Can I Have a Robot Friend?
Abstract
The development of autonomous social robots is still in its infancy, but there is no reason to think that it will not continue. In fact, the robotics industry is growing rapidly. Since this trend is showing no signs of abating it is relevant to ask what type of relations we can have with these machines. Is it for example possible to be friends with them? In this thesis I argue that it is unlikely that we will ever be able to be friends with robots. To believe otherwise is to be deceived, a trap it is all too easy to fall into since the efforts put on making social robots as human-like as possible and to make the human-robot interaction as smooth as possible are huge. But robots are not always what they seem. For instance, the capacity to enter into a friendship of one’s own volition is a core requirement for a relationship to be termed friendship. We also have a duty to act morally towards our friends, to treat them with due respect. To be able to do this we need to have self-knowledge, a sense of ourselves as persons in our own right. We do not have robots who display these capacities today, nor is it a given that we ever will.

Sammanfattning
# Table of Contents

1. Introduction ..................................................................................................................3

2. Friendship .....................................................................................................................6
   2.1. Special Duties of Friendship .................................................................................6
   2.2. Mutuality ................................................................................................................8
   2.3. Shared Activities ..................................................................................................11
   2.4. Self-knowledge .....................................................................................................12

3. Can a Robot have Self-knowledge? .............................................................................14

4. Deception and Other Problems ................................................................................17
   4.1. Deception .............................................................................................................17
   4.2. Sharing of Information .......................................................................................20
   4.3. Interspecies Relationships ................................................................................21

5. Conclusion ..................................................................................................................23

References ......................................................................................................................26
1. Introduction

In recent years we have seen an increase in the availability of robots designed to assist elderly or otherwise incapacitated people with physical tasks, cognitive issues, health management and psychosocial issues. At the same time there is a growing trend to develop purely social robots, which in a not too distant future might be bought on a commercial market and used for companionship and/or entertainment. So called sex robots would fall within this scope. In Japan and Korea autonomous robots are already operating as shop assistants in malls and they have been used as companions in kindergartens in Korea\(^1\). In Japan, a hotel is staffed by robots, although assisted by humans. What distinguishes them from traditional task-oriented robots, which are often found in factories and perform certain tasks in limited applications, is that these interaction-oriented robots are made to communicate with humans and partake in human society. They are often given human-like features or made to look like animals, primarily what we think of as "cute" ones. Examples of these are Aibo, which is developed by Sony and looks and behaves like a puppet, and Paro, a Japanese robot shaped like a baby seal which can move and make seal noises and has been sold in the thousands across Europe, the US and Asia for therapy use mainly in rest homes and in dementia care. A more human-like example is Kismet, developed by MIT and consisting of nothing more than a head, a so-called "face robot" capable of recognising and simulating facial expressions (Broadbent 2017: 629–631; Kanda & Ishiguro 2006: 370).

It is of course easier to connect to and interact with something that resembles a living, breathing creature than a machine arm or a drone, and these robots are intentionally made to look appealing and "harmless" in order to make the human-robot interaction go as smoothly as possible. When constructing human-like robots, robot engineers (or roboticists) are concerned not only with solving technical difficulties but also to define and model a range of complex human behaviours such as proxemics and gaze behaviour. Proxemics deals with the physical and psychological distance we place between each other and which differ depending on, for instance, gender, age and culture. To be better integrated in human society robots need to show appropriate distancing behaviour such as not standing too close or too far away from the person they are interacting with: you do not want to feel threatened when the robot hands you an item, so which is the appropriate distance from which to do this? Should they

---

\(^{1}\) In 2006, the South Korean Ministry of Information and Communication predicted that robots will be in every South Korean household sometime between 2015 and 2020. A somewhat optimistic forecast (Onishi, 2006, April 2).
approach you in a straight-line path or come at you from an angle? Gaze behaviour is concerned with the movement of the eye or "gaze acts," such as long and short glances, gaze aversion and gaze shifts: where to look, how long to look someone in the eye, whether to turn one’s gaze toward the thing one is talking about. This, together with head movement, affect how the human interprets the intentions of the robot (Broadbent 2017: 636).

The development of autonomous social robots is still in its infancy, but there is no reason to think that it will not continue. At least in countries with an ageing population where the incentives to ease the pressure on the health care sector are high or where a large enough segment of the general public are motivated enough to purchase a social robot, for pleasure or utility reasons, to make the development and marketing of such devices economically viable. Since this trend is showing no signs of abating and all signs of increasing it is relevant to ask what type of relations we could have with these machines, not least since they are becoming more and more human-like. Some people already spend most of their day in the presence of a social robot and it is reasonable to think that this will only be more prevalent in the near future. And because they are designed to interact with us, to help us when our bodies fail us, to keep us company when we are lonely – to be "social” – it is interesting to find out how far the interaction can go. Is it, for example, possible to be friends with a robot?

The purpose of this thesis is to examine if friendship with a robot is possible and I will argue that we are not now, nor is it likely that we will ever be able to be friends with robots. To believe otherwise is to be deceived, a trap it is all too easy to fall into, not least because of the efforts put on making social robots as human-like as possible and to make the human-robot interaction go as smoothly as possible. But the appearances belie the limits of the machines. For instance, robots would need to be able to act morally, to treat the friend with due respect, which requires moral agency. Also, the capacity to enter into and leave a friendship of one’s own volition is a core requirement for a relationship to be termed friendship, besides the fact that friends spend time together because they wish to do so. For robots to be able to do this they would need to have self-knowledge, i.e. to have a meta-perspective of themselves as persons in their own right. We do not have such robots today, and it is not a given that we ever will.

---

2 To be able to make moral judgments and to be held accountable for one’s actions.
I will first look at friendship generally. I argue that there is a moral aspect to friendship, that friends have special duties towards each other. There are also some themes which philosophers come back to when examining friendship, namely mutuality and shared activities. Then I will show what mental capacities robots have today. In order to be able to be friends with someone we must be aware of ourselves as persons in our own right: we have to have self-knowledge. Robots do not have this. Finally, I point at the risks of being deceived (of believing robots to be capable of things that they are not capable of), the trouble with information storing and sharing, and why the problems of interspecies relationships make it highly unlikely that we will be able to form friendships with robots, even if they one day will gain self-knowledge.
2. Friendship

In this part I will study the moral view of friendship; some argue that we have special duties towards our friends. Why is that so? I will also examine two themes, which usually recur when philosophers write about friendship. These are mutuality and shared activities. Helm (2017) lists them as mutual caring, intimacy and shared activities. To simplify things, I decided to merge the first two since the separation of them seems a bit arbitrary to begin with: caring and intimacy go closely together and mutuality, which I see as fundamental for friendship, covers them both. Caring and intimacy, and for that matter shared activities, also hints at love, which certainly is a part of friendship, but instead of listing love as a theme of its own, I trust that this account will show that love permeates both themes as well as the morality aspect. Finally, I will explain why self-knowledge is a necessary requirement for the participants to have in order to be able to form and maintain a friendship. Without self-knowledge you do not have moral agency, which would make the performance of your special duties towards your friends nothing more than just a performance, an act in accordance with a set of rules rather than a personally held belief of what is right and wrong. Mutuality and shared activities without self-knowledge would likewise account for not much more than an empty set of motions: you would do what you do because that is what is expected of you, but not because you want to or care for your friend.

2.1. Special Duties of Friendship

It is arguable whether we have a duty to have friends or not, but if and when we have them we seem to have certain duties towards them. There is, perhaps not surprisingly, no real consensus on what these duties are, more specifically. According to Kant we have a duty to love our friends: "I am his friend; he is my friend therefore means nothing else but, I cherish the inclination of well-wishing towards him from maxims = I love him from duty" (Kant 1997: 407). This well-wishing towards others is the universal duty of love since we must make it our maxim to promote goodness in others, and when the feeling is reciprocated, from one individual towards another, we have friendship (407–408). Moreover, we have negative duties towards our friends: not to impose ourselves too much on them. We shall not burden them with our requirements and we shall not let them fear that they will be called upon to benefit us. We shall preserve their dignity and independence and not ask them to disclose their innermost thoughts when it is not called for; a degree of modesty and delicacy is needed. We have no duty to reprove our friends’ moral faults since to point them out will necessarily
assail their self-respect. Our main positive duty is to love our friends as friends and to engage in a reciprocal development of our principles. It is hard work, and it follows that it is almost impossible to have many friends. Aristotle\(^3\) agrees: "It is not possible to be a friend to many in the case of the complete kind of friendship" (EN 1158a10–11). He also agrees that we should try to keep our friends from sharing in our hardships, but we ourselves should wish to share our friends' burdens (EE 1245b34–36). Besides, friends can help the young in avoiding mistakes, the older by caring for them, and direct people in their prime towards fine action since two can act and notice better than one. Basically, friends help each other stay on the right path by noticing and calling out each other’s mistakes (EN 1155a12–16). Telfer is of the opinion that friends should proffer advice and criticism, even when it is not asked for since it might be what is needed. We also have a duty to help a friend under attack or in any kind of trouble (Telfer 1991: 257).

But details aside, a more fundamental question is why we have moral duties towards our friends. Friendship and justice have to do with the same things Aristotle claims, and they involve the same persons. The justice or, for that matter, the injustice increases with the degree of friendship involved, so it is worse to cheat a friend than a fellow citizen and it is worse to strike your father than a stranger (EN 1159b25–1160a9). We have an obligation towards our friends that we do not have towards everybody. Just as it is one’s duty as a parent to care for one’s child first, it is our duty as friends to look after our friends before we look to others’ needs. If, Telfer says, a man has a choice between helping a friend decorate his flat and doing the same for an old age pensioner, the friend has if not a prior claim at least a competing claim. From a rule-utilitarian position our community will be best served if we look after our friends first as it is impossible for everyone to look after everyone else. The choice of whom to help has to be made somehow, and friends seem to be at least as good a choice as anyone, if not better. We are more likely to know our friends’ needs better than we do strangers’, and we already have a certain bond, we feel a special concern for them (Telfer 1991: 262–263). From a Kantian point of view it is rational to look after your friend since the action has moral merit and the reciprocity of the friendship more or less guarantees that your friend will answer in kind and also look after you (Kant 1991: 210–212). An issue with this

---

\(^3\) The philosopher who arguably wrote most extensively about friendship. He dealt with it both in his Nicomachean Ethics (EN) and Eudemian Ethics (EE). His theory on this subject is one of the most influential and the one many later philosophers refer to; if not explicitly stated, the seeds of Aristotle is usually apparent in their work. One notable example is Kant whose survey on friendship reveals a significant debt to his predecessor.
reasoning is of course that people without friends will be left out, they will have no-one who
cares for them. But the fact that there are people without friends does not detract from the idea
that friends have a special duty towards each other, just as the fact that there are orphans does
not detract from the idea that parents have a special duty towards their children. It is
incumbent on us to care for our friends more than we do for strangers. Brody (2014: 62–63)
argues that we have a greater duty of care for persons we have a close relationship with, not
because they are more important than others, but because they are more important to us than
others are. My children are more important to me than your children, which is not to say that
my children are more important than your children. My friends are more important to me than
other people, which is not to say that they are more important than other people. The
argument does not give us carte blanche to disregard other people, but it gives us permission
to take special measures to care for those closest to us.

2.2. Mutuality
What then is meant by mutuality in regards to friendship? It is crucial that both parties accept
the existence of the relationship for it to exist. To quote the Stanford Encyclopedia of
Philosophy (SEP), ”whereas we must make conceptual room for the idea of unrequited love,
unrequited friendship is senseless” (Helm 2017). It is perfectly reasonable to, for instance,
love, hate, fear, or be indifferent to another person without that person having the same
feelings towards you, but you cannot be friends with someone who does not accept you as a
friend. However much you idolise, revere, or wish to be friends with someone, friendship
does not occur until that person also wishes to be friends with you. Anything else is a form of
deception: you either deceive yourself into thinking that a friendship exists, or you are being
deceived by the other person (more on this in part 4). The mutuality of friendship is
fundamental: I can only be friends with you if you want to be friends with me.

Why is this? According to Aristotle, one of the essential foundations of friendship is
reciprocal influence (Sharp 2012: 233). He distinguishes ideal or virtue friendship from
friendship based on utility or friendship based on pleasure. Utility friendship arises from
different kinds of business relations: someone is able to provide something you need and you
form a relationship towards that end. Friendship based on pleasure arises when someone can
provide you with some sort of pleasure or entertainment: it might be based on a love of music,

4 It is even possible to have these feelings towards someone who does not even know that you exist.
drinking, sports or some other kind of mutual interest. Utility and pleasure are not “real”
friendships, but they have the potential of evolving into something more, they contain the
seed of a more virtuous relationship. Virtue friendship, then, is a relationship between
virtuous people who seek each other’s company primarily for the sake of achieving
eudaimonia – happiness, or the good life. This eudaimonia is a thing that is good in itself, it is
an end-goal, and to reach it we need things that promote this well-being. Having friends is one
such thing (Kraut 2018). Good people seek other good people in order to promote and
reinforce what is good in themselves and it works both ways: you want what is good for your
friend for their sake, and they want what is good for you for your sake. Without this mutual
well-wishing there is no friendship, only good will:

"Good will, then, is what these people evidently feel towards each other; but how could
one call them friends, if they are not aware of their mutual feelings? If there is to be
friendship, the parties must have good will towards each other, i.e. wish good things for
each other, and be aware of the other’s doing so.” (Aristotle: EN 1156a2–5)

Whereas love resembles an affection of the soul, Aristotle sees friendship as resembling a
disposition of the soul. Anyone can suddenly start to love or feel affection for anything,
people and things alike, whereas the reciprocal loving that friendship consists of always
involves an element of decision, and conscious decision-making comes from a disposition of
the soul. When friends wish good things for the other, not just for themselves, they show
something more than just an affection: they want to promote the good in themselves as well as
the good in the other: "… for the good person, in becoming a friend, becomes a good for the
person to whom he becomes a friend” (EN 1157b34-35). The good person is to his friend as
he is to himself, because the friend is another self. This can only come about if and when
friends perceive each other, recognise each other’s existence as persons and as friends (EN
1170b6–13).

Kant, much influenced by Aristotle, also sees perfect friendship as one where you treat the
other as yourself. There are two conflicting motives for action: love for oneself and love for
humanity. Self-love assures one’s happiness but is without moral merit, while love for
humanity has moral merit but neglects one’s self. The remedy is to treat your friends as you
would treat yourself while they in their turn treat you as they would treat themselves.
Everyone’s needs will be looked after and your actions will have moral merit (Kant 1991:
210–212). Kant distinguishes between well-wishing and well-liking where the former is the universal duty we have to love our fellow man while the latter is based on the appreciation you have earned through your actions and behaviour. Well-liking, however, could always be a one-sided love affair, an amor unilateralis, while the requited kind is an amor bilateralis. While it is possible to have an amor unilateralis for the entirety of humankind, only the reciprocated love will count as friendship: "This [amor bilateralis] involves the idea of friendship itself, is based on a reciprocal trust that the other also views my best interests as his own and is a bilateral bond" (Kant 1997: 407–8). Kant, as Aristotle, understands friendship as based on a moral disposition. While loving someone or something is a natural inclination, the reciprocal love demanded by friendship is an intellectual unification of feelings and thoughts. The reciprocal love must be coupled with a mutual respect for humanity in that friends must not demean each other. We are obliged both to preserve the respect for our own person and to ward off whatever could risk taking away the respect for the humanity of the other: "Now in this lies the mutual restriction of reciprocal love among friends, and it is this which makes the nature of friendship, namely the moral union of their feelings and thoughts, into the satisfaction of a purely intellectual need" (412–3).

We do not need to accept Aristotle’s notion that only good people can become friends, nor do we have to agree with the idea that the reason for getting friends it that we want to promote and reinforce what is good in ourselves. We do however need to agree with the idea that friends accept each other as friends. Friendship occurs only when both parties have accepted it. It is a relationship you enter into voluntarily: it is based on a moral disposition; it is an intellectual unification of feelings and thoughts, which involves an element of decision making. Forming a friendship demands an acceptance from both parts, and for maintaining the relationship friends have to treat each other with a certain kind of respect. It is not a master-slave relationship, it is not a parent-child relationship, it is a relationship between at least two persons of more or less equal standing. The friend is not merely an object for my affection, he or she is another self. Kierkegaard has a similar idea: in our friends we love the other self, a form of selfish love (1991: 241–242). But since the other is another self, and since I want what is best for me, in seeing myself in the other I care for him as much as I care

---

5 "So that since the primary friendship is grounded on virtue, friends of this sort will be themselves absolutely good” (EE 1237a 10–11). The idea being that the good is always good while the bad change their constitution over the course of the day, and you cannot be friends with someone who is constantly changing their nature for the simple reason that you are friends with someone who is like yourself."
for myself: the selfish aspect of friendship guarantees that we care for each other.

2.3. Shared Activities

Friends do things together, and for Aristotle this requires that friends live together. At least that is what it looks like at first glance, "for some are friends in so far as they delight in living together […] while others, because asleep or separated by geographical distance, are not actively friends but are disposed so as to be active in that way" (EN 1157b5–10). While the disposition for active friendship might be present, physical presence is a must for the friendship to actually come alive. However, living together does not have to be taken literally since it is defined as "conversing, and sharing their talk and thoughts” and not "feeding in the same location as grazing animals” (EN 1170b12–14). People are friends when they perform friendly services to one another; merely wishing to be friends is never enough (EE 1237b10–20). Friends enjoy a similar disposition, according to Kant, "an unconcealed sharing and participation in one another” (Kant 1997: 415). It is reasonable to understand shared activities as a sharing of one another’s time and presence, rather than putting the focus on doing things.

That friends do things together seems evident at first glance, but there is something that motivates friends to engage in joint pursuits, something which separates that from the activities we share with work colleagues and mere acquaintances. If you like to visit museums it is likely that you would rather visit museums with your friend than with just anybody who also happens to enjoy visiting museums. A shared interest might very well form the basis of a friendship, but the reason for engaging in joint activities with your friend is partly, or probably even mostly, because you are friends. Telfer points out that friendship not only depends on performing certain actions but the actions being performed for certain specific reasons: out of friendship rather than duty, pity or self-interest (1991: 251). Cocking highlights the relational self that is created by a shared activity. Friends do not just multiply value to the friendship, they generate value through their shared activities. If I have three friends and one of them dies, it is not a case of 33% of the friendship being lost. Neither is it a case of each remaining friend now making up 50% of what is left of the friendship, which would more likely be the case if one member of, say, a group that is formed for the sole purpose of visiting museums decided to leave the group. Rather, in losing friend A, friend B

---

6 Today, this might also involve contact over the internet, or video conferences, to use a more or less antiquated term. The ability to see each other is probably an important feature, since chatting or texting more resembles old-fashioned letter writing.
also loses A’s part in C, while C loses A’s part in B. For that reason it is not a simple matter of replacing a lost friend with another, it is more than mere arithmetics. Friends bring something to the friendship that is unique for them, as persons. It is not about bringing a list of ”love-worthy” properties to the friendship, something that would make the friend replaceable with anyone able to provide the same list. Instead, friends bring their irreplaceable personality into the relationship, and any change in the constituting parts, i.e. the members of the friendship, necessarily changes the nature of the friendship (2013: 2059–2061).

So friends do things together for the main purpose of spending time together. The main purpose of a shared activity with friends is not the activity itself; rather it is the sharing part, the spending-time-together part. Friends have to be persons in their own right with certain qualities that make other people want to spend time together with them, mainly because of who they are. And friends have to be able to feel the same towards other people, a longing to see them not just for pleasure or utility reasons – visiting museums or painting a house, say – but because it feels good to be with them because of who they are.

2.4. Self-knowledge

Being a close friend requires self-knowledge. Friends relate to one another and they share intimacies with one another, something that is impossible if you do not know who you are and, more to the point, that you are. Friendship also promotes self-knowledge. We get to know our friends during the course of the friendship, and in doing so we also come to know ourselves better since it is plausible that we will comment on each other’s behaviour and shared intimacies. Our close friends probably know us better than most and for that reason are able to teach us something about who we are (LaFollette 1996: 88). But self-knowledge is also required during the formation of a friendship. Friendships are preferential, according to Kierkegaard (1991: 240). You do not love your friend because you have to, as is the case with Christian love for humanity, but because you choose to. To just admire someone is neither love nor self-love, but to be admired by the object of one’s admiration creates a relationship

---

7 The example comes from C. S. Lewis’s *The Four Loves*, 1960, referenced by Cocking (2013).
8 Mostly defined as knowledge of one’s sensations, thoughts, beliefs and other mental states. However, it sometimes also refers to knowledge of the self, such as the ability to distinguish oneself from others (Gertler, 2017). That definition is hard to distinguish from self-consciousness, and when I henceforth use the term self-knowledge it is in this broader sense. I will only use the term self-consciousness when there are specific reasons to do so. More on this in part 3.
which turns back selfishly to the “I” that loves; therefore, love and friendship are forms of selfishness where I love my friends because they love me (242). While love for humanity is self-renouncing, preferential love is selfish: I choose the one who chooses me. This would not be possible without an idea of who I am. Telfer agrees that we choose our friends, at least to some degree. The initial spark might be what Telfer calls passion, an unexplained feeling towards the other, something which makes us want to be with that person, but to be passionate about someone is not enough in itself: it is necessary to decide whether to act on those passions. The decision might be more or less conscious, but the existence of the passion and the practice of acting on it have to be acknowledged by both parties. The acknowledgement involves an endorsement of a policy. This is what we do when we commit ourselves to our friends9 (Telfer 1991: 256–257). The friend we choose is not an object to be manipulated but someone who acknowledges and accepts the friendship. This requires a sense of self, of being a rational agent, capable of self-determination (259).

If we, as Kant claims, shall treat our friends as we would ourselves, we need to know how we would like to be treated. If, as Aristotle has it, the good person is to his friend as he is to himself because the friend is another self, he must be aware of himself. If, as Telfer and others propose, we choose our friends, then we have to know what we want. And if I have a moral obligation, some special duties, towards my friends, I need moral agency to be able to fulfil them. If I only pretend or act as if I have moral agency – because I am a psychopath or a programmed robot – my friends are either deceived into thinking that I care for them or they are aware of my lack of moral agency, which gives them no reason to trust that I care for them. Such a relationship, deficient in either honesty or in trust, is a poor excuse for a friendship. For all of these purposes I must have self-knowledge. Without knowing my inclinations and preferences and being able to make an independent choice for my own sake, and without being able to reflect on my own actions and having the capacity to explain and justify them, it would be difficult to form and maintain any kind of friendship: "Therefore to perceive a friend must be in a way to perceive one’s self and to know one’s self” (Aristotle: EE 1245a31–36). Whether it is possible for robots to have self-knowledge will be examined in the following section.

9 On a linguistic note: ‘making’ friends or ‘forming’ a friendship is different from ‘falling’ in love in that it points towards an activity: it takes (sometimes hard) work to build and sustain a friendship, while love is something you can fall in or out of, more or less on a whim. We do not usually make lovers or fall into friendship.
3. Can a Robot have Self-knowledge?

The question of whether robots are conscious or not demands a survey of the concept of consciousness, which would then probably only end in "it depends," the answer depending on what definition of consciousness we choose to apply. A more relevant question for this investigation is if robots can have self-knowledge since self-knowledge is necessary in order to form a friendship with another person. Self-knowledge in the broader sense of both knowing one’s sensations, thoughts, beliefs and other mental states, as well as having knowledge of the self, to distinguish oneself from others, is similar to the definition of self-consciousness as given by Smith (2017). Here, self-consciousness is explained as a condition where subjects are not only aware of what is happening to them, but are aware of themselves as themselves. They know that they are themselves the object of awareness. I see myself in a mirror and I know that it is myself that I see. I see an old photograph and I am not only aware of seeing an old photograph: I am aware of seeing a picture of myself in another place at another time. I am aware of myself as myself. In order not to confuse things too much, I will primarily use the term “self-knowledge,” but I will use it in the broader sense of both having knowledge of my own sensations and of myself as myself.

Can a robot be aware of itself? Hesslow & Jirenhed (2007) argue that it can at least have an inner world. "The mechanism underlying the appearance of an inner world in humans is an ability of our brains to simulate behaviour and perception” they say (85), and such perception has been simulated in a robot, albeit that robot was itself a computer simulation. The Kheperarobot was able to simulate a perception whereby it experienced obstacles in its surrounding even when no such objects were present. By first visually experiencing a room, the robot later used that information to navigate the room using its "mental" image of the room also when it could no longer "see” the room. It also has the capacity to make similar images of rooms it has not yet encountered. The authors argue that this is proof that a robot could have access to an inner world where it creates images of things that do not exist in the exterior world (86–89) but it still remains to be proven if it can be aware of itself. According to O’Regan (2012) there is no logical obstacle to this, even if it is practically difficult. He uses the term “conscious access,” understood as not only being aware of what is happening, which is to have cognitive access, but also being able to reason and communicate about the fact that you have cognitive access. You also need to be aware of the wider context in which this cognitive accessing is taking place. A chess-playing computer with this type of conscious
access would not only be able to play chess with you, but also to comment on your moves, explain the reasons behind its own moves, and strike up a conversation about the weather during a particularly boring part of the game. This is a two-level hierarchy of access, or a Higher Order Thought theory of conscious awareness where you are aware of being in a certain mental state (119). What a robot has today is "self-distinguishing" where it is able to distinguish its body from the outside world to stop it from bumping into other objects. Self-knowledge, which O’Regan interprets in the limited sense of being able to make plans and behave with a purpose, is something we can find in animals such as birds or mice but not yet in robots. The top-level, "knowledge of self-knowledge", is defined as the capacity to empathise with others and to interpret others’ actions in terms of beliefs, desires and motivations, which gives rise to social interaction involving concepts such as shame, embarrassment, pride and contempt (121). Even if O’Regan sees no logical difficulty in implementing this in a robot, it still requires an "I" who is doing the thinking and acting. It has been argued that this I is a social construct, a "narrative fiction." If that were the case the robot would also need to have access to it in order to experience itself as a person in its own right. Nothing implies that the robot could not create a similar construct, but it would have to live and act in an environment where such a fiction is useful, and it would need the capability to generate concepts and abstractions alongside with a body of common sense knowledge which we do not yet know how to implement in robots (121–122, 134).

Floridi (2005) makes the same type of distinction between first- and second-order sense of personal identity and between consciousness and self-consciousness. He calls the first order e-consciousness and it is environmentally oriented: you are not switched off, i.e. you are not asleep, not comatose, hypnotised, drugged and so on, and you are capable of processing information and of interacting with the environment. Animals and humans are usually conscious in this sense but may also be phenomenally conscious, having p-consciousness, which is internally oriented. When you have p-consciousness you experience qualitative, subjective, personal or phenomenological properties of your state of mind. This is what it is like to be someone. Finally s-consciousness, which is second order and also internally oriented, signifies self-consciousness and if you have this you are introspectively aware of both your personal identity and your perceptual or mental experiences: you know that you

---

10 O’Regans ”knowledge of self-knowledge” mostly resembles what others and I refer to as ”self-knowledge”.

11 Since Floridi uses the term self-consciousness, I will likewise use it in my account of his text.
think as well as what you think. A zombie or, for that matter, a robot might be externally oriented and to some extent conscious, in some sense aware of its surroundings, without being p- or s-conscious, i.e. without having a sense of what it is like to be itself and at the same time reflecting on that it is having a sense of what it is like to be itself (416-417). Having s-consciousness is to have what Floridi calls subjective reflectivity, where the agent and the "I" merge and see each other as one and the same. Whereas humans are decoupled from the environment (we are capable of seeing ourselves as entities separate from our surroundings) and coupled to themselves (we are capable of identifying with the image we see in the mirror), robots are decoupled from the agent they can issue reports (crash reports perhaps) about, i.e. from themselves (439).

To have self-knowledge is to be aware of oneself as oneself. I look in a mirror and not only do I see a person, I know that I see myself. I think about the way I look and I am aware of thinking about the way I look. I can also reflect on why I have these thoughts. I understand that other people might have similar thoughts and I can empathise with them. If I criticise how they look I can put myself in their position, reflect on how I would feel if someone criticised my appearance, and feel ashamed about my own behaviour. Robots do not have this type of self-knowledge. They can mimic it: they can be programmed or they can learn by trial and error to act as if they do care, as if they do feel regret and shame, as if they do feel joy. To an extent they can make independent choices such as whether to approach you or not, whether to look happy or not, whether to spend more or less time with you. But your robot lawnmower can also make independent choices, to an extent. Without self-knowledge they are just going through the motions, they are doing what is required of them and not what they want to do, since they do not want anything. For you, if you own a robot, that might be enough. You do not know the limits of the robot’s ability to understand itself so it does not bother you, or you are aware of it but you are fine with it, in which case it does not bother you either. But a form of play-acting is going on which is also a form of deception: you are either tricked into thinking that the robot thinks about you and cares for you, or you have convinced yourself that as long as it looks as if it cares for you that is enough. But it might not be enough. We will look at the problems with deception in the next section.
4. Deception and Other Problems

4.1. Deception

What is deception? A general definition is that you deceive someone if you cause that person to form a false belief. You deceive someone if:

- You do G
- p is false
- X infers that p is true as a result of your doing G

For example: you are bald so you wear a wig; the belief that you have hair is false; I believe that you have hair when I see you in a wig, and therefore I am being deceived by you. In this wide definition, no intention is required. Even if you deceive someone unwittingly, you have still deceived them (Matthias\textsuperscript{12} 2015: 177). Now, social robots are made to look and behave as humanlike as possible in order for humans to be able to interact with them more intuitively and naturally. Proxemics, or distancing behaviour, and gaze behaviour, as mentioned in the introduction, involve making the robots understand human behaviour such as what it means to fold or not to fold your arms, leaning back or leaning forward, touching or not touching, and so forth. To successfully interact with a human, a robot needs to understand what such gestures signal when it perceives them, and how to respond to them. This is partly due to the limited language abilities robots currently possess: it is easier to make a machine understand and mimic body language than speech patterns (Broadbent 2017: 635–636). Engineers make robots look human by imitating human behaviour and in the process influence human behaviour towards robots. That we apply social rules to machines has been shown in experiments, such as Germans rating a robot with a German name as warmer and better-performing than a similar robot with a Turkish name, or rating a less feminine-looking robot more suitable for male tasks. People have also been shown to demonstrate reciprocity such as showing more positive attitudes towards a robot that asked for help and returned the help with a favour (640–641). This amounts to a type of deception where we unknowingly ascribe humanlike qualities to a system that more or less crudely mimics human behaviour. We have a tendency to anthropomorphise, to attribute human characteristics to e.g. an inanimate object, so that when we regard an object that displays human-like behaviour, we tend to also impute

\textsuperscript{12} Matthias borrows the definition from Don Fallis’ "Lying and Deception" (2010).
mind-like qualities to that object since that is what we assume a human being would possess. And since persons in possession of human minds are usually moral agents, it is all too easy to ascribe moral agency to a robot that behaves in humanlike ways whether it has it or not. If it does not have it we should at least be aware of that fact, lest we are being deceived into believing that it does.

So does it? Malle (2016) assumes that it is possible to design robots with a set of elements that make up human moral competence, such as moral vocabulary, a system of norms, moral cognition and affect, moral decision-making and action, and moral communication. Such robots could then ”be trustworthy and productive partners, caretakers, educators, and members of the human community” (243). Here Malle makes a distinction between moral agency, being able to act in accordance of what one believes is right and wrong and to be held responsible for one’s actions, and moral competence, an orienting framework in which we function and which are made up of the aforementioned elements (244). However, Malle’s moral competence seems to leave room for pure simulation: I am morally competent if I act according to a set of rules whether I accept them, understand them, or, even, know that I am abiding by them. Malle does not go this far and argues that the competences should be acquired through learning rather than programming, since it is impossible for a programmer to foresee every possible situation a robot might end up in, and that this involves an element of meta-reasoning. Still, his language points at deceptive behaviour: ”In fact, if the robot can simulate in advance a possible human concern about the robot’s planned action and can conjure up an acceptable explanation, then the action has passed a social criterion for moral behavior” (252). Simulate human concerns and conjure up acceptable explanations, is that not enough? If it looks like a friend, speaks like a friend and behaves like a friend, can we not assume that it is a friend? I believe not. Let’s say a psychopath acted according to the elements, knowing they are working within a framework and following a set of rules, having moral competence but not moral agency, would we accept them as friends or would we think that there is a risk that we are being deceived? If we have any doubts regarding the moral qualities of their behaviour, it would be reasonable to have the same doubts regarding a robot possessing, or lacking, those qualities.

According to Matthias (2015: 178–189) there is a risk that deceptive behaviour both betrays the trust of the deceived and might lead to an erosion of trust in communicative acts. Trust involves a normative expectation where one individual trusts that another will act in a certain
way, for instance by telling the truth. Matthias, whose main focus is on robots in health care, argues that deception does not necessarily betray trust, that it depends on the circumstances. It probably does, but in the circumstances we are considering I take it that deception does indeed betray trust. Can you, for instance, trust a robot who says that it likes you for who you are? If you cannot for the simple reason that the robot does not have the mental capacity of making such a judgment (knowing what it wants from a friend, understanding what you are like, and drawing the relevant conclusions) your trust is betrayed unless you are aware of the limits of the robot’s mental capabilities. I also take it that deception erodes trust in communicative acts. Matthias has some reservations, such as that the trust is not necessarily eroded if the communication with the robot is clearly distinguishable from communication with a human or if the deception is clearly transparent. I find it hard to agree. It seems to me that trust in communicative acts does break down if you cannot be sure about what the robot means when it, for instance, says that it likes you. How does it understand ”like?” Does it even understand ”like” or is it just saying what is expected of it? If you are not aware of its limitations, you are obviously being deceived and your trust is betrayed. But even if you are aware of its limitations, if the deception is clearly transparent, trust is eroded and communication is stunted: what is the point of having an intimate discussion if you know that the robot cannot understand what you are talking about?

In his article “False Friends and False Coinage” (2015), Elder mentions the movie The Truman Show where the main character has lived his entire life in a TV-show, unbeknownst to himself: everyone in his life is an actor and he is the only one who is not in on the game. Imagine, Elder says, that you are given two options. In one, the Truman Show option, all the people you know are paid actors. In the other, the Genuine option, your friends are what they appear to be. Once you have made your choice you will forget ever having made it, and your life will seem entirely normal to you. Which one would you choose? Most of us would go for the Genuine option, Elder claims, because we value more than just appearances and instrumentality in friendship. We do not like to be deceived and we value friends not just for what they seem to be, but also for what they actually are. A real-life example, somewhat similar to the Truman one, is the Undercover Policing Inquiry in the UK.

---

13 This might of course be true in some human-human relationships as well, but the point is that in human-human relationships we assume that our friends have these mental capacities and that they are not lying to us, while in human-robot relationships it is always the case that we are being deceived if the robot has no self-knowledge.
investigating cases where police officers were employed to monitor political protest. In some cases, officers tricked women they were targeting into having sexual relationships with them. A few even ended up living together for years and had children together. When these women, decades later, realised that their partners had been monitoring them all along and in some cases having other, “real,” families on the side, they were distraught by having been deceived, and the fact that the relationships seemed genuine while they lasted mattered less in hindsight than the fact that they were based on a lie (Casciani 2018). Honesty matters to us. We want to know on what ground our friendship is based and for what reasons our friends spend time with us:

"Now if ever someone […] is deceived by the other’s pretending, he will be justified in accusing his deceiver—more justified, even, than if he were facing counterfeitters of the coinage, by the degree to which what is affected by the crime is the more precious.”

(Aristotle: EN 1165b8–12)

4.2. Sharing of Information

An additional trust-related problem is the sharing of intimate information. In the Truman case, every intimate moment of Truman’s life was broadcast to an audience of millions. In the case of the undercover police officers, the betrayal did not only consist in the officers pretending to be something they were not, but also in sharing information on their partners’ behaviour with their superiors. When you have a robot, unless you have built it yourself, which most of us are not able to do, you have probably purchased it from a company (or it has been given to you from the health care services) and whether you come to own it or just lease it there is always a risk that your interactions with it are being recorded in some form or other. Not necessarily for nefarious reasons, it might simply be a case of using the data to enhance the performance of the robot, to help it develop its interpersonal skills better to respond to your needs and wishes. Whatever the reasons, how willing are we to disclose intimate information if we know that we are being recorded, or at least that there is a chance of our conversations being stored for future use? And if we are not aware of it, it is even worse; then we are truly being deceived. Your smart TV can easily be hacked and it is most likely collecting data of your viewing habits (Consumer Reports 2018). In 2015, it was reported that talking doll Cayla, which answers children’s questions by looking up the answers on the Internet, could be hacked and made to say inappropriate things (Moye 2015). Amazon’s voice-activated assistant Alexa has been hacked in order to record everything it can hear (Murnane 2018).
Basically, anything connected to the Internet can be hacked and advanced electronic devices are probably collecting data of user habits. The company that makes the tool might assure you that they are concerned about your privacy, but as the recent Facebook-Cambridge Analytica data scandal proved, where millions of users’ private information was harvested for different purposes, such assurances have to be taken with a pinch of salt (Solon & Laughland 2018). Your private information (conversations, preferences, user habits) might be used to enhance performance, sold to third party companies or accessed by rogue states. Whatever the purpose, if you cannot be sure that what you share with the robot stays between you and the robot, and the chances are that you cannot, there is a lack of trust such as with a friend you know cannot keep a secret.

4.3. Interspecies Relationships

Let us say that we will have robots with self-knowledge in the future; in that case it looks like it would be possible to make friends with them. However, we have no reason to believe that the robots will think and reason like we do. In his article “What Is It Like to Be a Bat?” (1974), Nagel says that if you have conscious mental states there is something that it is like to be what you are. If a bat has consciousness, there is something that it is like to be a bat. But we, as humans, can never know what it is like to be a bat. We can try to imagine what it is like but we are always restricted to the resources of our own minds – I can only imagine what it would be like for me as a human to behave as a bat, which is not what it is like to be a bat (439). To be able to have a reasonable objective assumption of another’s experience, we have to be sufficiently similar to that individual in order to be able to adopt their point of view (442). As Wittgenstein remarked, if a lion could speak we would not understand it. Since we are likely to anthropomorphise robots, we assume that they have human or at least human-like minds, but why would a robot with self-knowledge think like a human? That robot is not programmed to think and act in a certain way; it has an independent mind, free to make its own choices, free to think for itself. As Lawrence et al point out (2016), we assume that they are motivated by the same kind of considerations which motivate us (252). But why should self-fulfilment for a robot be the same as self-fulfilment for a human? Beyond survival we have little way of knowing what actually motivates them. Even if they come to live by the principle of reciprocity, or by the Golden Rule (do unto others as you would have them do unto you), we do not know what constitutes benefit and harm, or hurting and healing, for them (256–258). Who says that they are even interested in having friends, or if they are, why would they choose human friends over robot friends who have a greater chance of understanding
them and with whom they are more likely to form a close relationship? How can I be friends with a robot if I cannot understand how it thinks and feels, and how can a robot be friends with me if it cannot understand how I think and feel? We might be able to have a relationship of sorts, but real intimacy would be difficult without this kind of understanding.
5. Conclusion

The robotics industry is growing rapidly (Tobe 2017). There are high hopes of more advanced robots soon being able to care for our elderly and keep us company, as well as sex robots doing the work that prostitutes and victims of trafficking perform today. There is also the usual fears of our new "robot overlords” taking over, fuelled by the media industry producing motion pictures and TV series such as *Terminator*, *Ex Machina* and *Westworld*. When, and it is probably a question of when rather than if, such advanced robots are provided by the health care services and available to purchase on the open market, we will very likely spend a considerate amount of time in close proximity to them and it is relevant to ask what type of relationship we will be able to have. For that reason I decided to ask whether it is possible to be friends with a robot. To try and find an answer to that question, I first looked at friendship. I focused on a moral point of view of friendship, that we have special duties towards each other, as well as two recurrent themes that recur when philosophers write about friendship: mutuality and shared activities. I also pointed out that this requires self-knowledge.

In short, when I have friends I have responsibilities towards my friends. I look after them because they are my friends, not because they are more important than other people are but because they are more important *to me* than other people are. This is a moral duty: just as parents have a moral duty to look after their offspring, friends have a moral duty to look after each other. Brody (2014) stops short at saying that his argument gives us permission to take special measures to care for those closest to us, but I claim that it is stronger than that: without turning this into an obligation the friendship would lose in worth. If friends only have competing claims and not prior claims on us, as Telfer argues (1991), the friendship would lose in worth. If I found out that my friend decided to help a complete stranger move house instead of helping me move house, I would question the value of our friendship. If my work colleague decided to have a drink with their best friend instead of having a drink with me, I would find that reasonable. We help our friends because they are our friends and we do things together with our friends because they are our friends. Without mutuality there is no friendship since, as Helm writes (2017), unrequited friendship is senseless. If two individuals do not accept each other as friends, there is no friendship. Reciprocated love, Kant’s ”amor bilateralis,” is necessary for friendship: if I want to be friends with a robot, both the robot and I have to be able to provide it and we cannot provide it unwittingly. Since, as the saying goes, it takes two to tango, it is not enough if I love the robot as a friend. For friendship to arise the
robot also has to love me, as a friend. If it does not, because it cannot, then there is no friendship.

As regards shared activity, in order to keep the friendship alive we have to do things together, but friends do things together not just for the sake of doing them: it is the together part which is crucial rather than the doing part. I want to do things with my friends because they are my friends, and I might even do things I do not particularly like, because my friends like them. I would hardly engage in activities I find boring or see no point in together with complete strangers, but I could very well do them with my friends, for their sake rather than for mine. In this sense, friends are not replaceable with each other. They do not just bring a list of love-worthy properties to the friendship; they bring themselves, their personality – which of course requires of them to have one in the first place. But it is still a two-way street: they also do things with me and for me because they love me for who I am. None of this would not be possible without self-knowledge, if I was not a person with thoughts, beliefs and motivations, and if I did not know that I was a person with thoughts, beliefs and motivations. I could love a robot that does not have a sense of self, but a robot without a sense of self could not love me. And without reciprocated love, there is no friendship.

A robot might seem to be able to do all that is required from a friend, but it would be a kind of deception (and deception does not have to be intentional – if you unwittingly deceive someone, you have still deceived them). The robot would only seem to care about you, much as, say, a psychopath could seem to care about you even if they did not. Unintentionally, since it has no intentions, it displays all the signs of caring and affection, but this is a complete lack of caring masked by a pretence that it does indeed care. If you are not aware that this is happening you are being deceived into thinking that the two of you are friends. If you know that you are not friends but decide to play along, if you suspend your disbelief, you are willingly engaging in something which has the signs of friendship but which is not friendship. This robot has no self-knowledge and therefore no moral agency, only possibly moral competence. But let’s say that we will have advanced robots with self-knowledge sometime in the future; then it seems as if we would be able to be friends. But would we? It is far from certain that they would want to be our friends. And if they would want to, since robots are not humans and humans are not robots, the difference between the two ”species” makes it uncertain that we will be able to understand what it is like to be the other, in which case it is questionable whether a friendship is at all possible.
As Harvey puts it in his colourful language, “once our machines look like us, talk like us, move like us, smell, feel and taste like us, and then, in addition, fuck us silly as well, it will prove difficult indeed to resist falling in love with our robots” (2015: 91). Once our robots display all the traits of a true friend we will probably not be able to resist forming what to us looks like friendship – we could make robots who seem to understand us perfectly, who never let us down, who are always there for us. Who would not want that? But if it is only an act, if it is merely a machine, an automaton that is not aware of what it is doing, there is no friendship. And a robot with a mind of its own, who knows what it wants? What says it would want to be friends with us? And even if it did, if it cannot understand me and I cannot understand it because we are too different – the distance is too big to adopt the other’s point of view – then friendship is at least difficult to form if not entirely impossible. How are you supposed to be intimate with someone you do not understand and who cannot understand you? So the answer is most likely no, I cannot have a robot friend.


Sharp, R. ”The obstacles against reaching the highest level of Aristotelian friendship online”. Ethics and Information Technology, 14:231–239, 2012

