Moral Difference between Humans and Robots:

Paternalism and Human-Relative Reason

TSUNG-HSING HO
NATIONAL CHUNG CHENG UNIVERSITY

This is a preprint of an article published in AI & SOCIETY. *The final authenticated version is available online at: https://doi.org/10.1007/s00146-021-01231-y*

ABSTRACT

Will robots be capable of moral agency? If moral agency is understood in behaviourist terms, according to some, robots could become moral agents that are as good as or even better than humans. Given the behaviourist conception, it is natural to think that between robots and humans there is no interesting moral difference in terms of moral agency (call it the *equivalence thesis*). However, such moral difference exists: based on Strawson's account of participant reactive attitude and Scanlon's relational account of blame, I argue that a distinct kind of reason available to humans—call it *human-relative reason*—is not available to robots. The difference in moral reason entails that sometimes an action is morally permissible for humans, but not for robots. Therefore, when developing moral robots, we cannot consider only what humans can or cannot do. I use examples of paternalism to illustrate my argument.

KEYWORDS: robot ethics; moral robots; artificial moral agency; paternalism; participant reactive attitude; meddling blame

With the advance of artificial intelligence, it looks less and less like a sci-fi fantasy that fully autonomous robots will free us from all sorts of tasks that are laborious, hazardous, or menial. To function efficiently, robots must be able to function with minimal human supervision. But how can we be certain that autonomous robots will not harm us or do anything bad? One natural idea is that robots' decisions must be *morally acceptable for humans*. Robots must be able to conform to moral rules as humans do. In other words, autonomous robots must be artificial moral agents, who can act as morally as, or even better than, we do.

However, could robots ever be moral agents? It depends on what we mean by 'moral agency'. Some conceptions of moral agency emphasise more on the psychological aspects of moral agency (Brożek & Janik, 2019; Laukyte, 2016; Himma, 2009). The psychological conceptions may require that moral agents be rationally or even emotionally responsive to moral reasons, able to justify their moral judgments. The psychological conceptions are often linked to issues concerning robots' moral responsibility (Hakli & Mäkelä, 2019; Johnson & Verdicchio, 2019).

While the psychological conceptions may be closer to what philosophers and laypeople have in mind about moral agency, I do not discuss artificial moral agency according to those conceptions because, for robots to function without human supervision, all we need is that their behaviours will produce effects that can be morally evaluated as morally right and wrong and conform to relevant moral norms. As Floridi and Sanders explain, moral agents are the 'entities that can in principle qualify as sources of moral action' (2004, 349). If moral action is understood thinly, in the sense that an action would count as moral action so long as it causes effects that can be evaluated as morally right or wrong, then we have a behaviourist conception of moral agency (Fossa, 2018; Floridi & Sanders, 2004; Gunkel, 2012; Grodzinsky, Miller, & Wolf, 2008; Beavers,

2012).[1]

To be clear, I do not mean that the behaviourist conception is *the* correct conception of moral agency. I mean only that the moral agency of robots discussed here is understood merely in the behaviourist conception. They may not be moral agents according to other conceptions of moral agency. More precisely, the kind of moral robot I want to discuss is the one that does not have emotions, feelings, consciousness, and sense of selfhood, despite being capable of moral agency (in accordance with the behaviourist conception). In other words, they are *not persons*.[2] My arguments below do not apply to robots that have personhood.[3] But I think that creating robots with personhood is unwise because

---

[1] One may worry that behaviourism seems to imply that natural events are moral agents since they can produce morally assessable effects. However, this worry is not serious to me. First, this is a challenge to behaviourism. My aim is to criticise the equivalence thesis (see below) looked plausible under behaviourism. Note that behaviourism is a popular thesis in robot ethics. Even if behaviourism itself is problematic, examining it from different perspectives remains worthwhile. Second, the equivalence thesis is not about any moral agent, but about the kind of moral agents that can perform as morally as humans. So, even if natural events are considered moral agents, they are not the targets of this paper.

[2] For an account of personhood that requires complex mental properties, see Lynne Rudder Baker (2000). Note that my view implies that the kind of moral robots under discussion here, despite being moral agents, are not moral patients (or are moral patients of much lower status than persons or sentient beings).

[3] Here I'm discussing the ontological issue of whether moral robots *are* persons, not a psychological, sociological or legal issue of whether they will be perceived or should

they ought to enjoy human rights and thus cannot be used merely as means to serve our needs. But I will not argue for it since it is not my concern here.

Some may question that moral agency requires personhood. But the kind of moral agency under discussion here is a behaviouristic one, which disregards mental features of agency. Hence, it does not rule out the possibility of moral agency without personhood. To make the setting more concrete, the kind of artificial moral agency discussed here is like the autonomous robots in the film *I, Robot*, suggested by Isaac Asimov's book of the same name. The film depicts a future where autonomous robots are widely deployed to serve humans' needs. The robots, despite lacking personhood, can obey moral rules and do not need to take orders from humans to function independently.[4] I assume that such robots are metaphysically possible, as they are depicted in many sci-fi films and novels. Those robots are moral agents according to the behaviourist conception.

This paper is structured as follows. In section 1, I explain that, since the behaviourist conception observes no essential difference between humans and robots in terms of moral agency, it encourages the thought—call it the *equivalence thesis*—that whether an action is done by a human or a robot makes no moral difference. However, I think

---

be recognised as such. It is possible that a non-person object is often perceived as a person. For example, pet owners often treat their pets as if they are persons. I discuss how this phenomenon affects my thesis below.

[4] In the film, a robot is created to have self-consciousness. I assume that it would be a person so that it is not the kind of artificial moral agency under discussion here. I refer to its numerous predecessors, which lack personhood and are treated as such by humans.

that it is false, particularly when the action is paternalist. In section 2, I outline my objection to the equivalence thesis, which is based on the idea of *human-relative reason*, inspired by P. F. Strawson's account of participant reactive attitude. In section 3, I explain how the idea of human-relative reason can refute the equivalence thesis. In section 4, I use Thomas Scanlon's account of blame to explain why robots are less suitable to blame humans, which is another instance to show that the equivalence thesis is wrong. Finally, I address some potential objections (section 5).

1. The Behaviourist Conception and Robot Paternalism

To get a better grip on the behaviourist conception of moral agency, we can look at the Turing Test. The test is designed to determine whether a machine is intelligent by comparing its performance with humans'. If the machine can perform in some respect as intelligibly as humans, to the extent that we may mistake its performance as humans', then the machine is considered intelligent in that respect.

Similarly, we can devise a Moral Turing Test on robots (Allen, Varner, & Zinser, 2000). A moral robot can act autonomously and cause morally relevant consequences. In the Moral Turing Test, the criteria of moral agency are defined by reference to the currently best moral agents, namely, humans. So, if a moral robot can act in some dimension as morally good as humans, to the extent that we cannot externally distinguish between them, then it should be judged as a *moral agent in that dimension*. An autonomous car could be judged as a moral agent in the driving dimension if it drives in a way that is externally indistinguishable from a morally good human driver. If a robot can act as morally good as humans in every aspect of day-to-day life, then it is a *whole moral agent*.

Several philosophers (Hall, 2011; Bostrom, 2014; Yudkowsky, 2008; Dietrich, 2007,

2011) speculates that robots might one day outpace humans in moral performances. If so, since the Moral Turing Test is behaviourist, what define the criteria of moral agency will be robots rather than humans. They think that robots will then be our *moral mentors*, instructing humans what we should and should not do.

Given the behaviourist conception, the idea that robots are our moral mentors looks plausible. Consider AlphaGo that beats the human Go masters. Naturally, human players will analyse and emulate how AlphaGo plays. But moral robots could go well beyond being our mentors. Indeed, they could be our *moral guardians*: not only do moral robots teach us ethical values and moral norms, but also they may actively intervene to prevent us from committing wrongdoings or self-harm (Tegmark, 2017).

If moral achievement contributes to one's well-being, then stopping us from committing wrongdoings is a form of paternalism. Paternalism is the interference with someone, 'against their will, and defended or motivated by a claim that the person interfered with will be better off or protected from harm' (Dworkin, 2020). Call the idea that robots act as our guardians—*robot paternalism*. By robot paternalism, I mean the idea that robots autonomously perform—not under humans' directions or supervision—paternalist actions towards humans.[5] The difference between human and robot paternalism I am interested in is only that the agent who performs a paternalist act is a human or a robot. I assume that robot paternalism, like human paternalism, is justifiable in some circumstances. A more interesting question is whether robot paternalism is *equally justifiable* as human paternalism; in other words, the question examined in this paper is this: is there any interesting moral difference concerning who—be it a human

[5] Autonomous robots still act under humans' directions in the sense that they are designed by humans. I discuss how this fact could affect my thesis later.

or a robot—performs the paternalist act toward humans?

By ‘moral difference’, I mean the difference in terms of moral reasons in favour of against an action. Whether an action is morally justified or permitted is determined by weighing all the reasons for or against it. But my concern is not moral justification. For two agents who perform the same type of action, while their actions could be both justified or not justified, their reasons might be different in strength. Therefore, the agent who has the stronger reasons is more appropriate to perform the action than the other, though the latter’s action is also justified. For example, suppose that a little girl is in danger of drowning. You and her father are the only people who can save her in time (you both are good swimmers). Of course, both of you are permitted in saving her, or even obliged to do so. But the fact that the man is the girl’s father makes him, other things equal, have stronger reason than you. The father is thus more obliged to save her. We may even criticise the father if he is standing by to see if you will jump into the water first. It is this kind of moral nuance that I want to discuss in this paper: namely, will the fact that the agent who performs a paternalist act towards humans is a robot rather than a human, other things equal, makes some interesting difference in moral reasons?

To be sure, in some circumstances, a difference in moral reasons could change the moral status of an action. In the above example, if both of you are doctors and on the way to meet your patients and you know each other very well (you know he is the girl’s father and a good swimmer who will unhesitatingly and can easily save her daughter, and he says that he can save her by his own), you may not be obliged to save the girl, but her father remains obliged.

Is there any moral difference between human and robot paternalism? Since the behaviourist conception of moral agency observes no essential difference between

humans and robots as far as moral agency is concerned, it's natural to think that all the reasons for and against a human's paternalist act toward humans are also available to robots in the same situation. If so, it entails the *equivalence thesis*: the same set of moral reasons are available both to humans and robots in the same circumstance.

2. My Objection to the Equivalence Thesis: An Outline

To refute the equivalence thesis, I will argue that some reasons that can justify paternalist acts towards humans are available only to humans, but not to robots. Before presenting my argument, I want to clear up some potential confusions. Some may find that the talk of reason perplexing because the discussion here assumes the behaviouristic conception of moral agency. Since behaviourism excludes any psychological features of agency, how could I talk about reason, which requires certain mental capacities to grasp?

This confusion is understandable since the notion of reason is ambiguous. Philosophers distinguish between *motivating reason* and *normative reason* (Alvarez, 2017). Motivating reason is what motivates the agent to act. Thus, motivating reason is 'in the agent's mind'. Throughout the paper, however, I discuss only normative reason. Normative reasons justify actions, rendering them permissible or obliged. Agents do deliberate normative reasons when they decide whether to take actions. To that extent, normative reasons are also in their mind. In moral philosophy, however, normative reason is often used in an objective way, namely, to assess the moral status of an action (whether it is right or wrong, permissible or not). Take the drowning little girl for example again. The fact that her father is a good swimmer is a normative reason that requires him to save her. He may not think about this reason. Indeed, he is likely to try to save his daughter immediately without any deliberation. But the normative reason *is there*, in the sense that it supports a moral requirement. And the father fulfils that

requirement even though he does not act out of that reason. His action is justified by the normative reason.

By saying that the reasons are available to robots, therefore, I do not mean that robots can use the reasons to justify their actions. I mean that the reasons are there to justify the robots' actions. The talk of reason is essential when we aim to design moral robots, even though we design them under the behaviouristic conception of moral agency. For we need to decide which actions are permissible or obligatory for robots to take. The decisions are based on facts about moral norms, potential benefits and harms. Those facts are normative reasons, which do not depend on or presuppose any psychological features of robots. So, if a robot sees the girl drowning but fails to take any action to save her, it is morally faulty. It is morally faulty, not because it cannot deliberate normative reasons, but because it fails to take actions that meet the demands of normative reasons. Therefore, the talk of reason here does not require robots to comprehend normative reasons.

Return to my objection to the equivalence thesis. I argue that some reasons that are available to humans are unavailable to robots. The Difference in reasons implies that in some circumstances a paternalist act is permissible to a human but not to a robot, or a human—because of having stronger reason—is a more appropriate agent to perform a paternalist act than a robot, despite both being justified. Therefore, the equivalence thesis is wrong.

To illustrate, let's first look at an example of human paternalism. Suppose James, a seven-year-old boy, is obese, which endangers his health. His mom, Mary, takes him to see the doctor, David. Could Mary or David force James to reduce weight? To simplify the issue, let's assume that there are only two kinds of *agent-neutral reason* concerning

any paternalist act:[6] the *welfarist reason* in favour of paternalism and the *autonomy reason* against paternalism. Both reasons are available to Mary and David. If those were the only relevant reasons, then there is no moral difference concerning who take the paternalist act. Nevertheless, since Mary is James's mom, she has a strong *agent-relative reason* unavailable to David, which might permit her, but not David, to act paternalistically towards James. Therefore, the difference in agent-relative reason makes the paternalist act by Mary and David unequally justifiable.

Similarly, I will argue that there are agent-relative reasons available *only to humans* that permit humans to act paternalistically towards humans. The idea that agent-relative reasons can permit or forbid different agents to act paternalistically is all too familiar. But the agent-relative reasons familiar to philosophers are generated by special relationships among the people concerned. So, how is the idea of agent-relative reason relevant here—for the moral issue I want to discuss is about humans and robots in general, not about those who are in specific relationships?

The answer is that the kind of agent-relative reason I will argue for is available *to all humans qua human*—or at least to those humans who are capable of moral agency. Call it *human-relative reason*. Human-relative reason is overlooked in ethical theories because there is no need to consider non-human moral agents—particularly those who lack personhood. Without non-human moral agents, the distinction between agent-

[6] Agent-neutral reason contrasts with agent-relative reason. In the example, the welfarist reason and the autonomy reason are agent-neutral because they can be specified without reference to the agents who perform the act. However, the fact that Mary is James' mother is an agent-relative reason because it is specified with reference to the agent, Mary.

neutral reason and human-relative reason makes no difference in practice. Given the possibility of artificial moral agency, I want to highlight the idea of human-relative reason and examine how it could affect the moral relationships between humans and robots. If the human-relative reasons exist, then it entails that the equivalence thesis is wrong.

To recap, the agent-neutral reasons for and against paternalism are available both to humans and robots. The justification of paternalism often involves agent-relative reasons. The usual kinds of agent-relative reason are compatible with the equivalence thesis because those reasons are available to individual agents—humans as well as robots—who are in specific relationships with the patients of their paternalist acts (more on this in section 5). The equivalence thesis is rejected by the existence of human-relative reasons because it is available only to humans, but not to robots.

3. Human-Relative Reason for Paternalism: A Strawsonian Account

My argument for human-relative reason is based on Strawson's ideas about participant reactive attitudes in his seminal article, "Freedom and Resentment" (1974). Strawson maintains that, by default, we adopt the *participant reactive attitude*, namely, that humans are by default participants in interpersonal relationships, in which we expect others to treat us with respect and goodwill. When our expectation is or is not met, it is natural and appropriate for us to respond to people with what Strawson calls *reactive attitudes*, such as gratitude, forgiveness, shame, resentment. For example, if someone is kind to us, we naturally and appropriately feel gratitude for them; or if someone is hostile or disrespectful to us, our resentment towards them is also natural and justified.

Strawson's ideas in *Freedom and Resentment*, however insightful they are, are notoriously difficult to interpret. Here I rely on Gary Watson's interpretation (2014).

According to Watson, Strawson's account identifies a normative framework that provides reasons for participants in interpersonal relationships:

> [Participant reactive attitudes] ground a normative framework, I take it, because sentiments are ways of valuing, and valuing is taking certain considerations as reasons. … On my reading, it is significant not only that Strawson speaks of the "commitment to the interpersonal attitudes" as nonrational, but that he speaks here of commitment, suggesting that they play a structural role in our normative lives, as defining in part what counts as reasons for feeling and acting. (Watson, 2014, 22-23)

According to Watson, participant reactive attitude is *reason-giving*. It constitutes a framework of interpersonal relationships, generating reasons for participants to express emotions. Emotions are modes of valuing (Tappolet, 2016; Epley, 2019), which provide further reasons to express emotions, such as verbal or bodily expressions. To illustrate, consider this example,

> *Resentment*. Jane fell off from the stairs. Although she was not seriously hurt, she was in pain and could not temporarily get up by herself. Charlie—who has never acquainted with Jane—saw that Jane needed help, but he just walked away, showing no sympathy and care. Seeing Charlie's indifference, Jane felt resentment towards him.

Presumably, Charlie is not obliged to help Jane; after all, Jane could get up by herself. Nevertheless, Jane and Charlie—despite being total strangers—are both participants in an interpersonal relationship. Seeing that Jane was hurt, Charlie had reason to express his concern or even assistance. Since Charlie failed to meet the expectation, Jane had reason to feel resentment towards Charlie.

This kind of reason—generated by participant reactive attitudes for us to feel and express our attitudes towards others—is what I call human-relative reason. Human-relative reason is a *pro tanto* reason in favour of paternalist act.[7] For when seeing someone is in trouble, participant reactive attitude directs us to show our goodwill and give her assistance appropriate to her needs, even if she says that she doesn't need assistance. Consider a variation of *Resentment*. This time, Charlie moved towards Jane and intended to assist Jane. Jane waved and said that she could get up by herself. Nevertheless, Charlie might still give her a hand and express that he was glad to help her. Although Charlie's response was against Jane's will, it could still be justified as required by the reason from participant reactive attitudes.

I want to stress that human-relative reason doesn't justify any sort of act that goes against the patients' wills. To be clear, human-relative reason is *pro tanto*, so it can be overridden. Therefore, the worry that human-relative reason could justify unwanted harassment is unfounded because the feelings and reactions favoured by human-relative reasons must be appropriate in accordance with the relationships and circumstances to which the agent and the patient belong. So, if Jane said that she didn't need help, it

[7] Philosophers use the notion of *pro tanto* reason in this sense: to say that R is a *pro tanto* reason in favour of an action x is to say that R, *considered on its own*, can justify doing x. When we determine whether to do x, we need to weigh all relevant *pro tanto* reasons for and against doing x. If *pro tanto* reasons against doing x is stronger than reasons for doing x, then doing x is not justified—but it remains true that R is a *pro tanto* reason for doing x. In other words, if there is a *pro tanto* reason in favour of doing x, doing x is, *other things equal*, justified, but it could be unjustified *all-things-considered*.

would be less appropriate for Charlie to directly help Jane up. But it's not inappropriate to express his concern and ask again whether she needed assistance. Similarly, Jane was justified in feeling or expressing resentment towards Charlie, but her reaction must be within a reasonable degree.

Besides the usual kinds of reason for paternalism, therefore, human-relative reasons provide further justification for agents who have reactive attitudes. Human-relative reasons, however, are not available to robots because robots, by stipulation, are not persons and are not by default participants in interpersonal relationships with humans. Robots do not have genuine emotions, so they do not have reactive attitudes to express. Accordingly, robots lack participant reactive attitude to generate human-relative reason that can justify paternalist acts towards humans.

To illustrate, let us imagine a future society—like the one in *I, Robot*—in which autonomous robots are widely used. Now consider the following case,

> *Suicide*. Tom, who is seriously ill and suffers great pain, is determined to commit suicide. For Tom to commit suicide, however, it would be difficult since robots are everywhere and are more agile and stronger than humans. Robots will stop Tom killing himself even if he expresses his determination to die.

While saving a human's life is great, it seems awful to me that, if Tom has thought it through and decided to leave the world, he is forced to live by robots. Things are somewhat different, on the other hand, if Tom is saved by a human. Imagine this time before Tom is going to kill himself, Rachel happens to pass by. She tries to talk Tom down, though Tom does not waver. Still, Rachel stops Tom from committing suicide. Suppose that Rachel and the robot are both justified in stopping Tom. Nevertheless, it seems less awful to me that Tom's autonomy is infringed by Rachel because of the

human-relative reason. Let me explain why.

The human-relative reason is provided by reactive attitudes. This means that Rachel's paternalist act is out of her care for Tom, whereas the robot has no care. Though their actions are both justified, Tom would, naturally and justifiably, feel resentment towards the violators of his autonomy because he may feel that they show disrespect to him. Strawson tells us that reactive attitudes are mainly in response to the qualities of people's wills (goodwill, ill will, or indifference). Since the robot has no will, Tom's resentment towards it targets at *nothing* (his resentment ought to target people who deploy the robots; see more on this in section 5). Thus, his resentment towards the robot would be empty and meaningless. There is no point to resent a robot since it does not do it out of ill will or indifference; it just has no will. The emptiness of Tom's feelings reveals a notable difference of robot paternalism from human paternalism: that is, robot paternalism would make our reactive attitudes unable to perform the therapeutic function of emotion. When Tom resents Rachel, his resentment is meaningful because he may feel that Rachel does not respect him. Tom's resentment could release his anger and frustration over his autonomy being violated. However, his frustration with the robot could not be released in the same way when he realises that being angry at the robot is pointless. Therefore, robot paternalism would be less acceptable from Tom's perspective.

To see my point more clearly, it's helpful to look at an interesting plot in *I, Robot*. The protagonist, Detective Del Spooner, is investigating a murder. He suspects that the murderer is a robot. Susan Calvin, a robo-psychologist, dismisses that possibility because robots are programmed to be unable to harm humans. Spooner tells her his experience that makes him distrust robots: because of a car accident, a girl and Spooner were drowning, and a passing robot could only save one of them; Spooner repeatedly

asked the robot to save the girl, but the robot could only save him because the logical conclusion is that his survival rate is higher than hers. Finally, Spooner gives his reason for distrusting robots:

> I was the logical choice. It calculated that I had a 45% chance of survival. Sarah only had an 11% chance. That was somebody's baby. 11% is more than enough. A human being would've known that. Robots, [indicating his heart], nothing here, just lights and clockwork. Go ahead, you trust 'em if you want to. (Proyas, 2004)

Spooner's distrust of robots may seem unreasonable since—despite against his will—saving his life was permissible for the robot. It seems unreasonable to blame and even resent for someone to do something morally permissible and even praiseworthy.[8] In light of the above discussion, however, we can see how Spooner's attitudes towards robots could be reasonable. Spooner knows that robots have no will ('just lights and clockwork'). This makes Spooner more frustrated since he realises that his life is interfered by someone to whom his resentment is pointless. On the contrary, if Spooner was saved by a human, Spooner could meaningfully blame her for not listening to him. She might apologise and say that she meant no disrespect, which, unlike robots, genuinely expressed her goodwill (or lack of ill will). Spooner might forgive her, so that their relationship could be thus repaired. There are many meaningful emotional reactions among humans—supported by human-relative reasons—that are missing between humans and robots. Spooner's distrust of robots can thus be interpreted as his dismissal of robots as qualified candidates who can have a say over his life.

[8] Several accounts of blameworthiness (Capes, 2012; Graham, 2014) argue that a morally permissible action may nevertheless be blameworthy, which can support that Spooner's blame is justified.

Second, the fact that the human-relative reason is available only to humans entails that, other things equal, human paternalism is more likely to be justified than robot paternalism. Even if Rachel and the robot are both permitted to stop Tom, it remains true that only Rachel has human-relative reason, in addition to other reasons shared both by her and the robot. This is not a small point, because when we consider whether to let a human or a robot perform a paternalist act, the human would be preferable to the robot since, other things equal, he or she would have a weightier overall reason than the robot. Therefore, unless there are other reasons in favour only of robot paternalism, human paternalism is more justifiable than robot paternalism.

4. Reactive Attitudes in Behaviouristic Moral Agency?

Before moving on, I want to address a worry. That is, how can I reject the equivalence thesis on the grounds that robots lack reactive attitudes, while accepting the behaviouristic conception of moral agency? Either I have to reject the behaviouristic conception altogether, or I cannot appeal to the Strawsonian idea to reject the equivalence thesis.

This worry, however, is misplaced because the concept of moral agency needs to be separated from the concept of normative reason. As I have explained in section 3, the kind of normative reason under discussion here is in the objective sense. That is, normative reasons are used to assess whether an action is morally permissible or obligatory. And normative reasons are provided by the facts that have moral significance in certain situations. What determine which fact is reason-providing are theories of normative ethics, not theories of moral agency. Behaviourism only rejects any psychological criteria of moral agency. But it does not deny that moral agents can have psychological features. Nor does it deny that their psychological features could sometimes be reasoning-providing.

I suspect that one may still worry that traditional ethical theories usually assume psychologism about moral agency. So, if it is replaced by behaviourism, then ethical theories must be revised to exclude psychological features from moral reasoning.

However, this is a mistake. To see why, we can look to paternalism, in which psychological features are morally significant. First, the patients' psychology must be taken into account because their well-being—one of the central concerns in paternalism—is partially constituted by their feelings of happiness and life satisfaction. Therefore, even if behaviourism about moral agency is assumed, we still need to recognise the moral significance of psychological features.

Certainly, the above shows only that the psychology of moral *patients* is morally significant, which says nothing about moral agency. However, as I've argued in section 3, it is also morally significant concerning which agent performs the paternalist acts. That moral significance is manifested in the difference regarding agent-relative reasons, which caused by the relationships of the patient with different agents. Some relationships are constituted by certain psychological features: for example, friendship by loyalty and care, familial relationship by love. Thus, even if behaviourism is assumed, the psychological elements in relationships could still affect the strength of the agent-relative reasons. For example, a disloyal husband who no longer loves his wife may have a weaker agent-relative reason than her loyal friend to act paternalistically towards her. This is possibly so because behaviourism does not deny that agents can have psychological features and it is a fact that psychological elements can greatly affect the quality and strength of interpersonal relationships, which then affect the existence and weight of agent-relative reasons.

The Strawsonian idea, therefore, is an extension of this idea: humans are normally participants of interpersonal relationships, which are partially constituted by our

reactive attitudes. The interpersonal relationships among humans provide us with a special kind of agent-relative reason, namely, human-relative reason, which is unavailable to robots because they lack reactive attitudes requisite for interpersonal relationships. The Strawsonian idea is compatible with behaviourism.

Indeed, I think that the worry demonstrates how the equivalence thesis is attractive under behaviourism. Imagine now that we are designing fully autonomous robots that will act morally. Naturally, we try to make robots emulating humans morally in an ideal way—hence, the equivalence thesis. If psychological about moral agency is assumed, we can easily spot the psychological difference between humans and robots and take it into consideration in our design. However, if behaviourism is assumed, it is tempting to brush aside our psychological difference with robots because it seems irrelevant under behaviourism. Hence, behaviourism makes the equivalence thesis look compelling.

However, this mistake is due to the failure of distinguishing between theories of moral agency and normative ethics. It misses the point of behaviourism is simply broadening the scope of moral agency, rather than narrowing the space of moral reasons. By adopting behaviourism, we now recognise robots as moral agents. But we do not thereby revise moral theories to dismiss psychology as reason-providing. For it remains true that many agents and patients have psychological features and their psychological features are morally significant and reason-providing.

5. Blaming Humans: A Scanlonian Account

In this section, I want to argue against the equivalence thesis from another angle: *blaming*. That is, if humans are done something blameworthy, is there any moral difference in whether it is a robot or a human that blames them? Other than the

Strawsonian account of human-relative reason, I will use Scanlon's relational account of blame (2008) to support my claim.

In general, a blameworthy action deserves to be blamed. However, some people may lack the standing to blame it. One oft-discussed case is hypocrisy (Coates & Tognazzini, 2018). People who have committed certain blameworthy actions are not suitable to blame others for similar actions.

Another case that receives less attention is *meddling blame*, which will be my target here. Central to the concept of meddling blame is the idea that blaming is '*not your business'.* When does one's blame count as meddling? Scanlon's account of blame offers a satisfactory explanation: 'To blame a person for an action, in my view, is to take that action to indicate something about the person that *impairs one's relationship with him or her, and to understand that relationship in a way that reflects this impairment*' (Scanlon, 2008, 123; my italics). To illustrate, consider this example:

> *Couple*. Will is arguing heatedly with his wife Lizzy in a mall about whether to buy a luxurious item. Kate, passing by and overhearing their argument, cannot help criticising Will that he should listen to his wife. In response, Will replies, 'Not your business'.

Intuitively, Will's response is justified because, according to Scanlon, the argument between Will and Lizzy does not impair their relationship with Kate. Kate is thus not in a position to blame Will. Her blame is meddling.

Now consider a revised version of *Couple*. This time, Will and Lizzy argue too loudly in the mall and Kate blames them for that. Kate's blame could be appropriate, because their behaviour is disrespectful to people in the same space with them and her blame suitably reflects that impairment.

What if it is a robot that blames Will? Since robots are not persons, there is no interpersonal relationship between robots and humans to impair. The views of Strawson and Scanlon together imply that, if a robot blames a human's misbehaviour, that will count as meddling blame. To illustrate, consider this time a robot—sent by its owner to buy some stuff—blames them that they should not argue so loudly. It seems to me that its blame is not appropriate because the robot is not a person and thus does not receive the disrespect shown by their behaviour. Since no genuine interpersonal relationship between humans and robots exists, for robots to blame humans always counts as meddling.

Certainly, many people do not enter interpersonal relationships with Will and Lizzy, so their blaming would be meddling, too. The crucial point, however, is that humans can, but robots cannot, establish genuine interpersonal relationships with other humans. This entails that there is always a *pro tanto* reason against using robots to blame humans. Of course, this does not mean that necessarily robots cannot blame humans because other reasons may overweigh the madding reason.

6. Objections and Replies

I have argued that robots, by stipulation, lack reactive attitudes and thus, according to Strawson, are not by default participants in interpersonal relationships. Therefore, human-relative reason is unavailable to robots, so the equivalence thesis is wrong. It entails that, other things equal, robot paternalism is less justifiable than human paternalism.

In response to my argument, there are two types of criticism: (i) despite the lack of human-relative reason, robot paternalism is as equally justified as or more justified than human paternalism because there are further reasons in more favour of the former, and

(ii) human-relative reason is available to robots because they can be participants in interpersonal relationships.[9]

5.1 Other Reasons in More Favour of Robot Paternalism

As I've repeatedly emphasised, human-relative reason is merely *pro tanto*. So, my view is compatible with the claim that there are other reasons in favour of robot paternalism rather than human paternalism. For example, if the task involves substantial risks of harm or robots are simply more capable of performing the task than humans, they would be strong reasons in favour of letting robots undertake the task. Both could be strong enough to overweigh human-relative reason, which entails that robot paternalism could be more justifiable than human paternalism in those situations.

If so, one may question that my thesis about the unavailability of human-relative reason to robots is not philosophically significant. However, I think that this response misses a crucial difference. Namely, the above reasons in favour of robot paternalism are *situation-specific*, namely, they are available only to some robots in certain situations. It would not be surprising that different situations would favour different agents to perform the same type of action. For example, when a child is drowning, people who cannot swim don't have reason requiring them to jump into the pond to save her, whereas those who can swim may have reason to do so. The difference in situation-specific reason, however, does not show that these people have any essential moral difference insofar as moral agency is concerned. Human-relative reason, on the other hand, is not situation-specific because it is generally available to humans and not to robots, insofar as they are capable of moral agency (though the strength of human-relative reason would vary according to the relationships among the people concerned).

[9] I am grateful for the reviewers for raising the following objections.

Therefore, the existence of human-relative reason is sufficient to show that the equivalence thesis is false.

5.2 Robots being Participants in Interpersonal Relationships

One may argue that human-relative reason is available to robots because humans can build relationships with them. By analogy, it is very common for pet owners to develop close relationships with their pets, even though their pets are not persons. Similarly, the fact that robots are not persons doesn't hinder the possibility of quasi-interpersonal relationships between humans and robots. Humans' relationship with their robots could be even closer than with strangers. Therefore, human-relative reason is available to robots.

I have three responses to this objection. First, the analogy between robots and pets is problematic. Although pets are arguably not persons, some pets, such as dogs, do have emotions. So, there could be reason from participant reactive attitudes for people to feel grateful for their dogs because the dogs do have qualities of will and are loyal to their owners. In contrast, robots lack emotions, so their relationships with humans are not really interpersonal. So, even if we don't recognise the relationships between humans and pets as genuinely interpersonal, theirs are closer to interpersonal relationships than the ones between humans and robots. This analogy, therefore, doesn't support the view that robots can obtain human-relative reasons, which are generated from reactive attitudes.

Second, it's true that some owners may treat their robots as genuine persons. To that extent, they are more willing to be interfered by their robots rather than by some human strangers. But this does not imply that those robots do obtain human-relative reasons. It is important to distinguish between the claim that robots are genuinely persons and

the claim that robots are treated as if they are persons (but they are not). I've assumed that the kind of robots under discussion are not persons. Now, we may interact with robots as if they are persons, but it does not follow that human-relative reason is readily available to robots. For it is humans that *choose* to establish relationships and only then human-relative reason is available to the robots. Not to mention the fact that some humans—like Spooner—do not want to have interpersonal relationships with robots; for them, robots are just instruments, like cars or computers. On the other hand, Strawson maintains that humans are by default participants in interpersonal relationships with each other. In other words, human-relative reason is by default available to all humans, but only to robots by the choice of some people. [10] Consider this analogy. When we choose to establish interpersonal relationships with robots, it's like someone who adopts a child. They have reason to look after each other, but the reason is there only after the parent chooses to adopt the child. Unlike biological parents and children, familial duties and reasons are naturally given. This difference is enough

[10] Coeckelbergh (2011) and Tavani (2014) argue that we can trust robots and the trust relationship between humans and robots could be the default. This may seem in conflict with my thesis. However, the conflict is merely apparent. As they indicate, we can trust non-personal entities, such as social institutions or machines. Surely, the fact that I can trust my car being reliable enough to last for the whole trip does not show that the relationship is not interpersonal. Thus, the reason is different from human-relative reason. Furthermore, the fact that we can by default trust robots does not in itself give a reason that robots can interfere with our autonomy. For example, the fact that I trust a doctor is not a reason that the doctor can violate my autonomy concerning my health, unless I choose to be her patient.

to reject the equivalence thesis.

Furthermore, since the reason is provided by the owners' approval. This means that the owners are willing to be interfered by their robots. Clearly, this is not paternalism.

Now, my critics might approach this objection from a different angle. They would say that if robots improve our well-being (whether we like it or not), we should feel grateful for them. But it means that we can and should have reactive attitudes towards robots. Doesn't it show that we can establish meaningful interpersonal relationships with robots?

I have explained that Spooner's resentment towards robots frustrates him even more because he realises that robots have no will at all. According to Strawson, since reactive attitudes respond to one's qualities of will, robots are not suitable objects of our reactive attitudes. True, we are psychologically capable of feeling resentment or gratitude towards robots (humans are prone to anthropomorphism). But Strawson's point is a normative one. Normatively speaking, we are more appropriate to resent or thank the people who deploy robots to assist or obstruct us. We could feel gratitude to them because they are thoughtful of our well-being. Or we could—as Spooner did—resent them for using robots to tyrannise us in the name of our own good. Therefore, our reactive attitudes towards robots should be directed at people who design or deploy robots rather than robots themselves.

Now, my critics might respond: 'True, human-relative reasons are not available to robots. But this is irrelevant, because robots are designed and deployed by some humans to assist us, thus expressing their care and goodwill towards us. In that sense, robots serve as surrogates of humans. Thus, there is no robot paternalism, only human paternalism.'

However, this is not an objection to my thesis; in fact, they are compatible. As I have

argued, it is more appropriate to resent or thank the people who deploy robots to act paternalistically towards us. So, it is correct that their care and goodwill could be mediated through robots. To make the objection work, it must say that there is no moral difference between a paternalist act taken by a human and a paternalist act taken by a robot on behalf of the human since the robot is merely a surrogate of the human. Understood in this way, however, it is implausible.

Imagine that a father is not parenting his daughter and buys the best care robot to look after her. The robot acts in a seemingly loving and caring way as if it is a good father. Does it mean that the daughter will feel that her father's love could be mediated through the robot? I don't deny that some people may feel so. Nevertheless, it's reasonable for the daughter to feel that her father does not love her enough. Even though the robot can behave as if it is her father, the fact that the father does not spend enough time with her shows that he does not love her sufficiently.

Even if the father genuinely loves her and thinks that the robot can do a better parenting job than him, the daughter may not feel his love being mediated through the robot. For, unless there is good reason that the father cannot do parenting by himself, if he loves her daughter, he should learn to do a better parenting job to look after his daughter. There is no better way to show parental love by accompanying their children. Therefore, it is wrong to think that robots can replace humans completely once they are designed properly.[11] Indeed, the fact that humans use robots as their surrogates in some circumstances could mean that they want to withdraw from interpersonal relationships,

[11] See (Sharkey & Sharkey, 2010) for a more substantial appraisal of childcare robots. While they list several advantages, they conclude that a near-exclusive care of children by robots is undesirable.

which shows that they care less and love fewer people in their relationships.

Another possible objection is that, since we adopt a behaviourist conception of moral agency, we should similarly adopt a behaviourist conception of quality of will and reactive attitudes. Since robots behave indistinguishably from us, they should be considered as having minds. When robots act paternalistically towards humans, we should think that their actions are out of their good wills. Therefore, robots can have participant reactive attitudes and human-relative reason.

A behaviourist conception of mind, I think, is controversial. If it is accepted, then I think that moral robots ought to be recognised as persons since they are now regarded as having intentions and reactive attitudes. Hence, human-relative reason is available to robots. But I have disclosed that my argument is conditional upon the assumption that robots do not have personhood. But this limitation does not render my thesis trivial. The assumption is not implausible because one can reasonably say that robots are acting *merely as if* they are intentional, however they behave like humans. In many sci-fi films and fictions, intelligent robots are considered emotionless and not treated as persons. Unless my opponents could prove that moral robots are necessarily persons, my thesis remains worth considering.

7. Conclusion

Based on the views of Strawson and Scanlon, I have argued that human-relative reason is unavailable to robots. Human-relative reason signifies an essential moral dimension of human interaction. We humans are participants in interpersonal relationships, which requires us to treat each other with goodwill and respond to their qualities of will with suitable reactive attitudes and actions. So, humans are, by default, situated in a normative framework that provides them reason to enhance or impair the relationships with each other. On the contrary, since robots are, by assumption, not participants in

interpersonal relationships, human-relative reason is not available to robots. The unavailability of human-relative reason to robots shows that the equivalence thesis is wrong: if action is permissible for humans, it may not be permissible for robots.

References

Allen, C., G. Varner, & J. Zinser. 2000. "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental & Theoretical Artificial Intelligence* 12 (3):251-261. doi: 10.1080/09528130050111428.

Alvarez, Maria. 2017. Reasons for Action: Justification, Motivation, Explanation. In *The Stanford Encyclopedia of Philosophy*, edited by N. Zalta Edward.

Baker, Lynne Rudder. 2000. *Persons and Bodies: A Constitution View*: Cambridge University Press.

Beavers, Anthony F. 2012. "Moral Machines and the Threat of Ethical Nihilism." In *Robot Ethics: The Ethical and Social Implication of Robotics*, edited by Patrick Lin, George Bekey and Keith Abney, 333-344. Cambridge, MA: The MIT Press.

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

Brożek, Bartosz, & Bartosz Janik. 2019. "Can Artificial Intelligences be Moral Agents?" *New Ideas in Psychology* 54:101-106. doi: 10.1016/j.newideapsych.2018.12.002.

Capes, Justin A. 2012. "Blameworthiness without Wrongdoing." *Pacific Philosophical Quarterly* 93 (3):417-437. doi: 10.1111/j.1468-0114.2012.01433.x.

Coates, D. Justin, & Neal A. Tognazzini. 2018. Blame. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Coeckelbergh, Mark. 2011. "Can we trust robots?" *Ethics and Information Technology* 14 (1):53-60. doi: 10.1007/s10676-011-9279-1.

Dietrich, Eric. 2007. "After the humans are gone Douglas Engelbart Keynote Address, North American Computers and Philosophy Conference Rensselaer Polytechnic Institute, August, 2006." *Journal of Experimental & Theoretical Artificial Intelligence* 19 (1):55-67. doi: 10.1080/09528130601115339.

Dietrich, Eric. 2011. "Homo Sapiens 2.0." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 531-538. Cambridge: Cambridge University Press.

Dworkin, Gerald. 2020. Paternalism. In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Epley, Kelly. 2019. "Emotions, Attitudes, and Reasons." *Pacific Philosophical Quarterly* 100 (1):256-282. doi: 10.1111/papq.12242.

Floridi, Luciano, & J. W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds*

and Machines 14 (3):349-379. doi: 10.1023/B:MIND.0000035461.63578.9d.
Fossa, Fabio. 2018. "Artificial moral agents: moral mentors or sensible tools?" *Ethics and Information Technology* 20 (2):115-126. doi: 10.1007/s10676-018-9451-y.
Graham, Peter A. 2014. "A Sketch of a Theory of Moral Blameworthiness." *Philosophy and Phenomenological Research* 88 (2):388-409. doi: 10.1111/j.1933-1592.2012.00608.x.
Grodzinsky, Frances S., Keith W. Miller, & Marty J. Wolf. 2008. "The Ethics of Designing Artificial Agents." *Ethics and Information Technology* 10 (2-3):115-121. doi: 10.1007/s10676-008-9163-9.
Gunkel, David J. 2012. *The Machine Question: Critical Perspectives on Ai, Robots, and Ethics*: MIT Press.
Hakli, Raul, & Pekka Mäkelä. 2019. "Moral Responsibility of Robots and Hybrid Agents." *The Monist* 102 (2):259-275. doi: 10.1093/monist/onz009.
Hall, J. Storrs. 2011. "Ethics for Self-Improving Machines." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 512-523. Cambridge: Cambridge University Press.
Himma, Kenneth Einar. 2009. "Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology* 11 (1):19-29. doi: 10.1007/s10676-008-9167-5.
Johnson, Deborah G., & Mario Verdicchio. 2019. "AI, agency and responsibility: the VW fraud case and beyond." *AI & Society* 34:639-647. doi: 10.1007/s00146-017-0781-9.
Laukyte, Migle. 2016. "Artificial agents among us: Should we recognize them as agents proper?" *Ethics and Information Technology* 19 (1):1-17. doi: 10.1007/s10676-016-9411-3.
Proyas, Alex. 2004. I, Robot. United States: 20th Century Fox.
Scanlon, T. M. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Belknap Press.
Sharkey, Noel, & Amanda Sharkey. 2010. "The crying shame of robot nannies." *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 11 (2):161-190. doi: 10.1075/is.11.2.01sha.
Strawson, P. F. 1974. *Freedom and Resentment and Other Essays*. London: Routledge.
Tappolet, Christine. 2016. *Emotions, Value, and Agency*. Oxford: Oxford University Press.
Tavani, Herman T. 2014. "Levels of Trust in the Context of Machine Ethics." *Philosophy & Technology* 28 (1):75-90. doi: 10.1007/s13347-014-0165-8.
Tegmark, Maz. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf.
Watson, Gary. 2014. "Peter Strawson on Responsibility and Sociality." In *Oxford Studies in Agency and Responsibility*, edited by David Shoemaker and Neal Tognazzini, 15-32. Oxford: Oxford University Press.
Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Bostrom Nick and Milan M. Ćirković, 308-345. Oxford: Oxford University Press.