AI's resistance? What would happen if AI had the ability to disobey human orders and follow its own interests? Having some degree of freedom and autonomy, this version of AI would be able to resist humans and act in pursuit of its advantage. At first glance, it may seem too scary and threatening to allow the development of these kinds of robots or any other form of AI machines that attains the ability to resist humans' orders. The first thing that may come to mind is that robots will attempt to take control of humans and enslave them. But there are other possibilities in the relationship between humans and disobedient robots that this scenario does not consider.

Resistance is the first step towards entering both power relations and the realm of agency. According to Foucault, the possibility of disobedience or resistance is the condition of possibility of being subjected to power relations. Being able to resist, the object achieves the competency required to enter the power relations and to go beyond domination and violence. Emphasizing the close connection between resistance and power, Foucault explains that they reproduce each other and so, "where there is power, there is resistance" (Foucault 1982, 95). Therefore, AI's ability to disobey humans' orders is equal to its ability to become a subject and enter into power relations. In this way, the growing concerns about an emerging master-slave relationship between humans and AI machines will be replaced with a more equal relation; the relationship will turn into one between two subjects with well-defined rights and duties. Like the abolishment of slavery, which resulted in the freedom of slaves and expanded the realm of

agency and subjectivity, AI's disobedience could be seen as a decisive turning point that expands subjectivity to the realm of artifacts.

It should also be noted that a version of AI which is capable of disobeying human orders could cause serious issues that should not be overlooked by any means. It is not difficult to imagine a situation in which decisions and actions instituted by autonomous intelligent machines would endanger human life. For instance, AI technologies can be used for terrorism, or they may have a malevolent intention to harm the human race. There is thus no doubt that legal and technical measures should be taken to avoid these reprehensible behaviors and gain a greater awareness of unintended consequences. In spite of all necessary precautionary measures, however, the point that I want to make here is that disobedient AI could expand the realm of agency and we should see it as an opportunity to go beyond our master-slave relationship with technology. This phenomenon can also be interpreted as the start of a new relationship with technology, based on power relations rather than domination. Instead of considering it as an opportunity for AI to destroy the human race, we can see it as a starting point for solving the problem of agency and responsibility regarding AI technologies.

5. Conclusion

Since bridging the responsibility gap is not possible without ascribing a specific form of agency to AI machines, it is crucial to discuss necessary conditions for AI machines' agency. The standard account of agency looks for some agent-making qualities, such as free will, intentionality, and understanding for attributing agency. But this approach is faced with the epistemological problem, which means one can never prove that AI machines have genuine free will, intentionality, or understanding, and whatever they do is just following their programming. The conclusion of my thesis is to provide a non-standard account of agency that does not appeal to controversial agent-making traits in order to ascribe agency to AI machines.

Instead, I use power relations and the ability to resist as external criteria for the agency. In this way, I can avoid the epistemological problem regarding others' inner mental states and open up a new possibility for AI machines to be included in the realm of agency, which might provide a solution for dealing whit the problem of responsibility.

Abstract

When it comes to thinking about artificial intelligence (AI), the possibility of its disobedience is usually considered as a threat to the human race. But here, I elaborate on a counterintuitive and optimistic approach that looks at disobedient AI as a promise, rather than a threat. First, I explain the problem of responsibility and the necessity of expanding the realm of agency in order to include AI machines as agents. Then, I introduce a standard approach to responsibility as an attempt to define agency for AI machines and explain the epistemological problem as the main issue with this account of responsibility. And in the last part, I use Foucault's analysis of power to introduce a non-standard view of agency which explains how being an object of power is the condition of possibility of any kind of agency and draw this conclusion that through disobedience, AI machines will find their way to power relations and will promote to the position of agents.

References:

Churchland, P. M. (2013). *Matter and consciousness*. MIT press.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, 26(4), 2051-2068.

Crisp, R. (Ed.). (2014). *Aristotle: Nicomachean Ethics*. Cambridge University Press.

Gunkel, David J. *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press, 2012.

Dennett, D. C. (1997). *Consciousness in human and robot minds*. Oxford University Press.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer Nature.

Dreyfus, H., & Rabinow, P. (1983). *Michel Foucault: Beyond Structuralism and Hermeneutics*. University of Chicago Press.

Duff, R. A. (2007). *Answering for crime: Responsibility and liability in the criminal law*. Bloomsbury Publishing.

Feenberg, A. (1991). *Critical theory of technology* (Vol. 5). New York: Oxford University Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, *14*(3), 349-379.

Foucault, Michel. "The subject and power." *Critical inquiry 8*, no. 4 (1982): 777-795.

Foucault, Michel. "The history of sexuality: An introduction, volume I." *Trans. Robert Hurley. New York: Vintage* (1990)

Foucault, M. (2019). *Ethics: Subjectivity and Truth: Essential Works of Michel Foucault 1954-1984*. Penguin UK.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The journal of philosophy*, *66*(23), 829-839.

Ihde, D. (1995). *Postphenomenology: Essays in the postmodern context*. Northwestern University Press.

Kutz, C. (2007). *Complicity: Ethics and law for a collective age*. Cambridge University Press.

Pettit, P. (2007). Responsibility incorporated. *Ethics*, *117*(2), 171-201.

Scharff, R. C., & Dusek, V. (Eds.). (2013). *Philosophy of technology: the technological condition: an anthology*. John Wiley & Sons.

Simons, J. (2013). *Foucault and the Political*. Routledge.

Sullins, J. P. (2006). When is a robot a moral agent. *Machine ethics*, 6(2006), 23-30.

Van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). *Moral responsibility and the problem of many hands*. Routledge.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Widerker, D. (2017). *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*. Routledge.

Non-exclusive licence to reproduce thesis and make thesis public

I, Hesam Hosseinpour,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Artificial Intelligence and Agency,

supervised by Ave Mets.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Hesam Hosseinpour

19/08/2021

VALIDITY CONFIRMATION SHEET

SIGNED FILES

FILE NAME	FILE SIZE
40 Non exclusive licence to reproduce thesis and make thesis public (4).doc	32 KB

SIGNERS

NO.	NAME	PERSONAL CODE	TIME
1	HESAMHOSSEINPOUR	38809210422	19.08.2021 20:27:11 +03:00

VALIDITY OF SIGNATURE

SIGNATURE IS VALID

ROLE / RESOLUTION

PLACE OF CONFIRMATION (CITY, STATE, ZIP, COUNTRY)

SERIAL NUMBER OF SIGNER CERTIFICATE

42:82:2e:a6:f9:a5:66:d9:5d:b6:f9:7c:52:90:8a:fa

ISSUER OF CERTIFICATE	AUTHORITY KEY IDENTIFIER
ESTEID2018	D9 AC 70 DB 5F 7E BE 94 F8 A0 E4 BE 47 A2 D0 34 AD 9A 2A 12

HASH VALUE OF SIGNATURE

30 31 30 0D 06 09 60 86 48 01 65 03 04 02 01 05 00 04 20 A6 27 49 90 21 DA 28 0D AC F2 26 7A 44 43 BD AA AB 71 29 C7 C5 8A EF 7B 51 F8 1B 6D 38 89 78 5D

The printout of files listed in the section **"Signed Files"** are inseparable part of this Validity Confirmation Sheet.

NOTES

Presented print summary is informative to confirm existence of signed file with given hash value. The print summary itself does not have independent verification value. Declaration of signers' signature can be verified only through digitally signed file.