Robot Betrayal
A Guide to the Ethics of Robotic Deception

John Danaher, NUI Galway

(Forthcoming in *Ethics and Information Technology*)

Abstract: If a robot sends a deceptive signal to a human user, is this always and everywhere an unethical act, or might it sometimes be ethically desirable? Building upon previous work in robot ethics, this article tries to clarify and refine our understanding of the ethics of robotic deception. It does so by making three arguments. First, it argues that we need to distinguish between three main forms of robotic deception (*external state deception*; *superficial state deception*; and *hidden state deception*) in order to think clearly about its ethics. Second, it argues that the second type of deception – superficial state deception – is not best thought of as a form of deception, even though it is frequently criticised as such. And third, it argues that the third type of deception is best understood as a form of betrayal because doing so captures the unique ethical harm to which it gives rise, and justifies special ethical protections against its use.

1. Introduction

What should we do about deceptive robots? Since Alan Turing first defined the field of AI research in terms of the Imitation Game, there has been an uneasy relationship between reality and fakery in the design and operation of robots (Turing 1950; Gunkel 2018, 145-150)**.** This has become particularly problematic in the field of social robotics, where the use of anthropomorphic cues — up to and including deceptive cues — is thought to be integral to the project of building a socially acceptable robot (Wagner and Arkin 2011; Damiano and Dumouchel 2018; Isaac and Bridewell 2017). This has put some ethicists and social commentators on edge as they worry about the ethics of such robotic deception and implore engineers and policy makers to ensure greater transparency in the design and operation of social robots (Kaminsky et al 2017; Leong and Selinger 2019; Turkle 2007 & 2010)**.**

There is, however, some confusion pervading this debate. There is a tendency to conflate the different kinds of deception and fakery that can arise, and to rush to judgment in both condemning and rationalising robotic deception. The purpose of this article is clear up some of this confusion and provide a more nuanced understanding of the ethics of robotic deception.

It does so by making three main arguments. First, it argues that although there are many different forms that robotic deception can take, there are three general categories that must be distinguished in order to think clearly about its ethics: (i) external state deception; (ii) superficial state deception and (iii) hidden state deception. The first is practically significant but relatively uninteresting from a philosophical and ethical perspective; the other two are more philosophically significant. Second, it argues that superficial state deception is best interpreted through the lens of a philosophical theory — here called "ethical behaviourism" (Danaher 2019) — which significantly alters how we ought to think about it — specifically by implying that it is not really an ethically worrisome form of deception in most cases. And third, it argues that hidden state deception is the most ethically worrisome and is best understood as a form of betrayal. Indeed, applying the concept of betrayal to this category of deception not only offers the best insight into the unique harm it can cause to human-robot relations, but also helps to make a compelling case for special protections to be put in place to prevent it from arising.

The article proceeds in four main stages. In the next section, I briefly review some of the existing contributions to the ethical debate about robotic deception. I follow this by defending the first argument regarding the three main categories of robotic deception. I then make the case for applying an "ethical behaviourist" lens to superficial state deception, before finally

concluding by arguing that hidden state deception should be viewed as a form of betrayal. The end result is a 'user's guide' to robotic deception, one that clarifies how best to think about it and signposts ethically sound practice in relation to the design and operation of deceptive robots.

2. Who's afraid of robotic deception?

Let's start by getting the lay of the land and by defining the concept of a robot. A robot can be defined as an embodied artificial agent, i.e. an artificial, computer-programmed, entity with the capacity to sense, process and act upon its surroundings (Gunkel 2019, Ch 1). So defined, robots are becoming more prevalent in our societies. They are being used to perform ethically high stakes actions in medical, military and automotive context, as well as more mundane actions in our homes and offices.

I mentioned in the introduction that people are concerned about deception performed by such robots, but why so? The concept of deception will be defined in greater detail in the next section but, as a rough first pass, we can say that deception involves the use of signals or representations to convey a misleading or false impression. Usually the deception serves some purpose other than truth, one that is typically to the advantage of the deceiver and the disadvantage of the deceived, though sometimes deception can be mutually advantageous, as in the case of so-called "white lies". Following this definition, we can say that robotic deception arises whenever a robot, as an embodied artificial agent, makes a representation or sends a signal that creates a misleading or false impression among those who interact with the robot. The signal/representation may be in the form of speech, behaviour or physical appearance. There are three general trends one can observe in the literature about such robotic deception to date.

First, there are those who think that some robotic deception is tolerable, perhaps even necessary (Wagner and Arkin 2011; Wagner 2016; Shim and Arkin 2016; Isaac and Bridewell 2017**)**. This is not surprising since it is the view that most people have of human deception. As a rule of thumb, we think that it is wrong to mislead others, but we also recognise that deception can serve a higher purpose that ethically justifies its use. For example, someone who lied to the Nazi authorities during WWII, to conceal the fact that they were harbouring Jews in their basement, would be viewed by most people as a moral hero, not a moral villain. If we want our robots to be ethical, it's plausible to argue that we should both expect and demand that they do the same. In addition to this, psychologists and evolutionary biologists have long recognised that deception and concealment is at the core of social intelligence (Trivers 2011; Simler and Hanson 2018). Getting along with your peers sometimes demands that you are frugal with your true opinions, that you flatter and conceal what you really think. This helps build alliances and enable coordination. Honesty can be a virtue but not if it is unrelenting. Building upon this observation, several roboticists have argued that if we want to build social robots that are capable of integrating smoothly into human society (and that's an "if" that we may wish to reconsider) we will have to equip them with some deceptive capacity (Wagner and Arkin 2011; Wagner 2016; Shim and Arkin 2016; Shaw 2015**).** This position has been endorsed by the robot ethicists Alistair Isaac and Will Bridewell (2017) who defend the view that robotic deception is acceptable whenever it serves some greater good, including the good of smooth social integration. All of these authors can draw support from recent research about the moral expectations people have about robotic agents. Studies by Malle et al (2015) and Voilkis et al (2016) have found that people do apply moral norms and rules to robotic agents, but that they tend to apply a more consequentialist (as opposed to deontological) set of expectations to robots than fellow humans. This could

mean that people will expect (and perhaps even demand) that robots lie if this serves the greater good.

Second, there are those who are more sceptical and think that at least some forms of robotic deception need to be stamped out. The fears here primarily relate to the use of "dishonest anthropomorphism" in the design and operation of robots (Kaminsky et al 2017; Leong and Selinger 2019; Zawieska 2016). This refers to the tendency for robots to use anthropomorphic appearance and behaviour to 'trick' people into believing that they are human-like. Sherry Turkle (2007; 2010) has voiced particular concerns about this. She argues that simulated affect in social robots — e.g. robots that cry, laugh or express concern for their users — is ethically dubious because it tricks people into thinking that there is some mutuality in the relationship they have with a robot when there really isn't. She suggests that all intimate or caring relationships with robots that are initiated on the basis of such signals are illusory, and that while simulated thinking might be a genuine form of thinking, "simulated feeling is never feeling, simulated love is never love" (Turkle 2010, 4). Others have voiced similar concerns about the dangers of fake or counterfeit relationships between humans and robots (Elder 2015 & 2017; Sharkey and Sharkey 2010).

Closely related to this is the fear articulated by Kaminsky et al (2017) about the use of anthropomorphic cues to conceal non-anthropomorphic capacities. To illustrate the fear, they give the example of a social robot with a pair of eyes that appears to look away from a user while actually using a concealed video camera to record what the user is doing. Brenda Leong and Evan Selinger (2019) build upon this and develop a detailed taxonomy of the different ways in which anthropomorphic cues can give misleading impressions of what a robot is really up to. They worry that such dishonest anthropomorphism can be leveraged by

malicious actors to surveil and manipulate humans in undesirable ways. In support of this, they point to studies from cognitive and social psychology that highlight how humans are automatically triggered into anthropomorphic modes of thought by the presence of human-like features (a point also emphasised in Zawieska 2015). When in this automatic mode of thinking, humans have certain expectations about what their interlocutors are going to do. Robots can easily violate these expectations because of their concealed non-human-like capacities (e.g. the capacity to 'see' in infra-red). This makes humans more vulnerable and less able to safeguard themselves from malicious forms of deception. Consequently ethical rules and policy guidelines need to be crafted to protect humans from dishonest anthropomorphism. In particular, Leong and Selinger advocate for greater transparency in the design and operation of social robots, suggesting that robots should clearly signal to their human users what they are doing and how they may not be truly human like in their functioning. In this regard, Leong and Selinger tap into the widely expressed desire for greater transparency in how robots and AI more generally operates. This desire perhaps finds its most authoritative expression in the EU High Level Expert group's principles for creating "trustworthy" AI (2019**)**. It should be noted, however, that the desire for greater transparency is not only expressed in response to deception as it is defined here; sometimes transparency is desirable even when there is no representation or signal that creates a misleading or false impression, as in the case of discrimination or bias in robotic decision-making.

Third, and finally, there are those who take a different view of dishonest anthropomorphism, suggesting that although there are ways in which it can be misused, it is important not to over-ascribe dishonesty to the use of anthropomorphic cues. This is an important view since, as noted in the introduction, 'imitation' is built into the foundational fabric of robotics. Luisa Damiano and Paul Dumouchel have defended one of the more

nuanced views on this matter (2018). In basic outline, Damiano and Dumouchel maintain that the criticisms of deceptive anthropomorphic signals in robots often rest on an misguided dualistic view of human mental capacities. In other words, they argue that the assumption made by critics like Turkle, is that in order for anthropomorphism to be *honest* the robot must have some inner mental state/capacity that matches its outward anthropomorphic cues. If a robot lacks the requisite inner mental state, then it is being dishonest. But this dualistic view is outdated and disconfirmed by embodied and relational theories of cognition. Following these theories, Damiano and Dumouchel argue that human emotions and affect are best understood as phenomena that facilitate coordinating relations between two or more agents. The anthropomorphic cues of emotion and affect provide information that enable the parties to effectively coordinate their behaviour. The "truth" or "falsity" of these cues depends on their pragmatic value within that coordinative relationship, not on the presence or absence of some inner mental state.

Damiano and Dumouchel use this view to invert the logic of Leong and Selinger's fears about the use of deceptive anthropomorphic cues. Far from worrying about the automatic mode of thinking and how it can be exploited, Damiano and Dumouchel argue that this automatic mode of thinking gives us reason to believe that pragmatic reliance on those cues is not dependent on false beliefs, and hence not really the basis for a form of deception. As they put it:

"[Anthropmorphic cues] trigger immediate emotional reactions that do not need, or rest on, the complex process of interpretation which philosophy, psychology, and classic cognitive sciences postulate as necessary for a person to access others' emotions. This affective

coordination bypasses theory of mind and folk psychology. Applied anthropomorphism does not require any false beliefs."

(Damiano and Dumouchel 2018, 7)

Whether this is persuasive or not is something to which we shall return. For now, it is worth noting that this pragmatic approach to anthropomorphism quells some of the fears about dishonesty and deception, but not all. It implies that some uses of anthropomorphism are not really deceptive or dishonest, but others could still be, including the use of multiple cues to conceal ulterior motives, or the use of signals that are clearly and unambiguously false.

In summary, there are three main trends to be observed in the current ethical debate about robotic deception. The first suggests that robotic deception can be ethically acceptable, possibly even necessary to the smooth social integration of robots. The second suggests that we should be very concerned about some forms of robotic deception, particularly, the dishonest use of anthropomorphic cues, and that we need to put in place safeguards to ensure honesty and transparecny in relation to the use of such cues. And the third, pushes back against this by arguing that some allegedly dishonest uses of anthropomorphic cues are not really dishonest at all.

How are we to make sense of this? Who is right and who is wrong? To start answering those questions we need to clarify the concept of robotic deception.

3. Three Forms of Robotic Deception

Deception is a complex phenomenon. In my initial gloss, I suggested that deception arises whenever signals or representations are used to create a false or misleading impression. This suffices for a first pass but the phenomenon is more complex than that and deserves greater scrutiny.

Clarifying exactly what it means to say that someone is engaging in a deceptive act has been a professional preoccupation for generations of philosophers (Mahon 2015**).** Within philosophical circles, most of the debate centres on a definition of deception that requires several conditions to be satisfied simultaneously in order for an act to count as deceptive. Some of these conditions relate to the intentions, beliefs and desires of the deceiver, and some to the state of mind of the deceived, as well as the context of the deceptive statement. These definitions are useful but confusing when applied to robots. One reason for this is that one of the things that robots are alleged to be deceptive about is whether or not they have intentions, beliefs and desires. This suggests that if we want to understand robotic deception we'll need to have the flexibility to go beyond the standard philosophical accounts.[1]

A good place to start is with the account of robotic deception developed by Isaac and Bridewell (2017). Their account focuses on overt speech acts — statements, claims, directives and so on — that might be issued by a robot in conversation with a human. To illustrate, you might imagine the statements made by AI companions when they are prompted for responses by their human users through a natural language interface: "Hey Siri, what's the weather like today?", "Are there any good movies in the theatre at the moment?" and so

[1] The philosophical obsession with consciously intended deception has been criticized by others. Robert Trivers, in his natural history of deception, points out that "If by deception we mean only consciously propagated deception—outright lies—then we miss the much larger category of unconscious deception, including active self-deception" (Trivers 2011, 3). This seems right and is the more appropriate approach to take when looking at robotic deception.

forth. The statement made in response is the speech act that is capable of deceiving the user by creating a false or misleading impression.

But how does this happen? Isaac and Bridewell argue that conversations come with *standing norms*, i.e. a set of common expectations about the purpose of the conversation and the rules that the conversational partners have to follow when participating in that conversation. The linguistic philosopher Paul Grice developed one of the best-known accounts of such standing norms with his theory of conversational implicature and its associated maxims (Grice 1975**).** According to Grice, most human-to-human conversations are guided by maxims of *quantity*, *quality*, *relation* and *manner*. That is to say, speakers ordinarily try to be succinct and informative, to stay on topic, and to avoid obscurity or irrelevance. All of these maxims serve the goal of truth telling. But conversations can be guided by other norms as well. In some contexts people want to be polite or build rapport. In those contexts, truth-telling might be subservient to other goals. This doesn't mean that speakers are free to tell outright falsehoods, but they can be less focused on truth-telling than might otherwise be the case. Isaac and Bridewell give the example of a friend who asks you what you thought of their poem, to which you reply "you must have put a lot of work into it" (2018, 162). This is not exactly false, but not exactly true either. It is an evasion but one that is, according to Isaac and Bridewell within the "expected standards of conversation" and maintains the rapport with your friend (2017, 163).

The critical point for Isaac and Bridewell is that deception arises when a speaker makes an utterance that violates the standing norms of the relevant conversation, and serves some other agenda in the process (i.e. some ulterior end). Since most conversations are governed by norms of truth-telling (as per Grice), the most obvious cases of deception involve

statements that are false, but this is not the only form that deception can take. Isaac and Bridewell give the examples of *paltering* (distracting someone by uttering an irrelevant truth), *bullshitting* (demonstrating a reckless disregard for the truth) and *pandering* (bullshitting in order to flatter a specific audience) as other forms of deception.

Whatever the truth value of the utterance in question, it will not count as a form of deception, according to Isaac and Bridewell, unless it serves some ulterior end. It is the presence of the ulterior end that turns the statement from an innocuous violation of the norms of conversation into something more ethically troubling. Nevertheless, and as should be clear from the previous section, Isaac and Bridewell are adamant that not all instances of deception are unethical: sometimes the ulterior end is ethically preferable and trumps the standing norms. This is why they think we should want robots to deceive us, at least in some cases.

Isaac and Bridewell's account is useful but incomplete. By focusing explicitly on speech acts, they overlook the dishonest anthropomorphism that some people find so disturbing. This form of deception involves the use of anthropomorphic cues (appearances and behaviours) to distract and mislead humans regarding the true nature and purpose of the robot. Prima facie, this is very similar to what Isaac and Bridewell are talking when they talk about deception since it involves a violation of the standing norms and expectations associated with anthropomorphism, and so it should be included in the discussion. It could well be, of course, that Isaac and Bridewell's speech act-oriented theory can be extended to cover behavioural signals and body language — sometimes people talk about speech acts in these expansive terms — but it is best to be explicit about their inclusion rather than leaving it open to interpretation. That's what I propose to do here.

The other feature of their account that requires some further discussion is the nature of the ulterior end that the deception serves. Isaac and Bridewell refer to these ulterior ends as 'goals' and 'motives'. One might think this is problematically mentalistic in the case of robots: can robots have such things? Isn't that effectively what Turkle and others deny? Maybe, but there is nothing particularly outlandish or philosophically suspect in suggesting that a robot has a goal state that it is trying to achieve. Even the simplest algorithm can be described in such terms: it will have some output that it 'aims' to produce by following a series of steps. The more important question is: whose goals are these, really? It is possible, in the case of highly sophisticated and autonomous robots, that the goals belong to the robot themselves. A sophisticated robot may, for example, learn deceptive instrumental goals on its route to achieving some more general final goal — this is, indeed, a common story-telling motif in science fiction and a concern among those who worry about AI risk (Haggstrom 2019; Bostrom 2014; Omuhundro 2008). In other cases, however, it is quite likely that the ulterior ends can be traced back to the goals of the original designers and manufacturers of the robot. Indeed, the presence of such third party deceptive goals seems to be the major concern that Kaminsky et al (2017) and Leong and Selinger (2019) have in mind when they talk about dishonest anthropomorphism.

With this in mind, I propose the following account of robotic deception:

Robotic Deception: Arises whenever a robot (a) uses some signal (speech act; anthropomorphic cue) in a way that (b) violates the expectations/norms we usually associate with the use of such a signals (most commonly by using the signal in a way that is objectively false or misleading), where (c) this serves some ulterior end that can either be traced to the robot themselves or some third party.

The first two conditions in this definition relate to the *content* of the signal (i.e. what it is taken to be about) and its relationship to the broader normative context associated with such signals. The third condition relates to the purpose it serves. In combination, they give us a framework for thinking about the different varieties of robotic deception. In principle, there are many such varieties — indeed, the varieties of robotic deception could be as vast (if not vaster) than the varieties of human deception. Nevertheless, there are a few high-level categories of robotic deception that seem to be worth distinguishing for ethical purposes.

These high-level categories emerge from the different possible *contents* of robotic deception — i.e. what the deceptive signals issued by the robots are about. Using the human case as an analogy, there would seem to be two things that a deceptive signal could be about. It could be about some state of affairs in the world that is external to the agent using the signal, e.g. a statement of historical fact or geographical fact. Or it could be about some feature of the agent themselves, e.g. a signal about their intentions, desires, capacities or identity. The latter category corresponds, roughly, to the form of deception previously described as "dishonest anthropomorphism" though it is broader and could include signals that are not obviously anthropomorphic. Looking at this second category in more detail it would seem that, in the case of robots, there are a couple of different forms it could take. A robot could, for example, signal that it has some capacity or internal state of mind that it actually lacks (this is what concerns Sherry Turkle about dishonest anthropomorphism) or it could be concealing the presence of some capacity or state that it actually has (this is what concerns Kaminsky et al and Leong and Selinger). This leads to the suggestion that there are three high-level forms that robotic deception can take:

External state deception: the robot uses a deceptive signal regarding some state of affairs in the world external to the robot.

Superficial state deception: the robot uses a deceptive signal to suggest that it has some capacity or internal state that it actually lacks.

Hidden state deception: the robot uses a deceptive signal to conceal or obscure the presence of some capacity or internal state that it actually has.

External state deception is practically important. A medical diagnostics robot, for example, that gave a misleading impression of your health and well-being in order to tempt you into unnecessary medical treatment would be highly problematic. Nevertheless, there is nothing particularly philosophically or ethically unique about this form of robotic deception. Presumably, whatever ethical rules apply to humans who engage in external state deception should apply to robots as well. If it would be wrong for a human to deceive another human about the relevant external state, then it would be wrong for a robot to do the same. If it would be right for a human to do so, then it is not clear why a robot should be held to a different standard. This is, I suspect, true even if people have different moral expectations of robots, as the study from Malle et al (2015) suggests: the moral standards that apply to deception shouldn't vary just because a robotic agent is the mediator of the deception. That said, issues could arise concerning who is responsible for the relevant deception (who determined the ulterior end?), and I will reconsider the claim that robots should not be held to higher standard again toward the end of the article after I have defended the ideal of robotic betrayal.

The other two categories of robotic deception are more philosophically and ethically interesting. They are both more unique and troubling in the robotic case because of the disputes about the ontological status and nature of robots. These forms of deception seem to be what is most disturbing and unsettling to critics. These critics worry that robots violate the expectations and norms associated with human-like entities and thus can be exploited by malicious actors. These two forms of deception are, however, distinct and there has been an unfortunate tendency to conflate them in the debate to date (e.g. Leong and Selinger 2019, 304**)**. This is apparent in the tendency to lump them together into the general category of dishonest anthropomorphism. This is unfortunate because it is possible for a robot to engage in superficial state deception without engaging in hidden state deception, and vice versa. A robot might look away from you while using a hidden video camera or voice recorder, but this does not mean that its 'eyes' are not also capable of seeing you (i.e. it could have hidden state deception without superficial state deception). Similarly a robot could appear to look at you and yet not possess the capacity to 'see' anything (i.e. it could have superficial state deception without hidden state deception). It is also unfortunate because, as I will argue below, superficial state deception is not, in many cases, an ethically disturbing form of deception. It consequently does a disservice to the harmfulness of hidden state deception to treat them as equivalent.

In addition to varying in terms of the content of the deceptive signals, robotic deception can also vary with respect to the ulterior ends it serves. There are many possible ends it could serve. Concealed spying that is facilitated through hidden state deception could serve the needs of corporations for marketing information or the needs of governments keen on identifying and preempting terrorism. It could also serve the needs of the robot itself, keen on masking its true intentions from its human interlocutors (cf Bostrom 2014 on the idea of the

"treacherous turn"). It is difficult to say anything abstract and useful about the high-level categories here, apart from the banal observation that some purposes will be benign, perhaps even beneficial, and others less so.

All I will say is that, given the entangled relationship between robot goals and the goals of third parties that control and design the robot, it is always important to ask what might be called the 'Cyrano de Bergerac' question about robotic deception. The question is named in after the play by Edmond Rostand. In this play, the title character (Cyrano) helps a more attractive younger colleague (Christian) to woo a woman (Roxane) with whom he (Cyrano) is in love. He does this by writing letters on behalf of Christian and telling him what to do. The net result is that Roxane falls in love with a fiction, a version of Christian that does not really exist, and is a front for the intentions and desires of Cyrano. The ruse is only made known to Roxane when it is too late. There is a tragedy in this since she was in love with the person who wrote the letters, not the person who signalled them to her. She would have loved Cyrano if only he had not mediated his intentions and desires through the medium of Christian. The Cyrano de Bergerac question then is: whose motives do the robot's signals really serve? Are we dealing with a *benign-Cyrano*, who we can trust and accept, or a *malicious-Cyrano*, who we cannot? That will be crucial to assessing the ethical status of all forms of robotic deception.

4. An Ethical Behaviourist Approach to Superficial State Deception

Let's now focus on the specific sub-categories, starting with superficial state deception. The view I defend here is that this is typically not an ethically disturbing form of deception. Indeed, in most cases, it is not a form of deception at all. Superficial states can be incomplete or inconsistent, but not directly deceptive in and of themselves.

Reaching this conclusion is dependent on first accepting a theory concerning how we ought to interpret the superficial states of a robot, at least when we interpret these states from an ethical perspective. This theory can be called 'ethical behaviourism'.[2] According to this theory, the ethical status of our interactions with a robot can be determined by their external behavioural states and cues only, and not by anything else. This is an epistemic thesis concerning the warrant for our ethical beliefs, not a metaphysical or ontological thesis concerning the ultimate grounding for those beliefs. According to ethical behaviourism, if a robot appears to have certain capacity (or intention or emotion) as a result of its superficial behaviour and appearances, then you are warranted (possibly mandated) in believing that this capacity is genuine. In other words, if a robot appears to love you, or care for you, or have certain intentions towards you, you ought, ceteris paribus, to respond as if this is genuinely the case. There is no inner state that you need to seek to confirm this. This means that, contrary to what Sherry Turkle and like-minded critics might suppose, simulated feeling can be genuine feeling, not fake or dishonest feeling. Consequently, if ethical behaviourism is true, then superficial state deception is not, properly speaking, a form of deception at all. This is because superficial states provide the best epistemic warrant for believing in the presence of the relevant mental states or capacities, at least for ethical purposes.

Why should someone accept ethical behaviourism? The simple answer is that behavioural cues are the most compelling evidence we have for the reality of certain capacities or inner states. There are two main reasons for this. The first is that the primary source of evidence we have for such things is our own first person experience, but this is inherently private and

[2] This paper is not the first to defend this idea, though the theory has not always been explicitly named as it is in the main text. For similar arguments, see Neely 2014, Schwitzgebel and Garza 2015, and Danaher 2019a and 2019b. Each of these authors suggests, either directly or indirectly, that the superficial signals of a robot should be taken seriously from an ethical perspective, at least under certain conditions.

problematic when it comes to assessing the ethical status of our interactions with others. I know my feelings for my partner are genuine because I personally *feel* them — I know the longing and desire I feel for her — but how does she know about them? She cannot directly access my inner mental states. She can only go by my external behaviour. If that behaviour is not consistent with my professed longing and desire, she will be epistemically warranted in believing that I am not being genuine. She has to use that behaviour to reach an informed conclusion about the ethical status of our interactions. This is, of course, a familiar thought. It is at the heart of Turing's famous test for machine intelligence. Ethical behaviourism is simply an application of this to the ethical domain.

The second reason for endorsing ethical behaviourism is that other alleged evidential bases for assessing the ethical status of our relationships do not override or undermine the behavioural ones. For example, some people might argue that we should use the biological constitution of an entity to determine the ethical status of our interactions with that entity. If the entity is made from biological tissues and organs, then its behavioural cues have an ethical significance they would otherwise lack: a whimpering dog is genuinely expressing its feelings; a whimpering robot is not. But really there is no reason to think that evidence concerning biological constitution should trump or undermine behavioural evidence, at least if that behavioural evidence is consistent and complete. In other words, if a robot consistently acts in a way that suggests its whimpering is genuine then there is no reason to deny it an ethical significance that is granted to the dog, apart from an unjustifiable fealty to biology. In ethics, we have to err on the side of caution, of over-inclusivity not under-inclusivity, when it comes to determining to whom we owe duties and we ought to do (Neely 2014; Sebo 2018).

Similarly, some people might argue that the presence of a brain, with certain functional patterns of activity, is essential to determining the ethical status of superficial cues. If an entity lacks a brain, then its superficial cues, no matter how sophisticated, lack an ethical significance they would otherwise have. But, again, there is no reason to think that the presence of a brain should play such a decisive role in our ethical assessments. For one thing, there is at least some ambiguity here as to what would count as a brain for ethical purposes. Must an ethically significant brain be made of neurons and glial cells? If so, then we run into the previous problem concerning biological constitution. If not — if all it must do is be functionally equivalent to a biological brain — then we run into another familiar problem: functional brains look to be multiply realisable and could be present in a robot.

This, however, does not suffice to defend ethical behaviourism. After all, it could be that some people are convinced that the presence of a multiply-realisable, functionally-equivalent brain is necessary for us to conclude that a superficial state is a genuine expression of mental capacity. The response to this is that it is not clear why the presence or absence of a functional equivalent to the brain should play an ethically decisive role. Consider the case of a hydroencephalic person who lacks much of the brain tissue of an ordinary human being. If we look inside this person's head, we can immediately confirm this. But, from a behavioural perspective, this person is essentially no different (or not substantially different) from a 'normal' adult human. Such people really exist. Do we deny the ethical significance of their behavioural cues because they lack a lot of the brain tissue a 'normal' human has? Presumably not — presumably we think their signalled feelings and intentions are every bit as honest and genuine as those of other human beings. But now imagine if it turned out that they lacked brain tissue altogether and yet still looked and acted like a normal adult. What would happen then? You might say such a person would be metaphysically impossible, and

maybe that is the case, but they do appear to be epistemically conceivable and when we conceive of them it is difficult to see why we would deny ethical significance to their behaviours purely because they lacked a brain. This observation carries some epistemic weight, even if it is metaphysically uncertain because given this metaphysical uncertainty as to whether other features (e.g. brains and biology) really matter from an ethical perspective, there is no reason not to treat behavioural cues as the most compelling basis for determining the ethical significance of our interactions with others.

It is important that this position is not misunderstood. Ethical behaviourism is an epistemic thesis, not a metaphysical one. Its claim is not that capacities and mental states can be ontologically reduced to behavioural cues (a position once defended by logical behaviourists in the philosophy of mind — see Graham 2015). Ethical behaviourism is strictly agnostic on the broader metaphysical questions: it just claims that when it comes to determining the genuineness of capacities and mental states, superficial cues are the most compelling evidence we have.[3]

This also does not mean that it is impossible to be misled or get the wrong impression of what a robot is capable of doing on the basis of superficial states. It most certainly is. Painting a pair of eyes onto a robot should not convince you that the robot can 'see' you. A more thorough investigation of its behavioural repertoire (a larger set of superficial cues) will

[3] Ethical behaviourism is also consistent with, but distinct from, the science of machine behaviour that Rahwan et al (2019) advocate. Rahwan et al's article make a plea for scientists to study how machines behave and interact human beings using the tools of behavioural science. In making this plea, they highlight a tendency to focus too much on the engineering details (how the robot/AI was designed and programmed) in the current literature. The result of this is that an important aspect of how machines work and the behavioural patterns they exhibit is being overlooked. I fully support their programme and the stance advocated in this article agrees with them insofar as (a) the behavioural perspective on robots does seem to be overlooked or downplayed in the current debate and (b) there is a significance to machine behaviour that is independent from the mechanical and computational details of their operation. Nevertheless, despite my sympathy for their programme, I would emphasize that ethical behaviourism is not intended to be part of a science of machine behaviour. It is a claim about the kinds of evidence we can use to warrant our ethical attitudes toward machines.

be required for that. Can it describe how you look? Can it talk about the colours or fabrics you are wearing? Does its gaze follow you around the room? Does it act in other ways that suggest it sees where you are and what you are doing? And so on. The genuineness of a capacity or mental state depends on both the richness of the set of superficial states from which you infer its presence (its *completeness*) and its *consistency*. A robot that signals some affection for its human user on some occasions, but on other occasions does things that are contrary to the interests and well-being of that human, may be a false friend or conniving companion. But this is not because the robot lacks some relevant inner capacity that undergirds its superficial states but because it is inconsistent in its behavioural repertoire. Similarly, a robot that signals some forms of affection, but does not perform all the acts of care and affection that we usually associate with human friends and companions, may lack a full human-level capacity or mental state. This does mean that its affection is false or deceptive; it just means that it is incomplete or unsophisticated.

It is important to bear in mind the need for completeness and consistency in assessing the significance of a superficial signal. People may get the wrong impression or jump hastily to conclusions about what a robot is or is not capable of, and it is possible that a malicious actor could take advantage of this tendency (as we will see below) but in the absence of evidence for a hidden capacity or state, this is always to be assessed by checking other superficial states, not by assuming the absence of some underlying inner state.

It's worth reflecting on whether this ethical behaviourist model is similar to the embodied-relational model favoured by Damiano and Dumouchel (discussed earlier). There is certainly overlap between the two positions. Both agree that the Turkle-style condemnation of simulated feeling and affect is misguided: there is no epistemically accessible underlying

mental state against which we can assess the veracity of a superficial signal; the superficial signals are the most compelling thing we have to go on. Nevertheless, there are some differences. The Damiano and Dumouchel view flirts with a form of non-cognitivism about cognitive states and capacities, suggesting that the ontological status of anthropomorphic cues is determined by their pragmatic value within the human-robot relationship. If they are useful and help the parties to coordinate with one another, then we go with that and assume they are genuine. If not, we might need to reassess. The ethical behaviourist view is more explicitly cognitivist in nature, arguing that the ascription of capacities and mental states to robots can be more or less correct, depending on how complete and consistent the relevant set of behavioural cues is. Furthermore, the ethical behaviourist view is ethically-oriented, not ontologically or metaphysically-oriented: it is about assessing the ethical status of our interactions with robots, not about making claims about the true nature of cognitive capacities and mental states.

If the ethical behaviourist view is correct, it has a number of significant implications. The most immediate is that it supports the conclusion I wish to defend I here, namely that: superficial state deception is not typically best described as deception at all. To be more precise, the conclusion it supports is that the use of an anthropomorphic cue or signal by a robot is not deceptive or false merely because it comes from a robot. By itself the superficial signal does not violate any expectations or norms we associate with that signal. It is only deceptive if the signal is inconsistent with other cues and signals for which we acquire evidence. This has another significant implication. It means that many of the relationships and connections we have with robots can be genuine sources of meaning and value purely in virtue of their superficial properties. Robots can be genuine friends and companions if they consistently and completely signal this to their human users. This means that our interactions

with robots have an *axiological* richness to them that is often denied by critics of dishonest anthropomorphism.

None of this, however, implies that we only need to care about surface appearances when it comes to understanding robotic deception. A superficial state could be genuine (in the sense it genuinely signals the presence of some capacity or mental state) but could also be used to conceal some other capacity or mental state. This is what happens in the case of hidden state deception and it is to that that we now turn.

5. Hidden State Deception and Robotic Betrayal

Hidden state deception is a serious matter. If a robot uses some superficial signal to conceal or misdirect attention away from some capacity or function that it actually has, then this is *prima facie* morally concerning. Go back to the example from Kaminsky et al (2017) discussed earlier in this article: the robot that averts its eyes and continues to record activities using a concealed video camera. This practice is undeniably deceptive because it violates the expectations and norms we ordinarily associate with the superficial signal. Whether it is an ethically disturbing form of deception depends on the ulterior motive this concealment serves. As a general rule of thumb, Isaac and Bridewell's principle would seem to be apposite in these cases: if the ulterior motive serves some greater good then it may be ethically permissible, otherwise it is not. It would be hard to appeal to the greater good in the case of a corporation mining consumer behaviour for marketing insights, but perhaps more feasible in the case of a government trying to predict the next terrorist attack (though, of course, all the standard arguments and objections to this practice of governmental surveillance in non-robotic cases would still apply and need to be considered).

It is tempting to leave the matter there. But that might be a mistake. Taking onboard the implications of the previous argument, a case can be made for thinking that stricter rules should apply to the use of hidden state deception by robots, particularly by social robots, and particularly where the hidden state deception is contrary to the interests of the robot's primary users. This is because the use of such deception is often best understood as a form of betrayal ('bot betrayal'). Understanding it in these terms allows us to capture the special ethical harm that is involved when a social robot conceals its capacities from its human users, and allows us to invoke special ethical protections against such betrayal tactics. To make this case more compelling, I'm first going to present a theoretical account of betrayal and then explain how it applies to the case of hidden state deception.

The theoretical account comes from Avishai Margalit (2017). Like much of Margalit's work, it hinges on the importance of a distinction between *thin* relations and *thick* relations in human life. Margalit doesn't offer precise definitions of these concepts — though he does say that "[u]sing "thick" as an attribute of human relations is a figurative extension of "thick" as physically dense, like trees in a thick forest" (Margalit 2017, 52). He prefers instead to illustrate what they mean by describing some paradigmatic forms. A paradigmatic form of a thin relation would be a market exchange, based on rules of contract. Such a relation is governed by some basic moral and legal rules, but the moral commitments are minimal: it is intended to be a mutually beneficial exchange of limited form and duration (specified by the terms of the contract). Contrast that with the paradigmatic forms of a thick relation: the relations between family and friends. These are governed by a more comprehensive set of ethical norms and rules. They are not limited in duration or intended to serve a particular purpose. They might be mutually beneficial — in the sense that people 'get' something out of these relations — but they are not primarily thought of or conceived in these terms. The

parties to a thick relation are bound together by what Margalit refers to as “glue”. In the paradigmatic case, this glue involves a shared sense of belonging, shared history/memory and shared meaning (Margalit 2017, ch 3). In other words, the parties to the thick relation will feel at home with one another, will be characters in a shared narrative, and will find that this narrative gives them a sense of meaning and purpose.

According to Margalit “thick relations are the relations we care most about” (2017, 53). They have a special place in our lives and so are governed by special duties and ethical protections. In particular, they are governed by duties of loyalty and obligations of trust: the parties to the relation must protect one another’s interests and trust one another not to work against those interests.

Betrayal is the ‘ungluing’ of thick relations (2017, 47). Anything that gives people good reason to reevaluate the meaning of a thick relation can amount to betrayal (2017, 88-94). This means betrayal can take many forms, but one of the most common is when the betrayer sends false and misleading signals to the betrayed that convince the betrayed of the thickness of their relation, while at the same time saying and doing things that are contrary to the interests of that thick relation. To put it another way, one of the most common forms of betrayal is facilitated by hidden state deception. In the paradigmatic case, this will involve the betrayer saying and doing things with a third party that undermines the relationship they have with the betrayed (e.g. betraying a spouse by having an affair). As Margalit puts it “[b]etrayal is a ternary relation. Betrayal in its paradigmatic cases involves a third party on top of the betrayer and the betrayed” (2017, 70).

When it comes to the ethical assessment of betrayal there are easy cases and hard cases. The easy cases arise when the act of betrayal serves no positive end and would, in and of itself, be morally impermissible. Consider the French national who betrays their nation by facilitating the Nazi occupation and extermination of the Jews. We might think of what they did as an unethical act of national betrayal, but it was also deeply immoral. We do not pause to condemn it. The ethical demands of the thick relation pull in the same direction as the broader moral demands. The harder cases arise when there is some tension between the duties of loyalty within the thick relation and broader moral demands or civic duties. If your best friend confessed that they had an extra-marital affair, or if your son committed some petty act of vandalism, you might feel ethically torn about whether to stay loyal or give them up. In at least some such cases, the ethical thing to do might be to allow the demands of the thick relation to take precedence. Certainly most of us act this way on a daily basis: we are partial to our thick relations and favour them over our thin relations in most cases of conflict.

So how does this shed light on the problem of hidden state deception in robots? The argument is as follows. If human-robot relations belong to the world of thin relations, then the duties of loyalty and trust we can expect within those relations will be relatively minimal, largely to be set by contract and user agreement. Even if the duties specified by contract are strict they can be easily overridden by other moral considerations and interests.[4] If, on the other hand, human-robot relations belong to the world of thick relations, then special duties of loyalty apply. The use of hidden state deception (as long as it undermines the thick relation) is an act of betrayal and takes on a special ethical significance. It is not so easy to override this duty of loyalty by appealing to other moral considerations and interests.

[4] Easily overridden by other moral considerations, that is. They might still legally amount to a breach of contract.

The position being defended here is that, at least on some occasions, human-robot relations should be thought of as thick relations. There are several reasons for this. One obvious reason is that certain social robots are marketed and intended to serve in the role of thick relations. Robot companions, friends and carers, for example, are designed to take on a special significance in the lives of their users. They are intended to use anthropomorphic cues to encourage social acceptance and integration. The designers and manufacturers of such devices should not be allowed to shirk the duties of loyalty that are integral to thick relations simply on the grounds that the devices are robots. Second, although human family and friends are the paradigmatic cases of thick relations, there is some 'constructive flexibility' to the concept. Anything that creates the 'glue' that is the special marker of thick relations — shared narrative history, shared meaning and a sense of belonging — can help to create a thick relation. The claim here is that robots, by being integrated into our lives, by sharing moments with us, by reacting to us, laughing with us and helping us out, can create that glue. This may not be true for all robots, of course. Some robots may remain strictly in the world of thin relations (e.g. the autonomous vehicle or the robot barista that makes your coffee at your local coffee shop), but this does not mean some robots cannot enter the realm of thick relations. When they do, the use of hidden state deception will amount to a betrayal of that relation.

The advantage of taking this view of human-robot relations is that it allows us to explain the unique harm involved in acts of bot betrayal and to invoke the special ethical protections we expect in thick relations. What, after all, is so disturbing about hidden state deception in the general case? Why would it be especially wrong for a government or company to use an anthropomorphic robot to engage in concealed spying? Don't they do that to us anyway through other, non-robotic, digital technologies? Without commenting on the ethical

propriety of those other acts of digital spying, we can now say that there is something especially disturbing about the use of hidden state deception in the robotic case. The use of the superficial cues of friendship and companionship builds the glue that is needed for a thick relation and thereby builds up the ethical expectations and demands we associate with thick relations. The use of hidden state deception unglues the thick relation. It is consequently more ethically disquieting than in cases where the glue has not been built. This vindicates the need for special forms of transparency and trust building in human-robot relations.

Many people will object to this. Margalit himself might object to it[5] — at one point he suggests that thick relations are uniquely human affairs (2017, 65). If so, this is presumably because he believes robots don't have the capacities needed to build up thick relations with human users. In this sense, Margalit might share the suspicion voiced by critics of superficial state deception. He might believe there is something illusory or fake about the relations we have with robots. But the argument in the previous section should call that suspicion into question. If that previous argument is right, then robots can have the capacities needed to build thick relations with humans purely on the basis of superficial states. There is, consequently, an irony to the view being defended here. We can say that there is something especially disturbing about hidden state deception in the robotic case, something that demands special ethical protection, but only if we accept that superficial states by themselves are not deceptive. If we assume that they are, and that all human-robot relations are fake, then we cannot easily explain the unique harm involved in acts of robot betrayal.

Others might object on the grounds that we have different ethical norms or expectations of robots. I mentioned earlier the studies by Malle et al (2015) and Voilkis et al (2016) that

[5] He also might not. He doesn't discuss the issue at all.

suggest that people expect to abide by a more utilitarian ethical code than humans. Someone could appeal to such studies and argue that they suggest that robots will held to the duties of loyalty expected in thick relations. There are two things to be said in response to this. First, the studies by Malle et al and Voilkis et al focus on how people assess the actions of robots in trolley-style dilemmas. These are a unique set of cases and do not involve close robot companions making decisions about their owners or primary users. The results of these studies thus might not extrapolate to the world of close robot companions. Second, even if they did, these studies are descriptive only: they are about the attitudes people have and not the attitudes they ought to have. The argument I am making here is a normative one. It is saying that if robots are marketed as, designed as, or function in the world of thick relations, then they ought to be held to the duty of loyalty we normatively demand of thick relations.

Let me wrap up this argument with two final comments. First, note that the argument developed here doesn't apply only to hidden state deception. It could also apply to external state deception. Hidden state deception is more likely to undermine a thick relation since it means that the robot is deceiving you about its own capacities and how they are, or are not, directed towards you. Nevertheless, it is possible that certain forms of external state deception will do the same. If a robot companion constantly misdirects you or gives bad advice, for example, you might come to question the thickness of the relationship you have with it. What's more likely, however, is that we will expect robots to engage in some acts of external state deception on our behalf, in order to keep the thick relation glued together. This too follows from the concept of a thick relation: sometimes the interests of the thick relation will be served by deceptive signals being sent to others outside of that relationship. Consider the father who conceals the truth on behalf of his vandalistic son. This does, however, give rise to another problem: with whom should a robot be deemed to be in a thick relation? Could

the original manufacturers and owners argue that they are the ones in the thick relation with the robot and so it is okay if the robot deceives on their behalf? They could certainly try to argue this, but it would not be a plausible view. The thick relation is created between the robot and the end user, i.e. the one with whom the robot shares experiences, signals affection and builds a shared relationship narrative. The original manufacturers and designers are outsiders to the thick relation. There is, of course, some messiness to this ideal of thick relations between humans and robots. A robot could be in a thick relation with many end users and owe them all duties of loyalty. Sometimes those duties might conflict. But this is no different from a human who has more than one friend. The ethical norms are the same: you should not betray your thick relations to others or to one another.

Second, thick relations aren't the only relations that invoke special duties of loyalty. There are some purely professional and contractual relationships that give rise to such duties. In law, these are termed fiduciary relationships and classic examples would include lawyer-client, trustee-beneficiary and director-company relationships. It is possible that robots could take on such professional fiduciary roles and thus either they or their controllers and manufacturers could be bound by those special duties of loyalty. This might be a way of gaining special protection against robotic deception without making assumptions about the thick relations we could share with robots. This could work well in some cases but it is worth bearing in mind that (a) at present these relationships only cover a narrow range of professional interactions and would have to be extended to cover all the cases of human-robot interactions we might want to have special protections for; and (b) limiting special protections to these cases overlooks the potential benefits of welcoming robots into the world of thick relations.

Conclusion

Robotic deception is an ethical concern, but it is important to think clearly about its different possible forms. This article has tried to facilitate clear thinking by arguing that there are three distinct high-level forms of robotic deception -- external state deception, superficial state deception and hidden state deception – the latter two of which pose unique philosophical and ethical challenges in the robotic case. It has argued, in turn, that the second form of deception, superficial state deception, is not best thought of as a form of deception at all, and that the third form of deception is best thought of as a form of betrayal. While these three arguments enable us to see the distinctive harms that might be involved in cases of robotic deception, they leave plenty of questions left to be answered. For example, how exactly can we protect against robotic betrayal in practical terms? Could it be that we are just too vulnerable to such betrayal to allow for the use and manufacture of social robots? These are important topics for another day.

Acknowledgements: My thanks to David Gunkel for inspiring me to write this article and for his recommended reading about the topic.

References

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Damiano, Luisa and Dumouchel, Paul (2018). Anthropomorphism in Human-Robot Co-Evolution. *Frontiers in Psychology* Volume 9: Article 468

Danaher, John (2019a). The Philosophical Case for Robot Friendship. *The Journal of Posthuman Studies* 3(1), 5-24.

Danaher, John (2019b). Welcoming Robots Into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*. DOI 10.1007/s11948-019-00119-x

Elder, A. (2015). False Friends and False Coinage: A tool for navigating the ethics of sociable robots. SIGCAS Computers and Society 45(3): 248-254

Elder, A. (2017). Robot Friends for Autistic Children: Monopoly money or counterfeit currency? In Lin, Abney and Jenkins (eds) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence.* Oxford: OUP.

EU High Level Expert Group on AI (2019). Ethics Guidelines for Trustworthy AI. Brussels: European Commission. Available at https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

Graham, G. (2015). Behaviorism. In Zalta, E. (ed) *Stanford Encyclopedia of the Philosophy*, available at https://plato.stanford.edu/entries/behaviorism/ (accessed 10/7/2018)

Grice, Paul H. (1975). Logic and conversation. In Cole, P., and J.L. Morgan, eds. *Speech Acts*. New York: Academic Press, 41–58

Gunkel, David (2018). *Robot Rights*. Cambridge, MA: MIT Press.

Isaac, Alistair M.C. and Bridewell, Will (2017). White Lies and Silver Tongues: Why Robots Need to Deceive (and How). In Lin, P., Jenkins, R. and Abney, K. (eds) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford: Oxford University Press

Kaminsky, Margot, Ruben, Matthew, Smart, William and Grimm, Cindy (2017). Averting Robot Eyes. *Maryland Law Review* 76: 983.

Häggström, Olle (2019). Challenges to the Omohundro-Bostrom Framework for AI Motivations. *Foresight* 21(1): 153-166

Leong, Brenda and Selinger, Evan (2019). Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism. *FAT* Conference 2019*, DOI:10.1145/3287560.3287591

Neely, Erica L. (2014). Machines and the moral community. Philosophy & Technology 27 (1): 97–111. doi:10.1007/s13347-013-0114-y

Mahon, James Edwin (2015). The Definition of Lying and Deception. In Zalta, E (ed) *Stanford Encyclopedia of Philosophy*, available at https://plato.stanford.edu/entries/lying-definition/

Malle, Bertram F., Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. (2015) Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pp. 117-124.

Margalit, Avishai (2017). *On Betrayal*. Cambridge, MA: Harvard University Press.

Omohundro, S. (2008). The basic AI drives. *Proceedings of the First AGI Conference Artificial General Intelligence 2008*, Wang, P., Goertzel, B., Franklin, S. (Eds), IOS, Amsterdam, pp. 483-492.

Rahwan, Iyad, Cebrian, Manuel, Obradovich, Nick, Bongard, Josh, Jean-François Bonnefon, Jean-Francois, Breazeal, Cynthia, Crandall, Jacob W. et al. (2019) Machine Behaviour. *Nature* 568: 477-486.

Schwitzgebel, Eric, and Mara Garza. 2015. A defense of the rights of artificial intelligences. Midwest Studies in Philosophy 39 (1): 89–119. doi:10.1111/misp.12032.

Sebo, J. (2018). The Moral Problem of Other Minds. The Harvard Review of Philosophy 10.5840/harvardreview20185913

Sharkey, A. and Sharkey, N. (2010). Granny and the robots: ethical issues in robot care for the elderly. Ethics and Information Technology 14(1): 27-40

Shaw, Katherine (2015). Experiment on Human Robot Deception. Available at http://katarinashaw.com/project/experiment-on-human-robot-deception/

Shim, Jaeeun and Arkin, Ronald (2016) Other-Oriented Robot Deception: How Can a Robot's Deceptive Feedback Help Humans in HRI? *International Conference on Social Robotics* 2016 DOI:10.1007/978-3-319-47437-3_22, available at https://www.cc.gatech.edu/ai/robot-lab/online-publications/ICSR2016_JS_camera_ready.pdf

Simler, K. and Hanson, R. (2018) *The Elephant in the Brain*. Oxford: Oxford University Press.

Trivers, Robert (2011). *The Folly of Fools*. New York: Basic Books.

Turing, Alan (1950). Computing Machinery and Intelligence. *Mind* 49:433-460.

Turkle, Sherry (2007). Authenticity in the Age of Digital Companions. *Interaction Studies* 8: 501-507.

Turkle, Sherry (2010). In Good Company. In Wilks, Y. (ed) *Close Engagements with Artificial Companions*. Amsterdam: John Benjamins Publishing.

Voiklis, John, Boyoung Kim, Corey Cusimano, and Bertram F. Malle. "Moral judgments of human vs. robot agents. (2016) In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 775-780. IEEE.

Wagner, Alan and Arkin, Ronald (2011). Acting Deceptively: Providing Robots with the Capacity for Deception. *International Journal of Social Robotics* 3(1): 5-26

Wagner, Alan (2016). Lies and deception: Robots that use falsehood as a social strategy. In Markowitz, J (ed) *Robots that Talk and Listen: Technology and Social Impact*. De Grutyer https://doi.org/10.1515/9781614514404

Zawieska, Karolina (2015) Deception and Manipulation in Social Robotics. The Emerging Policy and Ethics of Human-Robot Interaction. Workshop Paper at *The 10th ACM/IEEE International Conference on Human-Robot Interaction* (HRI2015), available at https://www.researchgate.net/publication/272474319_Deception_and_Manipulation_in_Social_Robotics