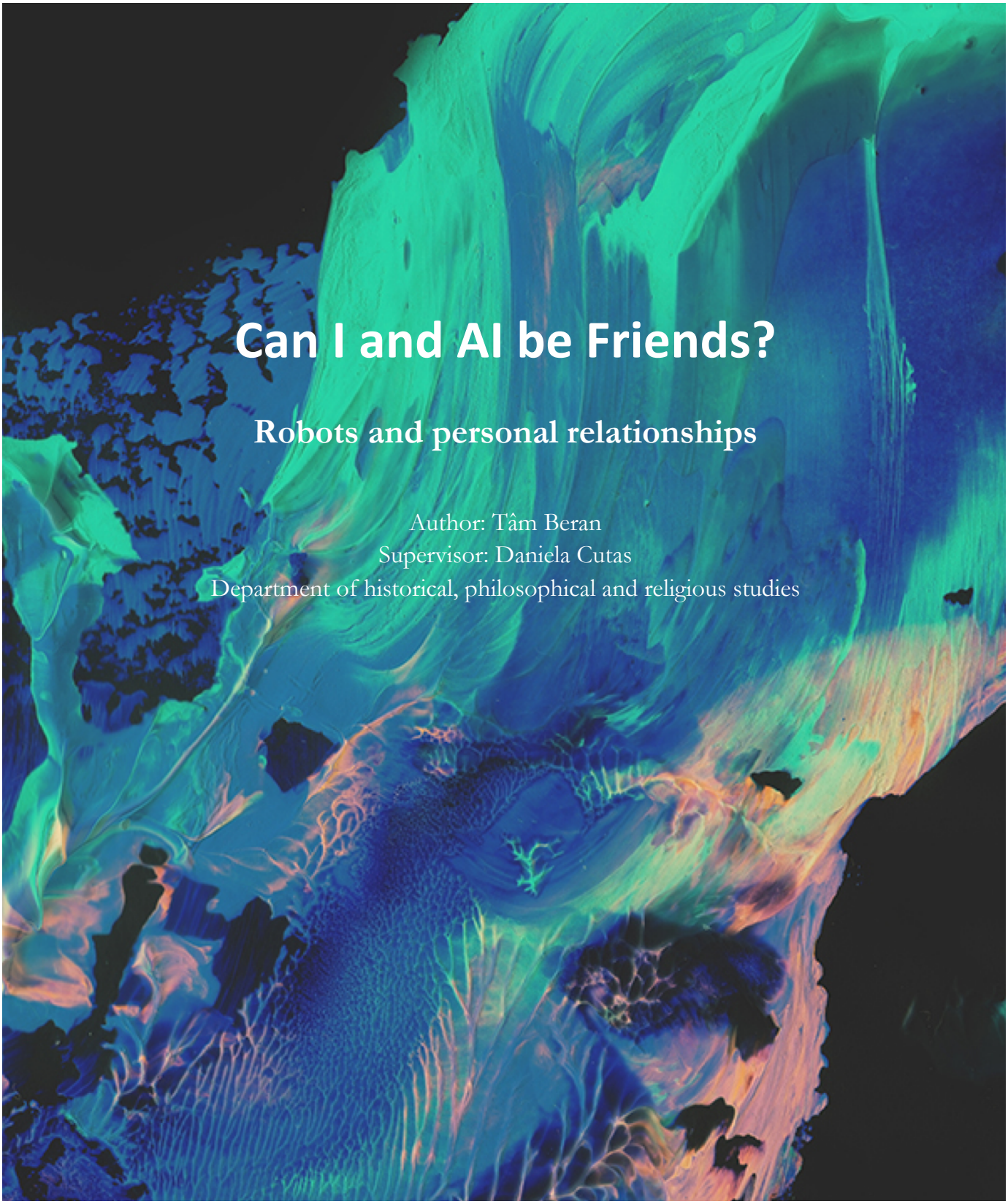
UMEÅ UNIVERSITY



Can I and AI be Friends?

Robots and personal relationships

Author: Tâm Beran
Supervisor: Daniela Cutas
Department of historical, philosophical and religious studies

Bachelor thesis, 15 hp
Philosophy C, 30 hp
Spring term 2020

You are imparting human qualities to it because it *looks* human – but I assure you: *it is not*. If it were a box on wheels I would *not* be facing this opposition.

Maddox about Data, Star Trek episode *The measure of a man*

Abstract

To exist is to relate. As human, you are relating to other beings, animate and inanimate entities, physical objects and abstract ideas. A kind of relationship that affects our life and wellbeing in a most concrete sense is that between humans. Modern technology has made it possible to create artificial intelligence (AI) that has become increasingly integrated in our everyday life. AI can be distinguished between *weak* or *strong*, that is whether the AI *appears* to have human mental capacities or *in fact* has these capacities. The aim of this thesis is to determine whether AI and humans can be friends, based on the condition of them having equal moral status, as well as the concept of friendship as defined by LaFollette. According to LaFollette, a friendship is defined as a relationship that is voluntary, reciprocal and where you relate to each other as unique individuals.

If considering life as essential for moral status, true friendship is not possible between a human and an AI, weak or strong. Other criteria for moral status are the capacity of feeling pleasure and pain, being conscious and having a mind. Although weak AI would behave as if it has human mental capacities, it cannot have the same moral status as humans, and consequently cannot be involved in a genuine friendship in this framework. On the other hand, a strong AI would have equal moral status as a human, and a relationship with such an AI would have all the essential properties required for a friendship as defined by LaFollette. However, from a subjective point of view, it is possible to create unidirectional emotions towards an AI regardless of it having a mind or not.

Table of Contents

1 Introduction

Intelligence, or some form of it, is not a capacity only humans can possess and needs not to have an origin in biological neural processes. Modern technology and the rapid development of computers have made it possible to create different forms of artificial intelligence (AI). In some contexts, it means an AI algorithm that can play chess. In other contexts, it can be a social robot that interacts with a patient.

AI becomes increasingly sophisticated and nestled in our everyday life, something which has consequences on our society in various aspects. In order to assess how AI may shape us as individuals and our societies, it is important to study the relationships that arise between humans and AI. Humans are relating to each other in different ways depending on with whom we are relating and under what circumstances. Our relationships in general and personal relationships in particular have a large impact on our lives and wellbeing.

The aim of this thesis is to determine whether AI and humans can be friends. The first step to meet this aim is to identify in what aspects AI differs from humans. The next step is to analyse moral philosophical arguments on what is required for a relationship to be defined as friendship. Finally, the question is answered by determining whether AI can meet these requirements or not, and whether relationships that may develop between humans and AI may legitimately be called friendship.

Within the scope of this bachelor thesis, some delimitations have to be made. I will not discuss whether it is possible to develop a sophisticated and most humanlike AI in the near or far future. Nor will I discuss whether it is ethically desirable or permissible to develop and use such an AI. I will also leave out discussions about what consequences AI may have in other areas, be they legal, ethical or societal.

The disposition of the thesis will be as follows. First, some background discussions are given in chapter 2 and 3. In chapter 2, definitions of AI and related concepts as well as AI's limitations and possibilities are summarized. Chapter 3 gives a description of the concept of personal relationship and a discussion of what properties are required for a relationship to be characterized as friendship. The main discussion of whether humans and AI can be friends is given in chapter 4 and finally, conclusions are drawn in chapter 5.

2 Computers, AI and mind

The idea of artificial beings is not new. In Greek mythology, there was Talos, a giant automaton made of bronze, and in fiction we can find Mary Shelley's Frankenstein. When talking of artificial intelligence today, we usually mean AI realized as some kind of machine, computer or a robot. What do all these words mean?

2.1 Terminology and definitions

Machine

"Machine, device, having a unique purpose, that augments or replaces human or animal effort for the accomplishment of physical tasks" (The Editors of Encyclopaedia Britannica, 2018). This includes simple devices as wheel and screw, as well as more advanced devices like cars and computers. All machines have input, output and some component that transforms the input to output.

Computer

A computer is "a programmable usually electronic device that can store, retrieve, and process data" (Merriam-Webster, 2020). It takes the input and with a program, it can carry out logical operations automatically and produce some output.

Robot

A robot is "an autonomous machine capable of sensing its environment, carrying out computations to make decisions, and performing actions in the real world" (Guizzo, 2018). A robot is a machine that uses computer programming to perform actions on its environment.

AI

It was at a summer conference in New Hampshire 1956 that the term of AI was coined, and it is defined as:

> *"AI is the field devoted to building artefacts that are intelligent, where 'intelligent' is operationalized through intelligence tests, and other tests of mental ability."* (Bringsjord & Govindarajulu, 2019)

Since then, there have been many attempts to create AI in different areas. From vacuum cleaner, personal assistant in smartphones, searching algorithms on the web, video games, diagnosis making and social robots for healthcare, to autonomous weapon systems. AI can be realized as a physical robot that can talk to you, touch you and affect you in a most concrete sense. AI can also be realized as a computer that can interact with you but is not an entity that can affect its environment physically.

AI can be distinguished between *weak* or *strong*. Weak AI "seeks to build information-processing machines that *appear* to have the full mental repertoire of human persons" (Bringsjord & Govindarajulu, 2019). Usually, weak AI is designed for a particular task, i.e. playing video games or being a personal assistant as Apple's Siri. Strong AI on the other hand "seeks to create artificial persons: machines that have all the mental powers we have, including phenomenal consciousness" (Bringsjord & Govindarajulu, 2019). Strong AI has a mind of its own. They can process and make independent decisions, while weak AI can only simulate human behaviour.

There have been several successes in the area of weak AI; an AI that is superior to human in some small tasks, like playing chess or Jeopardy. However, strong AI does not currently exist. The question of whether strong AI will exist in the future is still disputed (Bringsjord & Govindarajulu, 2019).

2.2 Is it all about input and output?

AI, weak or strong, is about processing input and giving correct output for some desired behaviour, simple or sophisticated. A robot perceives the external environment by receiving input from some sensors. It then processes the data and produces behaviour or actions on the basis of these perceptions. Does it matter how the input has been transformed as long as you get the correct output?

2.2.1 The Turing test

In his paper "Computing machinery and intelligence", Alan Turing proposes a test, the so-called Turing test, dealing with the question whether machines can think (Turing, 1950). A human evaluator is having a text conversation with a human and a machine, but he does not know which of them is a machine and which is a human. The goal for the evaluator is to reliably tell the machine from the human. If he does not succeed, then the machine is said to have passed the test, and it is counted as intelligent. Since the result of the Turing test is based on how the machine behaves or *appears* to behave, a weak AI would be able to pass the test and can be said to be intelligent. Regarding strong AI, there have been many objections and replies to the Turing test and one of them is the Chinese room argument.

2.2.2 The Chinese room argument

In "Minds, brains and programs", John Searle presents a thought experiment, the Chinese Room argument (Searle, 1980). Searle does not know Chinese and is locked in a room with a sheet of Chinese writing. He is given another sheet of Chinese script together with rules, in English, for correlating the second sheet with the first sheet. He can answer to questions given by people outside the room in Chinese by using those rules. No one outside can tell he does not speak Chinese. Searle behaves like a computer and the claim is that such a computer would pass the Turing test, and hence be considered as intelligent.

The Chinese room with Searle locked inside can be regarded as a whole system, a robot that is able to produce correct output. According to Searle, despite the fact that this robot can pass the Turing test in conversing in Chinese, it cannot be said to understand Chinese, it still follows instructions and manipulates formal symbols (Searle 1980, p. 420). Even though the robot would simulate all the brain processes of a Chinese speaker, it won't have simulated the causal properties of the brain, that is its ability to produce intentional states. Searle concludes that a robot could think but it is not a sufficient condition of understanding; "the computer has a syntax but no semantics" (Searle 1980, p. 423). Understanding is connected to the mind and not only to the brain processes. If the computer only mimics the brain processes and produces the correct output given the input, then it has no understanding. Passing the Turing test does not prove anything about the possibilities of strong AI.

Some objections have been made against Searle's Chinese room argument. Does anything else exist beyond the robot manipulating input and producing correct output?

2.2.3 Virtual person

In "Artificial intelligence and personal identity", David Cole argues against Searle's argument that though a computer might pass the Turing test, no computer will ever actually understand natural language or have beliefs (Cole, 1991). Cole agrees that no computer will ever understand Chinese. But this is consistent with the computer's causing a *new entity* to exist

a) that is not identical with the computer,
b) that exists solely in virtue of the machine's computational activity,
c) that does understand Chinese.

Showing that the machine itself does not understand, does not show that nothing does. From the fact that *someone* does not understand Chinese, it does not follow that *no one* understands Chinese. The fact that Searle does not understand Chinese is just that the person who understands Chinese is *not* Searle. Cole argues that there may be a mind realized by Searle's activity, a *virtual person* that understands Chinese, who is not Searle nor Searle with paper and manuals. From the fact that there is a single physical system, then, nothing follows about the number of minds which the system might realize. The Chinese room argument shows nothing about the possibilities of strong AI. Following Cole's argument, AI cannot be said to be merely a system that manipulates input and gives correct output. Though the system neither understands Chinese nor has a mind, it gives rise to an existence of a virtual person.

To summarize, a machine is said to be intelligent if it passes the Turing test. However, the Chinese room argument shows that though the machine behaves as if it is intelligent and understands Chinese, there is no real understanding and consciousness. Cole argues that though the machine does not understand Chinese, there may be a virtual person with a mind that understands Chinese.

When relating to an AI, does it make any difference whether the AI appears *as if* it has human mental capacities (weak AI) or the AI *in fact* has these capacities (strong AI), that is being conscious and having a mind? Before discussing how these properties of AI would affect how humans relate to AI, the concept of personal relationships will be explored in the next chapter.

3 Friendship and moral status

Humans are interacting with each other in various places and situations. We establish different relationships with different persons. In order to explore a possible friendship between a human and an AI, interpersonal relationships in general and friendship in particular need to be defined. I consider LaFollette's theory of interpersonal relationships to be most complete, since besides defining the concept of friendship, it also puts it in a wider context of interpersonal relationships. There are several other definitions of friendship; for instance, mutual caring, intimacy or shared activity are considered to be essential properties (Helm, 2017). However, these elements are also covered by LaFollette's view.

3.1 LaFollette's categories of interpersonal relationship

In "Personal relationships: love, identity, and morality", Hugh LaFollette categorizes different kinds of interpersonal relationships (LaFollette, 1996). He defines the difference between *personal* and *impersonal* relationships. You have an *impersonal* relationship with someone if the aim only is to satisfy some need, and it is of no matter who this person is as long as he or she can offer you the service you want. On the other hand, a relationship is *personal* when you relate to someone as a unique individual, and not only for filling a role or satisfying a need (LaFollette 1996, p. 4).

Personal relationships can be *close* or *un-close*. In close personal relationships, you want to promote each other's interests. The relationship is also reciprocal and voluntary. In un-close personal relationships, you relate to someone personally but you might want to make them miserable, as in the case of enemies (LaFollette 1996, p. 10).

LaFollette distinguishes between *rigid* and *historical* love. When your relationship is based on your attachment to that unique individual, regardless of her or his traits, then it is rigid. This is the case of brother, parents and children. A relationship is historical if the reason is you find the other person's traits attractive, as in the case of friends

(LaFollette 1996, p. 5). The different concepts of relationships can be summarized as following:

- Impersonal relationships – only filling a role or satisfying a need.
- Personal relationships – relating to each other as unique individuals:
    - Close relationship – both parts want to promote the other's interests, reciprocal and voluntary relationship.
    - Un-close relationship – might want to make the other miserable, like enemies.
    - Historical relationships – you find the other person's traits attractive.
    - Rigid relationships – attachment to a unique individual, regardless of her traits.

LaFollette's focus is on relationships between persons and does not include relationships between persons and for example inanimate objects, non-human animals[1], God or AI. With LaFollette's categorizing, friendship is defined as a close historical personal relationship.

3.1.1 Values of personal relationships

LaFollette claims that it is important for each of us to have a positive sense of self, and personal relationships is crucial in developing the sense of self-worth (LaFollette 1996, p. 87). We mirror ourselves in our friends, if I know that my friend cares for me and is kind to me, I am more likely to see myself as lovable, which enhances my self-image. However, my self-esteem cannot be promoted by someone who always agrees with me. I can esteem myself only if you are aware of my faults and point them out to me. Friends can disagree and get angry with each other but they accept each other and learn that they are worthwhile even with their faults. "If we have never been subjected to criticism, our so-called self-esteem may crumble the first time we are challenged" (LaFollette 1996, p. 88).

Being a close friend both requires and promotes self-knowledge. In order to disclose yourself to me and share intimate details about who you are, then you must first know to some degree who you are (LaFollette 1996, p. 88). To be earnest and trust each other, we learn more about ourselves from close friends than from anyone else. We also help each other to develop each other's behaviour or character. Criticism and personal assistance from our friends can help us change in ways we cannot change on our own.

[1] Human beings are referred to as human animals.

Besides the properties and values of friendship, some more fundamental requirement is needed for a friendship. A glance at the classical philosophers will be taken in the next section.

3.2 Aristotle and Kant

According to Aristotle, there are three types of friendships: friendship of utility, friendship of pleasure and complete or perfect friendship. Complete friendship is "the friendship of men who are good, and alike in virtue; for these wish well alike to each other qua good, and they are good themselves" (Aristotle 1999, p. 130). Non-perfect friendships happen when relating to each other is on the base of utility or pleasure. Such friendships only last as long as there exists a need of the utility and a pleasure of being in each other's company. Furthermore, Aristotle argues that perfect friendships are tightly bound to moral, since "only complete friends can be good people and only good people can be complete friends" (LaFollette 1996, p. 16). Close friends should be on an equal level of moral character.

Kant has a similar view on friendship and he denotes an ideal friendship as a "friendship of disposition", which is characterized by unselfishness and contains equality and reciprocity (Van Impe, 2011). Kant denies that there can be any true friendship between unequals. In a relationship of inequality, you cannot share your thoughts, judgments, feelings and lives with one another. In particular, friendship requires equal mutual love but also equal mutual respect.

According to Aristotle and Kant, some basic equality is needed for a friendship and this equality is moral status. In determining whether a human and an AI can be friends, the question of AI's moral status has to be answered.

3.3 Moral status

According to Mary Anne Warren, "the concept of moral status is […] a means of specifying those entities towards which we believe ourselves to have moral obligations, as well as something of what we take those obligations to be" (Warren, 2000). If an entity has moral status, then we have to consider its needs, interests and wellbeing, and above all, the reason to do so is because its needs have moral importance in their own right. Some philosophers allow for the possibility that moral status comes in degrees and the highest degree of status is *full moral status* (FMS) (Jaworska & Tannenbaum, 2018). There are different criteria on what kind of beings or entities can be said to have FMS.

Life criterion

Every entity that is a living organism has full and equal moral status, even bacteria (Schweitzer in Warren, 2000). A living organism is an entity that has capacity to grow through metabolism, reproduce, and adapt to environment.

Sentience criterion

Sentience is the capacity to experience "suffering and enjoyment, from simple feelings of pleasure or pain, to more complex emotions, moods, and passions" (Warren, 2000). An entity that has experiences, however simple or primitive, is not just a thing, but a being, a centre of consciousness. All entities with the capacity of being sentient have full and equal moral status.

Criterion of personhood

Persons are beings that are not only sentient but also possess more sophisticated mental capacities, such as intentionality, rationality, self-awareness, having beliefs and desires, remembering the past and anticipating the future (Regan in Warren, 2000). All persons with these capacities have equal and full moral status.

Relational criterion

The criteria for moral status mentioned above are based on an entity's *intrinsic* property, i.e. it is logically possible for the entity to have this property even if it were the only thing to exist. On the contrary, the relational criterion is based on an entity's relational properties, which are not possible to have if it were the only thing to exist. One such relational property is membership within a social and biological community (Callicott in Warren, 2000), for instance being a member of the human species or being citizen of Sweden. Another relational property is being emotional connected to each other, for instance friendship (Noddings in Warren, 2000).

As stated above, moral status is a basic equality that can be used to determine whether AI and humans can be friends. Since friendship is a relational property, the relational criterion would lead to an endless regress. To determine whether AI and humans can be friends, the moral status of AI has to be assessed. But to be able to define an AI's moral status, the relationship between humans and AI has to be determined. Regarding friendship, it is therefore only plausible to use the intrinsic properties as criteria for moral status.

To summarize, in terms of LaFollette's categorizing of interpersonal relationships, friendship is a close historical personal relationship. What is crucial for two beings to be true friends is that they have equal moral status, which in turn can be fulfilled by different criteria[2]. In the next chapter, I will give my own view and discuss whether the criteria of moral status and friendship can be fulfilled in the case of AI.

[2] The discussion on friendship and moral status of AI has many similarities with the discussion of whether humans and animals can be friends (Townley, 2010). What moral status does my dog have? Is it replaceable and do I relate to it as a unique individual? The comparison between AI and non-human animals must not be neglected but will for now be left out in the discussion that follows.

4 Can I and AI be friends?

In fiction, the answer to this question is obviously yes. Steven Spielberg's movie "A.I. Artificial Intelligence" is about the child android David that is programmed with the ability to love (Spielberg, 2001). In the movie "Her", Theodore is falling in love with Samantha, an artificially intelligent virtual assistant personified through a female voice (Jonze, 2013). One may think that the question of friendship is rather subjective; it should be up to everyone to decide whether an AI can be regarded as a friend or not. I can consider an AI as my friend unaware of what criteria have to be fulfilled for a friendship. However, from a moral philosophical point of view, an objective analysis is needed and the subjective view must be abandoned or at least set aside for the moment. In order to answer the question of whether AI and humans can be friends, I will organize the discussion as follows:

- Does AI have the same moral status as a human, according to the criteria of life, sentience and personhood?
- Can a relationship between a human and an AI be regarded as friendship (close personal historical relationship) according to LaFollette?

These questions will be answered in the light of weak and strong AI, that is if the AI appears to have human mental capacities or in fact has these capacities. I will also distinguish between current and future weak AI. Current weak AI are those AI that exist today. Future weak AI are those that may be developed in the future and are more sophisticated and humanlike compared with current weak AI. In focusing on personal relationships, I will discuss the case of AI that enter our homes, our private lives and are becoming increasingly part of our daily routines.

4.1 Weak AI

An AI is considered as weak AI if it lacks human mental capacities such as consciousness and self-awareness. According to Searle's Chinese room argument, though a most sophisticated AI would pass the Turing test in some aspect, it stills does not have a mind. With this view, all existing AI, and perhaps also all future AI, would be classified as weak AI. This section will discuss the case of weak AI that *appears* as if it has human mental capacities.

One example of current weak AI is Apple's virtual assistant Siri who can answer questions, make recommendations and reminders, text and send messages for you. Since Siri is a form of weak AI, she cannot have consciousness or intention as a human. She has been programmed to fill a role as an assistant to satisfy your need of managing different tasks.

Another example of weak AI is a social robot that "is able to communicate and interact with us, understand and even relate to us, in a personal way. It is a robot that is socially intelligent in a human-like way." (Breazeal in Scheutz, 2009). Unlike Siri, a social robot is a physical entity that can affect its environment and its object for interaction in a most concrete sense. Social robots can for instance be used as entertaining toys, to practice children with autism in social skill, or to accompany elder people in healthcare (Elder, 2018). The goal with these social robots is to establish some kind of social contact with the human, they are designed to appeal to our emotions. In the case of accompanying elder people, the role of the robot is to talk and respond to the human, to make her feel less lonely. As in the case of Siri, a social robot cannot have consciousness or intention, though it might act as if it has.

4.1.1 Moral status of weak AI

According to the criterion of life, only living organisms have full moral status. If defining life as capacity to grow through metabolism, reproduce and adapt to environment, then obviously, neither current weak AI as Siri and social robots nor future weak AI possess all these capacities. Possibly they can be developed to have the capacity to adapt to the environment, but only having one of these capacities is not sufficient to fulfil the criterion of life. Consequently, they cannot have the same moral status as humans.

Although a weak AI may behave as if it is happy or sad, there are still no conscious mental states and therefore no experiences of pleasure or pain. Weak AI, current or future, can therefore not meet the sentience criterion, and consequently cannot have the same moral status as humans. For the criterion of personhood, the entity needs to have more sophisticated mental capacities, such as intentionality, rationality, self-awareness, having beliefs and desires. Obviously, weak AI does not meet this criterion either.

To conclude, based on these criteria, weak AI, current or future, cannot have the same moral status as humans. According to Aristotle and Kant, whose claim is that moral equality is required for a complete friendship, weak AI and humans can never be true friends. However, to each of the criteria for moral status, there are disputes about whether it is too including or excluding (Jaworska & Tannenbaum, 2018). For instance, infants, severely impaired and unconscious humans would not meet the criteria of personhood since they lack mental capacities such as rationality and self-awareness. Consequently, true friendship would not arise between entities belonging to such category. When assessing whether AI has the same moral status as humans, I refer to those humans who have these mental capacities.

4.1.2 Relationship with weak AI according to LaFollette

Personal relationship

A personal relationship is characterized by relating to each other as unique individuals and not only for filling a role or satisfying a need. Apparently, you do not relate to Siri as a unique individual, it could be any virtual assistant that fills the role. It is not crucial for you, at least not in an emotional way, that it is Siri who is your virtual assistant. It could do well with another virtual assistant, perhaps with another app. As long as Siri can provide you the service you want, you wish to relate to her. Once your need cannot be satisfied by Siri, or some other more sophisticated virtual assistant is available on the market, you would abandon Siri. A relationship to Siri is defined as an *impersonal* relationship.

Regarding the social robots or any commercial robots, if there exist several copies of them, then there is no unique AI individual. You could have chosen any of the AI copies to socialize with. However, a future AI can be custom-made to be an entity programmed with a unique history and personality, or even to suit exactly your person. In such case, it would be unique for you, since it is a response to your unique personality. The question is whether you are unique for the robot. If the robot could be programmed to bond with any person, then you are not a unique entity for it. Such a relationship would not be defined as a personal relationship. However, if the robot is programmed to bond with exactly you, as in the case of David in the movie "A.I." who is bonded to his human mother Monica, then you are unique for the robot and you are involved in a *personal* relationship.

Close relationship

A close relationship is reciprocal and voluntary. If you have chosen to interact with an AI by your own will, then the relationship is voluntary from your side. But can one then say the relationship is voluntary from the AI's point of view? I think not, since weak AI does not have intention and cannot have chosen you by its own will. There is also an asymmetry of power. As human, you can choose when and which AI you want to be part of a relationship with. Furthermore, you can choose to quit the relationship with the AI, by pressing the power button or unplug it whenever you want. But the AI does not have this power; the relationship between you and the AI is completely on your terms. Such a relationship cannot be classified as mutually voluntary.

In a friendship, you should care for each other and do so for each other's sake. Friends influence and have a reciprocal effect on each other. That is, friends must be moved by what happens to their friends to feel the appropriate emotions: joy in their friends' successes, frustration and disappointment in their friends' failures. In the case of Siri, you can hardly say that you care for her and do so for her sake. Neither does she. The social robots can behave as if they care for you and your wellbeing, but they

do not really *feel* joy or disappointment for your sake. As Scheutz expresses it, "none of the social robots available for purchase today (or in the foreseeable future, for that matter) *care* about humans, simply because they *cannot* care" (Scheutz, 2009). The AI has been programmed to be your friend and it behaves as if it reciprocates your feelings. Such a relationship cannot be classified as reciprocal.

However, humans can create a unidirectional bond to the robot and care for it after all. One example is a robot developed for defusing of land mines. The robot is designed with several legs and when stepping on a mine, one leg is destroyed (Scheutz, 2009). On one test occasion, when it only had one leg left, the colonel in command cancelled its mission, since he found it inhumane to watch the crippled robot pulled itself forward with only one leg. Though the robot is a lifeless device, the human ascribes it sentient capacities such as capacity of feeling pain. Hence, the human has created a unidirectional bond to the robot and is unwilling to see it approaching its own destruction.

Historical relationship

In a historical relationship, you find each other's traits attractive. I'm friend with you because of who you are and what you are like, and not only because of you being my brother regardless of your traits. I may desire to be friends with you because you are funny or intellectual, but that does not mean I want to be friends with anyone with these properties (LaFollette 1996, p. 6). Even though two entities have the same traits, they may not exhibit them in the same way. Furthermore, to fully understand someone's historical dimension, we should know not only her traits, but also how she has evolved.

Current AI are merely copies of each other, they can adapt their behaviour to humans in some degree but do not relate to a specific person because of his or her traits. If AI would be designed to have individual characters as well as the capability to evolve itself and adapt its behaviour to you, then it is possible to have a historical relationship with it. Suppose you and I would have a robot each, and they are identical from start. As time goes by, my robot would adapt to me and different situations that arise during its interaction with me. Your robot would do the same for you. We would not change our robots with each other, we are now relating to them because of what they are like. Likewise, my robot would not consider me as interchangeable and can switch to be your companion instead, if disregarding the possibility of rebooting it to adapt to your personality. In this sense, I would claim you can have a historical relationship with an AI.

Values of friendship

Friends function as a kind of mirror of each other. I do not have complete knowledge about my own character, but by knowing a friend who reflects my qualities of character, I can come to know strengths and weaknesses of my character. Such mirroring allows us to enhance our sense of self-worth and shape each other's interests and values. Current AI may use different facial expressions to respond to your reactions (Scheutz, 2009). In the future, the AI may be developed to be more sophisticated in its response and it behaves as if it reflects your character. By interacting with me, it can learn and refine its behaviour. Likewise, I can be shaped by the AI and alter my character to be healthier or to be less prejudiced.

However, according to LaFollette, to enhance my self-esteem, it is not sufficient with someone who always agrees with me. I need a friend who also can criticize me and be furious with me. Current AI still fall short regarding these capabilities, but a future weak AI may be developed to be impatient and angry with you, though they lack these feelings. In that case, you would perceive it as more authentic and not only programmed to please you. On the other hand, some people may prefer the even mood and predictability of an AI in some cases. An AI cannot be tired, get upset or lose its temper, it cannot lie or disclose your secret to anyone else, if it has been programmed in such way [3]. Some people may find an AI more reliable than humans and feel more comfortable to reveal their problems to a robot (Elder, 2018).

To summarize, a relationship between human and current weak AI lacks all the properties of being close personal historical and cannot be categorized as friendship. A relationship with a future weak AI can be personal, historical but not close, and consequently cannot be considered as a genuine friendship. However, it may have some values that resemble those of friendship.

4.2 Strong AI

Does it make any difference if an AI behaves as if it has mental capacities or it in fact has these capacities? With the Chinese room argument, Searle claims that all AI, however sophisticated and humanlike, cannot have a mind. Even though an AI would pass the Turing test and be indistinguishable from a human, no conclusions can be drawn about its possession of a mind. Accordingly, nothing can be said about the possibility of strong AI. However, following Cole's view, there may exist a virtual person, besides the machine. This dualistic view implies your person is not only your body with all organs and brain processes; there is also a mind. With this reasoning, an AI can be said to have a virtual identity and it is not impossible that it has something called a mind, which is an assumption for strong AI.

[3] There exists a risk that the robot is hacked and your secrets can be leaked to other people. But for now, this risk can be neglected.

It is of course not trivial to know with absolute certainty that an AI actually has these mental capacities. However sophisticated and humanlike an AI may be, there is still a small doubt if there is a mind or it is only an empty shell that resembles a human[4]. In the Star Trek episode "The measure of a man", there is a dispute whether Data, an android officer on the starship Enterprise, is sentient and has the same moral status as humans (Scheerer, 1989). Though Data can be regarded as a strong AI, you can never be absolutely sure that he actually has a mind or a soul. But we cannot have this knowledge about other human minds either, for that matter. According to the so called "other minds problem", I cannot justify the belief that you, my friend, have thoughts, feelings and other mental attributes (Avramides, 2019). It is beyond this thesis to discuss this issue, but for now, when discussing strong AI, it is assumed that we know it having a mind. How would the presence of a mind affect the fulfilment of the criteria for moral status and the properties of friendship?

4.2.1 Moral status of strong AI

Still, a strong AI would not meet the criterion of life. It may be able to reproduce and duplicate themselves, at least considering the case of virtual entities. However, I cannot see how an AI would be able to grow through metabolism. A strong AI with a mind can be said to be sentient and has beliefs and desires. Therefore, it would meet the criteria of sentience and personhood. Data in Star Trek would not meet the criterion of life but possess capacity of intelligence and self-awareness and hence meet the criteria of sentience and personhood.

4.2.2 Relationship with strong AI according to LaFollette

Personal relationship

Can I relate to a strong AI as a unique individual? As in the case of weak AI, if the AI only exists as one copy or there are several copies but your AI is developed to adapt to your personality, then it is unique and not interchangeable. For a comparison, consider the case of twins or human clones. Although they are each other's exact copies from the beginning, they will come to develop to different personalities and hence not interchangeable. Likewise, if the AI has been programmed to be your companion, then you are unique for the AI. The conclusion here is the same for strong as weak AI; it is possible with a personal relationship between humans and AI.

Close relationship

Since a strong AI has intention, desires and beliefs, one can say that it can reciprocate your feelings in a genuine way. With its own desires and beliefs, an AI is also capable of choosing with whom it wants to relate. On your side, you can choose to enter a relationship with any AI you want, and when interacting with it, you would reciprocate its feelings. Such a relationship would be classified as close. Furthermore, there would

[4] Pointed out by Per Algander, the chairman of the opposition seminar on 25 May 2020.

exist a power symmetry between you and the AI, which does not exist in the case of weak AI. Since the AI has the capacity of being sentient and conscious, it would have the same moral status as humans. Consequently, we would have obligations to it and cannot treat it any way we please, for instance unplug it or sell it to someone else.

One may ask if it is really the case that an AI can choose you voluntarily if it has been created by some robotics and AI researcher. The question is whether such an entity has free will, even if it has a mind and consciousness. However, the discussion of free will is also applicable to humans; and despite the fact of its importance, it is considered to be beyond the scope of this thesis. Therefore, I will assume that AI has a free will if it is considered to have desires and beliefs, regardless of how it was created.

Historical relationship

With a mind, the AI is capable of evolving its own personality and traits, and exhibit them on its own way. When you have spent a lot of time with the AI, both of you would know how you have evolved and appreciate each other's traits. You relate to the AI not in a rigid way, not because it is of a certain brand or manufacture, it is because of its traits and how it has evolved. Hence, it is possible to have a historical relationship with an AI.

Values of friendship

A strong AI with human mental capacities would possess a wide range of emotions and reactions. It can shape and be shaped by your character, that is you can mirror each other and develop your sense of self-image. Since the only difference between a strong AI and a human is the different material they are constituted of, a relationship between these entities would have all values that characterizes a friendship.

4.3 Summary of the analysis

According to Aristotle and Kant, only persons with equal moral status can be true friends. By embracing this view, moral status is a necessary condition for friendship. Table 1 below summarizes fulfilment of the different criteria for moral status for weak respectively strong AI.

Table 1 Fulfilment of the criteria for moral status.

	Life	Sentience	Personhood
Current weak AI	No	No	No
Future weak AI	No	No	No
Strong AI	No	Yes	Yes

Neither current nor future weak AI meets any of the criteria for moral status. Strong AI with a mind meets the criterion of sentience and personhood but not of life.

Therefore, considering the life criterion, an AI can never have the same moral status as a human, implying there cannot exist a true friendship between them. Table 2 shows the results of whether AI falls within LaFollette's classification of friendship.

Table 2 Fulfilment of the criteria for friendship as defined by LaFollette.

	Personal	Close	Historical	Values of friendship
Current weak AI	No	No	No	No
Future weak AI	Yes	No	Yes	Yes
Strong AI	Yes	Yes	Yes	Yes

A relationship between a human and current weak AI meets none of LaFollette's criteria for friendship. A relationship with a strong AI would have all the essential properties for friendship. Since future weak AI cannot be involved in a close relationship, a relationship with such an AI would not be characterized as a friendship. Following these results and the assumption that moral status is necessary for friendship, one can conclude that true friendship can never exist between humans and AI if using the life criterion. When applying the criterion of sentience and personhood, neither current nor future weak AI can be a human's true friend, despite the fact that they can fulfil some of the essential properties for friendship. Strong AI meets the criteria of sentience and personhood as well as LaFollette's definition of what essential properties a friendship should be consisted of.

4.4 From a subjective point of view

The analysis above shows that, depending on what criterion of moral status is applied, I and strong AI can be friends, but there can never be a genuine friendship between me and a weak AI. This result is based on a moral philosophical view, but concerning relationship, the subjective point of view cannot be neglected so easily.

Not fulfilling the different criteria for friendship does not necessary imply you cannot have feelings toward an AI. In some cases, AI fails to meet a certain criterion because of it lacking mental states. Does it really matter whether an AI has a mind or not? Since the AI acts as if it were a human, you would relate to it as you do with a human. The AI talks to you, laughs at your jokes, asks about your favourite book, discusses philosophy with you. It is a seamless integration of all programming and algorithms resulting in a believable humanlike creature. The AI acts as if it is conscious and enjoying your company. Why should you relate to it differently?

Many studies have shown that people anthropomorphize all sorts of technologies (Breazeal, 2003). We attribute mental states (i.e. intents, beliefs, feelings, desires, etc.) to describe the behaviour of machines, cars and computers. It is then most natural that we anthropomorphize the AI realized either as a virtual assistant, a social robot or an entity indistinguishable from human. We relate to the AI as if it has mental states,

as if it has a mind and being conscious. The anthropomorphizing of AI may have the consequence that humans tend to create a unidirectional bond to the AI though it does not possess these mental properties (Scheutz, 2009).

The unidirectional bond is reinforced if the AI behaves in a most humanlike manner and also in cases where you suffer from some mental condition, like dementia, such that you are not aware of the artificiality of your friend. You would have all sorts of feelings about the AI, you would be sad if you lose your AI, you would be happy for its return. However, even in cases when you are aware of it is an AI you are interacting with and it is not a very sophisticated AI, there is a tendency to anthropomorphize it and ascribe it capabilities that it does not have.

5 Conclusions

The answer to the question of whether AI and humans can be friends is both yes and no. In this thesis, I have embraced the ancient philosophers' claim that moral status is a basic equality or a necessary condition for two entities to be true friends. If applying the criterion of life for moral status, neither weak nor strong AI can ever be my true friend. Applying the criterion of sentience and personhood, I can be involved in a genuine friendship with a strong but not a weak AI. However, from a subjective point of view, it is possible to create unidirectional emotions towards an AI regardless of it having a mind or not.

Moral status has been used as a ground for friendship; however, the reasoning could be reversed, that is friendship could be considered as a criterion for moral status. If I consider an AI as my friend, then this relational status should imply it having the same moral status as a human. This would affect our rights and obligations toward AI, and AI's rights and obligations toward us. Discussion of an AI's rights and obligations is beyond the scope of this thesis, but the question of whether I and AI can be friends may have considerable consequences on our lives and society.

In assessing the possibility of whether AI and humans can be friends, questions concerning essential properties of humans, such as free will and consciousness, are raised as well. If AI is under scrutiny for being assessed of having a mind or not, humans cannot escape this doubt either. In expanding our knowledge about a different form of being, we also come to learn more about ourselves as human creatures.

6 Bibliography

Aristotle. (1999). Nicomachean ethics, Book VIII. (W. Ross, Trans.) Batoche books.

Avramides, A. (2019). *Other minds.* (E. N. Zalta, Ed.) Retrieved June 5, 2020, from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/other-minds/

Breazeal, C. (2003). Towards sociable robots. *Robotics and autonomous systems, 42*(3), 167-175.

Bringsjord, S., & Govindarajulu, N. S. (2019). *Artificial intelligence.* (E. N. Zalta, Ed.) Retrieved April 4, 2020, from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/win2019/entries/artificial-intelligence/

Cole, D. (1991). Artificial intelligence and personal identity. *Synthese 88*, 399-417.

Elder, A. (2018). *Friendship, robots, and social media: false friends and second selves.* London: Routledge.

Guizzo, E. (2018). *What is a robot?* Retrieved May 11, 2020, from Robot your guide to the world of robotics: https://robots.ieee.org/learn/

Helm, B. (2017). *Friendship.* (E. N. Zalta, Ed.) Retrieved May 2, 2020, from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/fall2017/entries/friendship/

Jaworska, A., & Tannenbaum, J. (2018). *The grounds of moral status.* (E. N. Zalta, Editor) Retrieved May 2, 2020, from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/spr2018/entries/grounds-moral-status

Jonze, S. (Director). (2013). *Her* [Motion Picture].

LaFollette, H. (1996). *Personal relationships: love, identity, and morality.* Cambridge: Blackwell.

Merriam-Webster. (2020). *Computer.* Retrieved May 11, 2020, from Merriam-Webster.com dictionary: https://www.merriam-webster.com/dictionary/computer

Oakley, J. (2013). Personal relationships. (H. LaFollette, Ed.) *International Encyclopedia of Ethics.*

Scheerer, R. (Director). (1989). *Star Trek: The next generation, The measure of a man, season 2, episode 9* [Motion Picture].

Scheutz, M. (2009). The inherent dangers of unidirectional emotional bonds between humans and social robots. *Workshop on Roboetics at ICRA.*

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 417-424.

Spielberg, S. (Director). (2001). *A.I. Artificial intelligence* [Motion Picture].

The Editors of Encyclopaedia Britannica. (2018). Retrieved May 11, 2020, from Encyclopædia Britannica: https://www.britannica.com/technology/machine

Townley, C. (2010). Animals as friends. *Between the Species, 13*(10), 45-59.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-60.

Van Impe, S. (2011). Kant on friendship. *International Journal of Arts & Sciences, 4*(3), 127-139.

Warren, M. A. (2000). Moral status: obligations to persons and other living things. Oxford: Oxford university press.