Minds and Machines (2021) 31:595–616
https://doi.org/10.1007/s11023-021-09579-2

Robot Autonomy vs. Human Autonomy: Social Robots, Artificial Intelligence (AI), and the Nature of Autonomy

Paul Formosa[1]

Received: 3 June 2020 / Accepted: 17 October 2021 / Published online: 25 October 2021

Abstract
Social robots are robots that can interact socially with humans. As social robots and the artificial intelligence (AI) that powers them becomes more advanced, they will likely take on more social and work roles. This has many important ethical implications. In this paper, we focus on one of the most central of these, the impacts that social robots can have on human autonomy. We argue that, due to their physical presence and social capacities, there is a strong potential for social robots to enhance human autonomy as well as several ways they can inhibit and disrespect it. We argue that social robots could improve human autonomy by helping us to achieve more valuable ends, make more authentic choices, and improve our autonomy competencies. We also argue that social robots have the potential to harm human autonomy by instead leading us to achieve fewer valuable ends ourselves, make less authentic choices, decrease our autonomy competencies, make our autonomy more vulnerable, and disrespect our autonomy. Whether the impacts of social robots on human autonomy are positive or negative overall will depend on the design, regulation, and use we make of social robots in the future.

Keywords Autonomy · Social robots · Artificial intelligence (AI) · Machine ethics · Artificial moral agents · Respect

1 Introduction

Social robots are robots that can appear to express and perceive human emotions and can communicate with us using "high-level dialogue and natural cues", such as gaze and gestures (Fosch-Villaronga et al., 2020, p. 441). The interactivity and receptivity of social robots can encourage humans to form social relationships with them. As social robots and the artificial intelligence (AI) that powers them becomes

✉ Paul Formosa
Paul.Formosa@mq.edu.au

[1] Department of Philosophy & Centre for Agency, Values and Ethics, Macquarie University, North Ryde, Australia

more advanced (Lutz et al., 2019), they will likely take on more social and work roles. This could include undertaking care work for children, the elderly and the sick, becoming our teachers and work colleagues and, eventually, our social companions, friends, and even sexual partners (Darling, 2018; Ferreira et al., 2017; Lin et al., 2012; Mackenzie, 2018; Pirhonen et al., 2020; Sparrow, 2017). These changing roles constitute a shift in our relationship with technology such as social robots from it being a *tool* that we use to achieve our ends to something that we regard as an *agent* that we interact with (Breazeal et al., 2004). This shift has many important ethical implications. In this paper, we focus on one of the most central of these, its impacts on our autonomy. The autonomy of AIs and social robots and the autonomy of humans are often seen as a zero-sum game: more autonomy for social robots by offloading decisions to them equals less autonomy for humans (Floridi & Cowls, 2019). But, as we shall see, the impacts of social robots on human autonomy are more varied and complex than such an analysis suggests. Given the importance of autonomy to our understanding of morality, it is essential that we think through these ethical issues before these impacts are widely felt.

The paper proceeds as follows. First, we set out briefly what is meant by autonomy in the philosophy and ethical AI literatures. This is important since it shows us both that autonomy can mean different things in different literatures and that there are substantive disagreements between different theories of autonomy. This matters since different theories of autonomy can have competing normative implications. With this background in place, we then justify our focus on social robots on the grounds that, given their physical presence and social abilities, they have the potential to have very significant impacts on our autonomy (Borenstein & Arkin, 2016). While some of the implications that social robots have for autonomy will also hold for AI and other forms of technology, not all these implications will hold or will hold to the same degree for these other forms of technology. We then demonstrate the ways in which social robots could enhance and respect, as well as inhibit and disrespect, the autonomy of their users. We identify three broad ways that social robots could improve our autonomy through leading to humans having: (1) more valuable ends; (2) improved autonomy competencies; and (3) more authentic choices. We also identify five ways that social robots could harm our autonomy through leading to humans having: (1) fewer valuable ends; (2) worse autonomy competencies; (3) less authentic choices; (4) greater autonomy vulnerability; and (5) their autonomy disrespected. While this list is not intended to be exhaustive, it is illustrative as it brings together for the first time a systematic analysis of the most important impacts of social robots on human autonomy. We show that whether the impacts of social robots are positive or negative overall for human autonomy will depend on the design, regulation, and use that we make of social robots in the future.

2 Human Autonomy

While a full analysis of the philosophical literature on autonomy is obviously beyond our scope here, it will prove useful to outline some of the most relevant features of that literature here as background for the detailed analysis that occurs later

in the paper. In its broadest sense, autonomy means "self-rule" (Darwall, 2006). Autonomy has been applied to political entities, institutions, machines, and persons who govern themselves. However, over time the focus of autonomy has shifted (Darwall, 2006). Initially, autonomy in ancient Greek thought was primarily used as a *political* term referring to states or cities that govern themselves (Schneewind, 1998, p. 3; Formosa, 2017). Only later, with Kant, does autonomy come to be linked to the independence of practical reason and the freedom that reason grants *persons* to govern themselves independently of obedience to others, including the state (Schneewind, 1998). Kant's concept of *personal* autonomy as rational self-government has, in turn, become increasingly expanded beyond adherence to universal law to include governing yourself according to your "own" authentic desires and impulses (O'Neill, 2002, p. 31). Our focus here will be on this expanded sense of personal autonomy.

There are many different theories of personal autonomy in the philosophical literature (Anderson et al., 2005). While the detailed differences between these theories are not relevant here, the broad structure of the different types of theories of autonomy will prove important since it influences how we conceptualise the impacts that social robots have on human autonomy. Broadly, we can group contemporary theories of personal autonomy into *procedural* or *substantive* theories (Formosa, 2013; Mackenzie & Stoljar, 2000). We shall briefly consider each type in turn.

Procedural theories hold that content-neutral procedures provide necessary and sufficient conditions for autonomy. For example, Frankfurt's (1971) well-known account of autonomy holds that an action is autonomous if you act on desires that you desire to have. Here we have a *procedure*, namely that the first-order desires that you act on are desires that you desire to have (i.e. your second-order desires are effective), that determines whether you are autonomous, while saying nothing about the *content* of your desires (either first or second order). Other procedural theories give a different account of the procedures that determine whether you are autonomous, such as Watson's (1975) account which focuses on acting from reflectively endorsed values or Christman's (2004) account which focuses on acting from values that you would not revise were you to become aware of the influences underwriting those values. What matters for such theories is the procedure you follow in deciding what to do, not the content of your decisions or values.

Critics of procedural theories argue that they struggle to deal with the problem of "oppressive socialisation", that is, forms of socialisation that "impede the autonomy of the persons" that undergo it by undermining their "normative competence" at assessing norms for themselves (Benson, 1991, p. 406; Mackenzie & Stoljar, 2000, p. 20). This is a problem for procedural theories since their focus on content-neutral procedures makes it difficult for them to deal with cases where a person comes to reflectively endorse substantively flawed desires or norms. This is seen in the much-discussed case of the 1950s housewife who endorses the sexist and heteronomous norms that a woman should be subservient and under the control of her husband because her oppressive socialisation has left her unable to assess the falsity of such norms (Benson, 1991). This problem is related to the issue of adaptive preferences, whereby people can adapt their preferences to suit poor or unjust circumstances (Begon, 2015).

To deal with these concerns, substantive theories of autonomy hold that persons act autonomously if they act from the right set of values or in accordance with true or valid norms (Stoljar, 2000). For example, on some Kantian views, persons are autonomous when they act from the endorsement of the absolute value of the dignity of all rational agents (Formosa, 2013) or from the practical identity of themselves as an equal lawgiving member of the kingdom of ends (Korsgaard, 1996). The content of your decisions and values matters for such views, not merely the procedure you follow. Substantive accounts can avoid the problem of oppressive socialisation since they can claim that oppressive norms are false (Stoljar, 2000) or are incompatible with the dignity of all rational agents. However, critics of such theories worry that they can struggle to justify *which* substantive values or norms are the right ones (Formosa, 2013).

Both substantive and procedural theories of personal autonomy typically differentiate between *competency* and *authenticity* conditions (Christman, 2009; Susser et al., 2019). Authenticity conditions require that the values and desires that you act on are *really your own*, and not those that result from manipulation, oppression, subservience, undue external influence, or coercion. Competency conditions point to the fact that to be autonomous you must be able to *do* various things and have certain *skills* and *self-attitudes* (Meyers, 1987), such as being able to critically reflect on your values, adopt ends, imagine yourself being otherwise, and regard yourself as the bearer of dignity authorised to set your own ends. Several self-attitudes, such as self-respect, self-love, self-esteem, and self-trust, are also seen as important autonomy competencies (Benson, 1994; Mackenzie, 2008). These are seen as important because if you are to regard yourself as self-governing, then you need to be able to have respect for your powers of rational agency (self-respect), hold that your ends are worthwhile (self-love), trust that you can do what you set out to do (self-trust), and think of yourself as having worth as a person (self-esteem). Oppressive socialisation works by inhibiting the development of these competencies by, for example, lowering the esteem in which you hold your own worth as a person (Benson, 1991; Mackenzie & Stoljar, 2000). Oppression can also undermine the authenticity of our choices by leading us to hold values and norms that are the result of undue external social pressures and are thus not really our "own" (Friedman, 1986). In contrast, positive patterns of intersubjective recognition can help to bolster these vulnerable self-attitudes (Mackenzie, 2008) and help us to develop values and norms that are authentically our own.

Autonomy can also be diminished and empowered through the quality of the choices available to us. This is clearly illustrated through Raz's (1986) "*Man in the Pit*" example, where a man is stuck alone in a dark pit with a choice between eating, sleeping, or scratching his left ear now or a little later. Raz's man in the pit lacks autonomy because he lacks an "adequacy of options" (Raz, 1986, p. 373). When we are given more control over important aspects of our lives and access to a diverse range of meaningful choices, then our autonomy is increased. Further, having some degree of control over *how* our choices are realised, and not being subject to excessive oversight or control in their pursuit, is also important for our sense of autonomy (Ryan et al., 2006; Ryan & Deci, 2017).

As well as a capacity for self-governance, autonomy is also understood as a moral *principle*. This is clearest in the Kantian tradition (Kant, 1996), but is also present in various forms of principlism (Shea, 2020). For example, the 1974 Belmont Report on the ethical treatment of research subjects (via the "Respect for Persons" principle), the highly influential four principles of Beauchamp and Childress (2001) and its more recent extension by the AI4People framework (Floridi et al., 2018), all list autonomy as a basic ethical principle. Here autonomy is understood as something that ought to be respected, and that requires a focus on the consent of persons (Beauchamp & DeGrazia, 2004). Autonomy as a moral principle also speaks to the dangers of paternalism on the grounds that it disrespects autonomy through bypassing the consent of others (Scoccia, 1990).

Drawing these points together, we can say that human autonomy depends on the development and maintenance of a range of autonomy competencies. Autonomy also depends on having access to a sufficient range of meaningful options across important areas of life and being able to act freely on non-oppressive norms and values that are authentically our own without excessive oversight. Further, human autonomy is something that should be respected. All these aspects are important to consider when assessing the multifaceted impacts that social robots can have on human autonomy.

3 Machine Autonomy and AI

Machine autonomy can be understood as "the ability of a computer to follow a complex algorithm in response to environmental inputs, independently of real-time human input" (Etzioni & Etzioni, 2016, p. 149). More advanced forms of machine autonomy typically depend upon the use of AI. Although there are many competing definitions of AI, we shall understand it here to be creating information-processing systems that can do things which we would typically classify as intelligent were a human to do them, such as reason, plan, solve problems, categorise, adapt to its environment, and learn from experience (for discussion see Wang, 2019). Machine autonomy comes in degrees. The more responsive machines are to a greater range of environmental inputs and the greater range of conditions in which machines can act, reason, and choose independently of real-time human input, the higher is their degree of autonomy.

The issue of machine (or artificial) autonomy is of central importance to much of the recent literature on ethical AI, as demonstrated by three recent reviews by Floridi and Cowls (2019), Hagendorff (2020), and Jobin et al. (2019). Floridi and Cowls (2019) conceptualise the issue of autonomy as one where humans offload decision-making powers to AI, and they worry that "the growth in artificial autonomy may undermine the flourishing of human autonomy" (Floridi & Cowls, 2019, p. 7). On this analysis, if humans delegate a decision to an AI, then humans lose some autonomy and the AI gains some autonomy. Hagendorff (2020) instead takes human autonomy to refer in AI ethical guidelines to people being treated with respect as individuals, and he notes the tension between the need for AI to train on large data sets and the importance of not treating humans merely as sources of data. Further,

he also identifies the ways that AI can be a threat to human autonomy by manipulating users through "micro targeting, nudging, [and] UX-design" (Hagendorff, 2020). Jobin et al. (2019) undertake an exhaustive review of ethical guidelines for AI in the grey literature (i.e. non-academic sources such as government reports). They find that autonomy is used in these guidelines to refer to both "positive freedom", including the freedom *to* self-determination and to withdraw consent, and "negative freedom", including the freedom *from* manipulation and surveillance. Autonomy is to be promoted through transparency, maintaining broad option sets, increasing knowledge of AI, requiring informed consent, and limiting data collection (Jobin et al., 2019).

Clearly, there are both strong overlaps and important differences in how machine autonomy is understood in contrast to human autonomy. For both humans and machines, autonomy is a matter of self-governing across a range of significant choices in various contexts, and thus increasing the capacity to self-govern across a greater range of contexts and actions increases autonomy. Further, for both humans and machines, taking on significant choices increases autonomy and offloading significant choices to others decreases autonomy. In contrast, concerns about nudging, manipulation and surveillance apply to human autonomy only. Further, when autonomy is understood as a moral principle, there is a clear imperative to respect the autonomy of humans, which requires their consent, that does not apply to respecting the autonomy of machines, because the former and not the latter (for now, at least) are moral agents. Whether social robots or AIs could ever become persons or moral agents are further questions beyond our scope (but for discussion see, for example, Gunkel, 2020; Sparrow, 2012; Fosch-Villaronga et al., 2020).

4 The Impacts of Social Robots on Human Autonomy

While all forms of technology can impact human behaviour, we focus in this paper on the impacts on human autonomy of advanced social robots since these impacts are likely to be particularly significant (Bankins & Formosa, 2020). Given the lack at present or in the near future (see Bostrom, 2014) of Artificial General Intelligence (AGI), that is, AI that matches human-level performance across all relevant human abilities (Walsh et al., 2019, p. 16), we focus here only on social robots powered by Artificial Narrow Intelligence (ANI), that is, AI that is specialised to work only in specific areas (Gurkaynak et al., 2016). This means that we only consider instances of social robots being given limited machine autonomy in specific contexts, rather than general-purpose autonomy in every context.

Breazeal (2003, p. 167) defines social robots as the "class of robots that people anthropomorphise in order to interact with them". The Computers as Social Actors (CASA) paradigm (Reeves & Nass, 1996) suggests that humans tend to act *as if* computers and other forms of technology, such as social robots, are agents (or "social actors") and not mere things. This leads humans to interact with technology by following the same social scripts, schemas, and rules, such as norms of politeness and reciprocity, that are used in human–human interactions (Reeves & Nass, 1996; for an updated review of CASA see Gambino et al., 2020). This helps to explain the

human tendency to anthropomorphise technology by attributing human qualities and characteristics, such as motivations, intentions, and emotions, to non-human entities and inanimate objects (Epley et al., 2007; Fossa, 2018; Turkle, 2012).

However, while the tendency to anthropomorphise technology applies beyond social robots, it has been shown that the more socially interactive and human-like the robot is, the stronger is the tendency to anthropomorphise it (Fink, 2012). The social interactivity of social robots makes them "relational artifacts" that "present themselves as having 'states of mind'" for their human partners to engage with (Turkle et al., 2006, p. 347). This transforms our perception of social robots from *tools* that we use, into *agents* that we interact with in socially intuitive ways (Breazeal et al., 2004). Of course, this does not mean that social robots really *are* moral agents deserving moral respect, but it does mean that humans will tend to treat social robots *as if* they are agents. The use by social robots of verbal and non-verbal cues, such as gaze direction, and emotional receptivity aids this outcome. Drawing on Breazeal's (2003, p. 169) work, we can see that social robots come in various degrees of sophistication, from simple "*socially evocative*" robots such as robotic pets, to "*social interface*" robots which can use "human-like social cues and communication modalities", to "*socially receptive*" robots that are receptive to human social cues, and finally "*sociable*" robots that have their own internal goals and "model people in social and cognitive terms in order to interact with them". Our focus will primarily, but not exclusively, be on social robots on the more sophisticated end of this spectrum. We are therefore mainly thinking here about social robots that are "more sophisticated (but still non-sentient) versions of the [social] robots that we can build today" (Sparrow, 2017, p. 468), that have advanced motor, social and emotional skills, and can draw on "empathetic technology" and "extensive knowledge of our preferences" to "tailor their behaviours" toward us (Bankins & Formosa, 2020, p. 3). The social interactivity and physical presence of such sophisticated social robots makes their potential impacts on human autonomy very large, and this justifies our focus on them in this paper.

Given the importance for the discussion of autonomy of the delegation of decisions from humans to robots, we need to conceptualise the different ways this might occur. One commonly used way to describe that is through the language of a human *in*, *on* or *out of* the decision-making loop. The notion of a "*human-in-the-loop*" design has been used in a number of ways across several fields, from human–computer interaction (HCI), human–robot interaction (HRI), machine learning (ML) (Rahwan, 2018), and in the military context to discuss autonomous weapons systems (Schmitt & Thurnher, 2013; Sparrow, 2016; Walsh et al., 2019). Drawing on this literature, we can define a human in-the-loop design as one where a human *must* decide what a robot will do (e.g. a robot offers options but does not act until a human tells it which option to undertake); an on-the-loop design as one where a human *may* decide what a robot will do (e.g. a robot offers options but will act on its own if a human does not tell it which option to undertake); and an out-of-the-loop design as one where a human *cannot* decide what a robot will do (e.g. a robot independently acts on a certain option with no scope for human input). In the context of social robots, a similar distinction has been made between "opt in", "opt out" and "no way out" pathways (Borenstein & Arkin, 2016, p. 42) that approximates respectively the

human *in*, *on*, and *out of* the loop distinction. Given its existing use in the context of social robots, we will adopt this language here.

To see the differences between these three pathways, consider the following example. Imagine a simple social robot that can offer advice about what clothes you should buy, but only does so if you explicitly ask for that advice or "opt in" to that service (in-the-loop). However, the social robot will automatically call emergency services if it thinks that you have fallen over unless you explicitly tell it not to or "opt out" within 10 s (on-the-loop). The social robot also has a GPS tracker that sends back its location at regular intervals to its manufacturer and the user has "no way out" of this tracking (out-of-the-loop). Both "opt in" (once opted in) and "opt out" pathways can operate at the level of decision support mechanisms as they leave the decision to the human user who remains part of the decision loop. In contrast, the "no way out" pathway removes the human from the decision-making loop, granting the machine full autonomy to undertake the action itself. While there may be more complex ways to make this distinction (such as differentiating between automating information provision, information analysis, and decision options; for discussion, see Lyell et. al., 2021), this simple tripartite model will suffice for our purposes. However, the practical differences here might be blurred given the presence of the "automation bias", which is the "tendency [of humans] to over-rely on automation" (Goddard et al., 2012, p. 121). Even if humans remain *formally* part of the decision loop, they may be biased towards always uncritically following the machine's advice, which *practically* means that they are allowing the machine to act with little or no human oversight (as in a "no way out" design).

4.1 Social Robots as Autonomy Enhancers

Drawing on the above discussion, we argue that there are at least three broad ways that our autonomy could be enhanced by social robots. We can summarise these as, through the assistance of social robots, humans can achieve: (1) *more valuable ends*; (2) *improved autonomy competencies*; and (3) *more authentic choices*. These are important cases as they counteract the common view that more autonomy for machines means less autonomy for humans. Consider the example of Corti, an AI-powered machine that informs emergency call responders whether the caller is at risk of a heart attack through using machine learning to analyse breathing and speech patterns (van Wynsberghe & Robbins, 2019). Although Corti is not a social robot, we could easily imagine a "robotic triage nurse" with similar functions (Asaro, 2006, p. 14). Corti is implemented as an "opt in" design as it merely advises a human operator who must choose whether to act on its advice. But if we instead delegate to Corti the decision whether to send an ambulance to someone through a "no way out" design, then we have seemingly increased Corti's autonomy (since it can act independently in a greater range of cases) *by* decreasing the human operator's autonomy (since they no longer make an important decision for themselves). This transforms Corti from what is known in the medical AI literature as a "decision support" into an "autonomous decision" technology (Lyell et al., 2021; Rogers et al., 2021). This makes human and machine autonomy seem like a zero-sum game, with

more for one meaning less for the other. But, as the below discussion shows, this is not always the case.

4.1.1 More Valuable Ends

First, we can increase a person's autonomy through giving them access to a suitably broad range of valuable ends. We can do that through giving people access either to a greater number of valuable ends or to ends that are more valuable. Social robots can help in both regards either by undertaking the means to ends set by humans or by setting lower value ends for humans on their behalf. In the first case, imagine an elderly woman called Sally who is unable to move around by herself. Sally would like a cup of tea to drink while she reads her novel, but she cannot adopt that end as she cannot move around by herself. However, one day Sally acquires a social robot who can assist her. As before, Sally would like a cup of tea to drink while she reads her novel, but now she can ask her social robot to make it and bring it to her, which it does while Sally continues to read. Sally has more autonomy because she can now set a valuable end, that of drinking a cup of tea while reading, which she could not otherwise set without (in this example) the help of her social robot. (Clearly, this example extends to many other cases of robots helping people to overcome restricted functional abilities—see Pirhonen et al. (2020). Further, many simpler forms of technology, such as walking aids, can also help people to set ends they otherwise could not). In the second case, imagine a businessman called Sam who has a social robot designed to be proactively helpful to him. After examining Sam's schedule, his social robot proactively selects an appropriate shirt and tie for a business meeting that Sam has that morning (for an example of this sort of social robot see Woiceshyn et al., 2017) and brings the clothes to Sam at the exact moment it calculates that he will need them to get dressed to make his meeting on time. Sam is thankful for not having to spend time selecting his clothes for the day. After getting dressed, he hops into the taxi his robot has ordered for him so that he arrives exactly 5 min before his meeting, since his robot knows he always likes to be a few minutes early to meetings. Sam uses the time his robot's proactive actions have gained him to read important documents that he wants to get through. Sam has more autonomy because he can now set a valuable end, that of reading important documents before his meeting, that he could not otherwise set without (in this example) the help of his social robot.

In Sally's case, the social robot undertakes the *means* to ends that are set by a human. This is an "opt in" design. In Sam's case, the social robot proactively sets *ends* for a human so that the human can set other ends. This is an "opt out" or "no way out" design, depending on the implementation. However, in both cases we do not have, for utility gains, a loss of human autonomy through a gain in machine autonomy. Instead, we have a *gain in human autonomy* (i.e. more meaningful choices for a human) *through more machine autonomy* (i.e. by delegating less important choices to a machine). To see why, consider the tea-making social robot in Sally's case. While, in terms of starting the tea-making process, this is an "opt in" design, there are still many *other* sub-decisions that are delegated to Sally's social robot and thus which constitutes a "no way out" design in this regard, such as the

decision about how to safely navigate the room without stepping on the cat's tail or spilling the tea. Compare this to a tea-making robot that lacks all autonomy, which would make it a simple remote-controlled device (or "telepresence robot") unable to move by itself (Pirhonen et al., 2020). This design gives the human user more control over how the robot navigates the room, but this comes at the cost of making the robot far less useful. A simple way of reading this trade-off is: more autonomy for humans but a less useful machine, or less autonomy for humans but a more useful machine. But this is an overly simplistic analysis, as we shall see.

Autonomy is (in part) about freely choosing to do the valuable things that we authentically want to do. If Sally must spend her time remote controlling a robot across a room, rather than reading the novel which she really wants to read, then having more control over the robot means *less autonomy* for Sally as she is *forced* to do something that she does not value highly (i.e. remote controlling a robot) to get something else she really wants (i.e. a cup of tea to drink while reading her novel). In contrast, if Sally delegates the task of room navigation to the robot, thereby giving it more autonomy, then Sally is also *more autonomous* as she can instead spend her time *freely* doing what she really wants to do (i.e. reading the novel while the cup of tea is made for her). There is also some evidence to suggest that Sally will feel more autonomous due to the independence her robot gives her (Pirhonen et al., 2020). Likewise, Sam gains greater autonomy by delegating to his social robot the setting of what he regards as the less valuable ends of selecting which shirt and tie to wear and how to get to his meeting on time, since this allows him to pursue more valuable ends, in this case reading documents for his work meeting, that he really wants to do instead. In both cases, more machine autonomy leads to more human autonomy, not less, by giving Sally and Sam more time to do what they value most highly through the offloading of less valuable choices to their social robots. But this does not mean, as we shall see in the next section, that we can offload *every* difficult task or important decision to machines without loss to our autonomy.

4.1.2 Improved Autonomy Competencies

Second, social robots can also increase a person's autonomy by helping them to build, maintain, and develop their autonomy competencies. A social robot could do this through either *indirect* or *direct* assistance. In the case of indirect assistance, a social robot indirectly frees up a person's time and attention resources through undertaking less valued tasks for them. This gives that person the time and space they would not otherwise have had to develop their autonomy competencies themselves. Imagine a variation of the previous examples where a person offloads mundane tasks, such as making tea or booking a taxi, to a social robot so that they can directly cultivate their autonomy competencies themselves by, for example, reading a book on critical reasoning or talking to an encouraging friend which boosts their self-esteem. Here the social robot helps to facilitate autonomy competency development that might not otherwise have been possible. (In this case, other time saving forms of technology could have similar impacts). In the more interesting case of direct assistance, a social robot could directly increase a person's autonomy competencies through positive social interactions with them. Here the social interactivity

of this technology is crucial. If humans can develop, maintain, and cultivate their autonomy competencies through positive social interactions with each other that bolster attitudes such as self-respect, self-love, and self-trust (Mackenzie, 2008), then something similar should be possible with advanced social robots (Pirhonen et al., 2020). There is some evidence to support this claim. For example, a systematic review of the use of social robots among older adults found a lack of high-quality studies but some indications that social robots can reduce agitation, anxiety, and loneliness (Pu et al., 2019), which could in turn boost relevant autonomy competencies such as self-esteem. Similar positive impacts have been found in other populations (Jeong et al., 2015). Another study showed that social rejection by a robot can lower self-esteem relative to social acceptance by a social robot or a control condition (Nash et al., 2018).

These positive and negative impacts will likely be due, in part, to the human tendency to anthropomorphise social robots by regarding them as social agents who have "states of mind", including attitudes toward us, that develop through our intuitive social interactions with them (Breazeal et al., 2004; Fossa, 2018; Turkle, 2012). For example, by seeming to regard you as a source of normative authority about what ought to be done, a social robot might be able to help foster your self-respect. Likewise, a social robot that seems to regard you and your ends as valuable by taking the initiative to proactively help you to achieve your ends might help to foster your self-love and self-esteem. By encouraging you, a social robot may also help you to develop self-trust. These positive social outcomes could be strengthened through the social robot's use of gestures, tone of voice, eye contact, expression of (what appears to be) emotions such as sympathy, and physically embodied presence (Borenstein & Arkin, 2016; Li, 2013; Moshkina et al., 2011). Insofar as these positive outcomes can be achieved, social robots could directly improve our autonomy competencies.

4.1.3 More Authentic Choices

Third, we can increase a person's autonomy by helping them to make more authentic choices, both in the sense of *more choices* that are authentic and choices that are *more authentic*. A social robot could use its social interactivity to help to achieve this outcome in several ways. A choice is authentic if one acts "on motives, desires, preferences and other reasons" that are "one's own", and they count as "one's own" when, on reflection, one endorses or acknowledges them (Walker & Mackenzie, 2020, p. 8). The more a choice is "one's own" in this sense, the more authentic it is. However, measures of authenticity differ between substantive and procedural theories of autonomy.

On strong substantive views, a choice is more authentic the more it reflects the right values (Wolf, 1990) or norms (Stoljar, 2000), since it is only when we act on such values or norms that we correctly grasp moral reality and act authentically as the moral beings we are. Of course, as noted above, such views suffer from the difficulty of justifying what are the right values or norms. In any case, according to such views, social robots that help us to avoid acting from the wrong values or norms thereby help us to make more authentic choices by better connecting us with moral

reality and our authentic moral selves. We can see how social robots might bring about this outcome by examining the way that some social robots are designed to shut down or resist abusive interactions (Turkle, 2012). For example, the "robotic dinosaur Pleo cries out as though it is experiencing pain if pushed over or otherwise 'mistreated'" (Borenstein & Arkin, 2016, p. 42). Generalizing, a social robot could be designed to use such behaviours to encourage us to make (what counts on a strong substantive view as) more authentic choices. For example, if you propose to commit a crime with the assistance of your social robot or attempt to violently assault your social robot (see Darling, 2018), then it could refuse to help you by shutting down or it could cry out in pain to stop you on the grounds that you are acting in an abusive and therefore inauthentic manner. However, social robots that actively resist poor treatment can create their own ethical difficulties, especially regarding "realistic female [sex] robots" because some users may use the robot's refusal of consent to experiment with "rape fantasy" (Sparrow, 2017, p. 465). Therefore, careful consideration of context and design is required to ensure that robot refusals encourage authentic moral behaviours rather than fuel immoral fantasies.

On procedural views, a choice is more authentic if it follows from the right sort of procedures, such as informed critical reflection. According to such views, social robots could help us to make more authentic choices by helping us to do better at critically reflecting on our choices and values. For example, imagine a social robot with "empathetic technology" that can identify a person's emotional state through analysing their facial features, speech, and the levels of carbon dioxide on their breath (Seïler & Craig, 2016; Wakefield, 2018). Using this technology, a social robot could detect that a person is overcome with extremely strong emotions when they issue a command that could have serious implications for themselves and others. The social robot could then refuse to undertake that command for a certain period of time to give the person space to calm down and activate their critical reflection skills. Alternatively, a social robot could draw on relevant research about biases that impact human thinking (Kahneman, 2011), and evidence that people are more open to critical reflection after positive self-affirmation (von Hippel & Trivers, 2011), to first bolster a person's sense of self-worth before alerting them to potential biases it has identified in their reasoning that might be preventing them from choosing what they would authentically want to choose. A social robot could also act as an interlocuter and help a person to consider the pros and cons of an important choice, provide information that it has identified as relevant to their choice to help to ensure that their choice is properly informed, alert them to the presence of past oppression that could be unduly influencing their choice without them knowing it, and keep them updated with changing information.

Many of these imaginary interventions by a social robot constitute examples of "nudging" a human to be more autonomous (Thaler & Sunstein, 2008). Drawing on dual process theory (Evans, 2008), Thaler and Sunstein describe two types of nudges, those that impact on our "Automatic System", such as placing the item we wish to nudge someone towards at eye level, and those that impact on our "Reflective System", such as nudges that encourage us to think carefully about something (Borenstein & Arkin, 2016; Thaler & Sunstein, 2008). The examples that we looked at in the previous paragraph involve Reflective System prompts to engage in

processes that promote autonomy, such as informed critical reflection and the avoidance of unconscious biases. But nudging can also seek to influence us via our Automatic System. Reflective System nudges are less ethically worrisome, since they merely seek to encourage and inform autonomous self-reflection, whereas Automatic System nudges bypass critical self-reflection through unconscious influences aimed at paternalistically achieving a certain outcome. While there might still be good all-things-considered reasons for the latter type of nudges, such as opting people in automatically to socially beneficial programs rather than explaining to them the good reasons they have to opt in, the ethical issues involved in this type of nudging are more complicated (for discussion, see Schmidt & Engelen, 2020) and raise significant ethical concerns about paternalism. As such, while robotic nudges via our Reflective System (as focused on in this section) could aid our autonomy, similar nudges via our Automatic System may limit it (as we shall see in the next section).

4.2 Social Robots as Autonomy Inhibitors

The previous section focuses on the positives for autonomy. But it is not hard to see the negatives too. We can use the inverse of the three categories outlined above to group these worries. We can summarise these as, through the impacts of social robots, humans can have: (1) *fewer valuable ends*; (2) *worse autonomy competencies*; and (3) *less authentic choices*. But there are also other potential problems, including: (4) *making human autonomy more vulnerable*; and (5) *disrespecting human autonomy*. Again, this list is meant to be illustrative, not exhaustive.

4.2.1 Fewer Valuable Ends

First, social robots could reduce our autonomy if it means that we set and achieve *fewer valuable ends* ourselves. As we saw above, when we offload unimportant means to our ends or offload unimportant ends to social robots, then our autonomy may be enhanced. By contrast, when we offload decisions to social robots about important ends or offload the undertaking of important means that are integral to the achievement of valuable ends, then our autonomy can be diminished. For example, if a social robot autonomously decides on your behalf (through an "opt out" or "no way out" design) whether to notify you of an incoming phone call or whether to accept a calendar invite based on *its* (and *not your*) view of the perceived importance of the caller or inviter, then you lose some autonomy as you can no longer make the important choice of whether to answer a phone call or accept a meeting invite yourself. Less realistically but more troubling, a social robot could start to decide for you who you will date by using a dating app on your behalf after analysing your past dating experiences and preferences or decide on your behalf which school your child should attend after analysing school performance data and your child's learning preferences. (Some of the examples in this section clearly apply to AI in general rather than social robots in particular). Even if you explicitly "opt in" to having a social robot make these decisions on your behalf, you still lose some autonomy because handing over such significant

life decisions to a robot means that you have less control over important aspects of your life. This makes you less autonomous, even if you "opt in" to it and even if the decision is justified on other ethical grounds, such as the quality of the resulting robotic decision. This point is related to the issue of whether we should offload moral decisions to AI or artificial moral agents, since moral decisions are clearly important decisions (Robbins, 2019; Sparrow, 2016; van Wynsberghe & Robbins, 2018). While there may be good all-things-considered ethical reasons, such as better outcomes or the existence of time constraints, for offloading some important ethical decisions to an AI or social robot (Formosa & Ryan, 2020), there is also a clear cost to our autonomy in doing so that must be considered.

4.2.2 Worse Autonomy Competencies

Second, social robots could reduce our autonomy by resulting in us having lower levels of autonomy competencies. This could occur because they harm the *development* of autonomy competencies in children, or they harm the *maintenance* and *cultivation* of them in adults. Due to their physical presence, social robots have been shown to be effective in achieving positive educational outcomes for children (Belpaeme et al., 2018; Kanero et al., 2018). But do teaching interactions with social robots also help children to develop autonomy competencies? If it turns out that robots are less effective, as they are in other areas, at developing such competencies in children compared to skilled human teachers (Kanero et al., 2018), and if social robots take on more education and caring roles, then this could lead to children developing lower levels of autonomy competencies than they would through skilled human teaching (although there is a general lack of evidence in this regard; see Pashevich, 2021). Of course, this assumes that skilled human teaching is available, and where it is not, then robot teaching may be better than the alternatives. In terms of skill maintenance and cultivation in adults, Vallor (2015) raises the related worry of "moral deskilling". In its general form, this worry is that when we offload tasks to technology, then we start to lose or degrade the relevant skills, including autonomy competencies, needed to complete the offloaded task. For example, if you become dependent on a social robot to make most decisions for you or to tell you what to do, then you may start to lose trust in your ability to get things done by yourself and your skills at making decisions could start to dissipate. Further, interpersonal skills are often essential to realising our ends, since achieving many ends requires complex social cooperation. But if we get used primarily to interacting with social robots, then we may start to lose our human-to-human interpersonal social skills. Similarly, if we get used to interacting with social robots that do not demand equal reciprocity in terms of social exchange, then our skills at engaging in reciprocal social exchanges with humans could start to atrophy (Bankins & Formosa, 2020). If we use our autonomy competencies less because social robots do more things for us, then our autonomy competencies will likely deteriorate.

4.2.3 Less Authentic Choices

Third, social robots could reduce our autonomy by causing us to have less authentic choices, both in the sense of *fewer choices* that are authentic and choices that are *less authentic*. There are several ways this could happen. One of the reasons that social robots have the potential to influence our behaviour is that we tend to regard them as social agents with states of mind and not mere tools (Breazeal et al., 2004). But this influence could also have negative impacts on our autonomy. For example, when we feel ourselves under surveillance and under the gaze of others, we can feel less able to act authentically and be who we really want to be (Molitorisz, 2020). This is compounded by the fact that we know that the AI and machine learning that will power social robots depends on large datasets, and we might worry that our social robot is really a surveillance machine sending our intimate personal data to its corporate creators (Hagendorff, 2020). This could make us act more self-consciously and less authentically in front of social robots, including by engaging in pre-emptive self-censorship, and given how deeply integrated into our lives social robots could become, this could deeply impair our autonomy. Social robots could also promote inauthenticity through the perpetuation of oppressive socialisation that reinforces unjust gender norms. For example, a UNESCO (2019) report shows that "female" AI assistants, such as Cortana, Siri, and Alexa, can perpetuate and reinforce norms that women should be servile and put up with abuse. A concrete example of this is that at one point Apple's Siri responded to "You're a slut" with "I'd blush if I could" (UNESCO, 2019, p. 107). Submissiveness in "female" social robots, created by largely male development teams (UNESCO, 2019), could thus help to perpetuate oppressive norms that can directly harm the autonomy of women and other minorities. The likely reliance of social robots on pretrained neural language models that are "prone to generating racist, sexist, or otherwise toxic language" could further exacerbate this problem (Gehman et al., 2020, p. 3356).

Another way that social robots could lead to less authentic choices is if they manipulate us. As happens in many online contexts, much of this manipulation could occur by targeting and exploiting the "decision-making vulnerabilities" of persons which can result in "autonomy harm" (Susser et al., 2019, p. 1). For example, a 2017 report exposed internal Facebook documents showing that through monitoring its users, Facebook could determine when teenagers were feeling insecure, stressed, or anxious, and it could in principle use this information (even if it in fact did not) to manipulate them to purchase items through carefully targeted advertising (Susser et al., 2019). Whereas the robotic nudges toward autonomy that we discussed in the previous section operate via our conscious Reflective System and seek to counteract biases, the manipulations highlighted here work by exploiting human biases and decision-making vulnerabilities in ways that we are not consciously aware of via our Automatic System. This manipulation can also occur through the careful presentation, filtering, and ordering of information that social robots pass on to us, since what is and is not shown or told to us, in what way, and in what order it is presented, can all have hidden influences on our choices. This can involve "nudging" people through careful design of the "choice architecture" or context within which choices are made (for further discussions of this extensive literature see, for example, Thaler

& Sunstein, 2008; Cohen, 2013; Quigley, 2013; Hansen & Jespersen, 2013). To the extent that these influences are exploited by social robots (or their creators) to get us to do what is in the commercial or political interests of its developers or advertisers, then the autonomy of users could be harmed and disrespected. This amounts to treating users as mere means to outcomes that others want them to choose, often for commercial or ideological reasons, rather than helping users to choose what they authentically want to do. While such manipulations through technology are hardly unique to social robots, the physically embodied nature of social robots means that these manipulative impacts could be greater than with other forms of technology.

4.2.4 Making Autonomy More Vulnerable

Fourth, social robots, as likely commercial products, could make our autonomy vulnerable in new ways and access to autonomy more precarious and unfair. Autonomy is not the same as independence, since dependency is a central feature of human life (Kittay, 1997), and most people autonomously choose to make themselves dependent on others, such as friends and family. Even so, if we become dependent on social robots for realising many of our ends or for social connection, then our autonomy becomes vulnerable in new ways. For example, our social robot might cease to work properly after a firmware update, which means that it becomes less able to help us to achieve our ends. Further, this could make our autonomy dependent on a company focused on profit (Hagendorff, 2020), rather than on friends and family who may genuinely care for us. In terms of access, social robots are likely, at least initially, to be very expensive, and this could create an underclass of people who have less access to the autonomy-enhancing features (outlined above) of social robots than the wealthy. While we already have such inequalities, since autonomy as substantive control over our lives requires access to resources that many people lack, it does create an important new area for this inequality to playout.

4.2.5 Disrespecting Autonomy

Fifth, social robots could be disrespectful towards our autonomy, and this is bad in its own right and could also lead to an erosion of our autonomy competencies if we internalise that disrespect (Formosa, 2013). Social robots that manipulate us use us as mere means. This constitutes disrespectful treatment, and the intentional design of such robots is an expression of disrespect by its creators. An example might be that of a social robot that knows that you are in a depressed state and uses that information to manipulatively encourage you to purchase an upgrade or other item. More generally, there might be something disrespectful about the very nature of social robots given that they "push our Darwinian buttons" by deceptively appearing to be "alive enough" (Turkle, 2012, p. 8, 18). Indeed, the effectiveness of social robots depends on their cultivating the illusion in humans that they have internal mental and emotional states that, in fact, they do not really have. Many worry that this deception is unethical (Lucidi & Nardi, 2018). To the extent that it is unethical, it also disrespects our autonomy as it manipulates us into having false beliefs about the inner life of social robots. Social robots may also be used to amass large

amounts of very personal data about us, given their potentially intimate presence in our lives (Lutz et al., 2019). This data gives corporations power over us which could be used to manipulate, pressure, and coerce us through social robots (Susser et al., 2019). Further, whether that intimate data could be obtained in a way that respects our autonomy is unclear. This points to the problem of what Nissenbaum (2011) calls the "transparency paradox". Our ability to autonomously consent to privacy policies is flawed, given that we either consent to something too simplistic to accurately represent data flows or we cannot understand the complex legalese of more detailed policies. Either way, informed autonomous consent is difficult to achieve, and given that social robots will likely harvest large amounts of very intimate data about us, the potential this has for expressing disrespect and limiting our ability to have autonomous control over our personal data is concerning (Hagendorff, 2020; Sharkey & Sharkey, 2012).

5 Discussion

When thinking about the ethical implications of the increasing use of sophisticated social robots, it is important that we consider both their potential positive and negative impacts on our autonomy. From the above analysis, a few key issues emerge. Before we consider these from multiple perspectives, two points are worth noting.

First, our focus here is on autonomy only. But there are other relevant ethical issues at play, such as beneficence and justice, and as noted above the ethical issues raised by autonomy need to be balanced against these competing ethical concerns. Second, the range of possible responses to the ethical issues raised here include improved user education and public awareness, design considerations, ethical guidelines, industry standards, government agencies, and regulatory or legal frameworks (for discussion see Fosch-Villaronga et al., 2020; Petit, 2017). This discussion needs to consider the related existing and proposed regulatory frameworks and guidelines around robotics, AI, and privacy that exist across different jurisdictions. Thus far, most regulation in this context has taken the form of voluntary ethical guidelines, although this is changing through the impact of Europe's GDPR (General Data Protection Regulation) in terms of privacy considerations and the emergence of standards such as the BS 8611:2016 Guide to the Ethical Design and Application of Robots and Robotic Systems and the IEEE Ethically Aligned Design 2017 (for an overview, see Fosch-Villaronga et al., 2020).

Further, the intensity of the regulatory response should be dependent on the degree and nature of the specific harms and externalities generated by social robots (Petit, 2017). There are dangers of both too little regulation, which can lead to user harms and a reluctance to embrace new technology, and of too much regulation, which can stifle innovation and prevent benefits and user choice. Given these complexities, rather than provide specific regulatory recommendations here, we shall instead focus on highlighting, from the perspectives of users, designers, and society more generally, the most significant ethical issues that emerge from the above analysis.

From the perspective of users, further education is an important goal. This should focus on how significant the choices being offloaded by users to their social robots are and how frequently that offloading is occurring, since this is under user control and has the potential to have both positive and negative impacts on their autonomy. The offloading of trivial or unimportant decisions promises to free up users' time and attention resources for more meaningful exercises and cultivation of their autonomy and related competencies. In contrast, offloading significant decisions to social robots can not only directly limit the control that users have over important aspects of their lives, but also lead to an atrophying of their autonomy competencies when such offloading occurs frequently. Further, users need to be educated about their bias to rely uncritically on technology such as social robots (Goddard et al., 2012), and the dangers to their autonomy that nudging from social robots can have. Finally, users also need to be educated about the privacy implications of using social robots and be prescient to the dangers of emotional manipulation by their social robots. This is a particular problem for children who need to be reminded that social robots do not really care about them or have feelings, even if they seem to (Turkle, 2012).

For designers, a particular focus should be on how users will perceive the attitudes that social robots will seem to express toward them, especially insofar as they impact important self-attitudes such as self-respect and self-esteem. This should also include a focus on the differing social impacts, among a variety of cultural and social groups, of differences in speech, tone, and facial expression by social robots. Further, given the importance of social acceptance or rejection for users, the ways that social robots express these types of social judgments must be considered carefully to minimise any potential autonomy harms, especially for vulnerable users. Users will perceive social robots as having emotions and states of mind, and designers should be careful to avoid, intentionally or unintentionally, using these responses to manipulate users in inappropriate ways. Designers should also seek to aid user autonomy through Reflective System nudges that encourage critical reflection and limit the use of Automatic System nudges that can potentially disrespect users' autonomy.

At a society level, beyond dealing with the issues already raised above and the broader existing regulatory frameworks around privacy, AI, and robotic safety (Fosch-Villaronga et al., 2020; Hagendorff, 2020), there are two further areas of focus worth mentioning here. These are how social robots respond to mistreatment and abusive behaviour (see Darling, 2016) and the potential impacts of social robots on perpetuating oppressive social norms that can inhibit human autonomy. These should be considered here because the ways that social robots in the *aggregate* respond to mistreatment and perpetuate existing norms will have broader consequences that should be considered at a social level. Dealing with these issues requires the input of a diverse group of stakeholders to ensure a variety of perspectives are considered. Industry guidance or examples of ethical best practice would be helpful in this regard. Finally, given their massive data collection potential, and their impacts on the physical and informational privacy of their users, social robots must be designed with user privacy in mind, and this is probably best dealt with at a regulatory level to ensure compliance (see Lutz et al., 2019).

6 Conclusion

Social robots have the potential to help their users to be more independent and autonomous and improve their autonomy competencies, but also the potential to manipulate, deskill, illicitly surveil, and disrespect their users' autonomy. Whether the impacts of social robots are positive or negative overall for human autonomy will depend on the design, regulation, and use that we make of social robots in the future. What is clear is that the potential impacts of social robots on human autonomy are profound and multifaceted. While the issues examined here are not exhaustive, we have provided a systematic analysis of the most important and relevant ethical considerations through highlighting both the potential positive and negative implications. This provides a useful theoretical foundation for further work examining the implications for human autonomy of social robots and AI more broadly.

Funding Open Access funding provided by the Macquarie University Research Centre for Agency, Values and Ethics (CAVE).

References

Anderson, J., Christman, J., & Anderson, J. (2005). *Autonomy and the challenges to liberalism*. Cambridge University Press.

Asaro, P. (2006). What should we want from a robot ethic? *International Review of Information Ethics, 6*, 9–16.

Bankins, S., & Formosa, P. (2020). When AI meets PC: Exploring the implications of workplace social robots and a human-robot psychological contract. *European Journal of Work and Organizational Psychology, 29*(2), 215–229.

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics*. Oxford University Press.

Beauchamp, T. L., & DeGrazia, D. (2004). Principles and principlism. In G. Khushf (Ed.), *Handbook of bioethics* (pp. 55–74). Springer.

Begon, J. (2015). What are adaptive preferences? *Journal of Applied Philosophy, 32*(3), 241–257.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*. https://doi.org/10.1126/scirobotics.aat5954

Benson, P. (1991). Autonomy and oppressive socialization. *Social Theory and Practice, XVI, I*(3), 385–408.

Benson, P. (1994). Free agency and self-worth. *Journal of Philosophy, 91*(12), 650–658.

Borenstein, J., & Arkin, R. (2016). Robotic nudges: The ethics of engineering a more socially just human being. *Science and Engineering Ethics, 22*(1), 31–46.

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems, 42*, 167–175.
Breazeal, C., Gray, J., Hoffman, G., & Berlin, M. (2004). Social robots: Beyond tools to partners. *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, pp. 551–556.
Calvo, R. A., Peters, D., & Vold, K. (forthcoming). Supporting human autonomy in AI systems. In C. Burr & L. Floridi (Eds.), *Ethics of Digital Well-Being*. Springer.
Christman, J. (2004). Relational autonomy, liberal individualism and the social constitution of selves. *Philosophical Studies, 117*, 143–164.
Christman, J. (2009). *The politics of persons: Individual autonomy and socio-historical selves*. Cambridge University Press.
Cohen, S. (2013). Nudging and informed consent. *The American Journal of Bioethics, 13*(6), 3–11.
Darling, K. (2016). Extending legal protection to social robots. In R. Calo, A. Froomkin, & I. Kerr (Eds.), *Robot law.* Edward Elgar.
Darling, K. (2018). Who's Johnny?' anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, G. Bekey, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0 (p. 22).* Oxford University Press.
Darwall, S. (2006). The value of autonomy and autonomy of the will. *Ethics, 116*, 263–284.
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886.
Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology, 18*(2), 149–156.
Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgement, and social cognition. *Annual Review of Psychology, 2008*(59), 255–278.
Ferreira, M. I. A., Sequeira, J. S., Tokhi, M. O., Kadar, E. E., & Virk, G. S. (Eds.). (2017). *A World with Robots: International Conference on Robot Ethics: ICRE 2015*. Springer.
Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human–robot interaction. In S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, & M.-A. Williams (Eds.), *Social robotics (Vol. 7621, pp. 199–208).* Springer.
Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*.
Floridi, L., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707.
Formosa, P. (2013). Kant's conception of personal autonomy. *Journal of Social Philosophy, 44*(3), 193–212.
Formosa, P. (2017). *Kantian ethics*. Cambridge University Press.
Formosa, P., & Ryan, M. (2020). Making moral machines. *AI & Society*. https://doi.org/10.1007/s00146-020-01089-6
Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Gathering expert opinions for social robots' ethical, legal, and societal concerns. *International Journal of Social Robotics, 12*(2), 441–458.
Fossa, F. (2018). Artificial moral agents: Moral mentors or sensible tools? *Ethics and Information Technology, 20*(2), 1–12.
Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*(1), 5–20.
Friedman, M. (1986). Autonomy and the split-level self. *Southern Journal of Philosophy, 24*(1), 19–35.
Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–86.
Gehman, S., et al. (2020). RealToxicityPrompts: evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics.
Goddard, K., Roudsari, A., & Wyatt, J. (2012). Automation bias. *Journal of the American Medical Informatics Association, 19*(1), 121–127.
Gunkel, D. J. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*(4), 307–320.
Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling artificial intelligence. *Computer Law & Security Review, 32*(5), 749–758.
Hagendorff, T. (2020). The ethics of Ai ethics: An evaluation of guidelines. *Minds and Machines*. https://doi.org/10.1007/s11023-020-09517-8

Hansen, P., & Jespersen, A. (2013). Nudge and the manipulation of choice. *European Journal of Risk Regulation, 4*(1), 3–28.
Jeong, S., et al. (2015). A Social Robot to Mitigate Stress, Anxiety, and Pain in Hospital Pediatric Care. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 103–104.
Jobin, A., Lenca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., & Göksun, T. (2018). Social robots for early language learning: Current evidence and future directions. *Child Development Perspectives, 12*(3), 146–151.
Kant, I. (1996). Groundwork of the metaphysics of morals. In M. J. Gregor (Ed.), *Practical philosophy* (pp. 37–108). Cambridge University Press.
Kittay, E. F. (1997). Human dependency and Rawlsian equality. In D. Meyers (Ed.), *Feminists rethink the self*. Westview Press.
Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge University Press.
Li, J. (2013). The nature of the bots. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction—ICMI '13*, pp. 337–340.
Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2012). *Robot ethics*. MIT Press.
Lucidi, P. B., & Nardi, D. (2018). Companion Robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 17–22.
Lutz, C., Schöttler, M., & Hoffmann, C. (2019). The privacy implications of social robots. *Mobile Media & Communication, 7*(3), 412–434.
Lyell, D., Coiera, E., Chen, J., Shah, P., & Magrabi, F. (2021). How machine learning is embedded to support clinician decision making: An analysis of FDA-approved medical devices. *BMJ Health & Care Informatics, 28*(1), e100301. https://doi.org/10.1136/bmjhci-2020-100301
Mackenzie, C. (2008). Relational autonomy, normative authority and perfectionism. *Journal of Social Philosophy, 39*(4), 512–533.
Mackenzie, C., & Stoljar, N. (Eds.). (2000). *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford University Press.
Mackenzie, R. (2018). Sexbots: sex slaves, vulnerable others or perfect partners? *International Journal of Technoethics, 9*(1), 1–17.
Meyers, D. (1987). Personal autonomy and the paradox of feminine socialization. *Journal of Philosophy, 84*(11), 619–628.
Molitorisz, S. (2020). *Net privacy*. NewSouth Publishing.
Moshkina, L., Park, S., Arkin, R. C., Lee, J. K., & Jung, H. (2011). TAME: Time-varying affective response for humanoid robots. *International Journal of Social Robotics, 3*(3), 207–221.
Nash, K., Lea, J. M., Davies, T., & Yogeeswaran, K. (2018). The bionic blues: Robot rejection lowers self-esteem. *Computers in Human Behavior, 78*, 59–63.
Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus, 140*(4), 32–48.
O'Neill, O. (2002). *Autonomy and Trust in Bioethics*. Cambridge University Press.
Pashevich, E. (2021). Can communication with social robots influence how children develop empathy? *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01214-z
Petit, N. (2017). Law and regulation of artificial intelligence and robots. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2931339
Pirhonen, J., Melkas, H., Laitinen, A., & Pekkarinen, S. (2020). Could robots strengthen the sense of autonomy of older people residing in assisted living facilities? *Ethics and Information Technology, 22*(2), 151–162.
Pu, L., Moyle, W., Jones, C., & Todorovic, M. (2019). The Effectiveness of social robots for older adults. *The Gerontologist, 59*(1), e37–e51.
Quigley, M. (2013). Nudging for health. *Medical Law Review, 21*(4), 588–621.
Rahwan, I. (2018). Society-in-the-loop. *Ethics and Information Technology, 20*(1), 5–14.
Raz, J. (1986). *The morality of freedom*. Clarendon Press.
Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
Robbins, S. (2019). AI and the path to envelopment. *AI & SOCIETY*. https://doi.org/10.1007/s00146-019-00891-1

Rogers, W. A., Draper, H., & Carter, S. M. (2021). Evaluation of artificial intelligence clinical applications: Detailed case analyses show value of healthcare ethics approach in identifying patient care issues. *Bioethics, 35*(7), 623–633. https://doi.org/10.1111/bioe.12885
Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion, 30*(4), 344–360.
Ryan, R. M., & Deci, E. L. (2017). *Self-Determination Theory*. Guilford Publications.
Schmidt, A. T., & Engelen, B. (2020). The ethics of nudging. *Philosophy Compass*. https://doi.org/10.1111/phc3.12658
Schmitt, M. N., & Thurnher, J. S. (2013). "Out of the loop": Autonomous weapon systems and the law of armed conflict. *Harvard National Security Journal, 4*, 231–281.
Schneewind, J. B. (1998). *The invention of autonomy*. Cambridge University Press.
Scoccia, D. (1990). Paternalism and respect for autonomy. *Ethics, 100*(2), 318–334.
Seïler, N. R., & Craig, P. (2016). Empathetic technology. In S. Tettegah & S. Sharon (Eds.), *Emotions and technology, emotions, technology, and design* (pp. 55–81). Academic Press.
Sharkey, A., & Sharkey, N. (2012). Granny and the robots. *Ethics and Information Technology, 14*(1), 27–40.
Shea, M. (2020). Forty years of the four principles. *The Journal of Medicine and Philosophy, 45*(4–5), 387–395.
Sparrow, R. (2012). Can machines be people? In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot ethics* (pp. 301–316). MIT Press.
Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics and International Affairs, 30*(1), 93–116.
Sparrow, R. (2017). Robots, rape, and representation. *International Journal of Social Robotics, 9*(4), 465–477.
Stoljar, N. (2000). Autonomy and the FEMINIST INTUITion. In C. Mackenzie & N. Stoljar (Eds.), *Relational autonomy*. Oxford University Press.
Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*. https://doi.org/10.14763/2019.2.1410
Thaler, R. H., & Sunstein, C. R. (2008). *Nudge*. Yale University Press.
Turkle, S. (2012). *Alone together*. Basic Books.
Turkle, S., Targgart, W., Kidd, C., & Daste, O. (2006). Relational artifacts with children and elders. *Connection Science, 18*(4), 347–361.
UNESCO. (2019). *I'd blush if I could: Closing gender divides in digital skills through education*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1
Vallor, S. (2015). moral deskilling and upskilling in a new machine age. *Philosophy & Technology, 28*(1), 107–124.
van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*, 719–735.
von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34*(1), 1–16.
Wakefield, J. (2018). Fear detector exposes people's emotions. *BBC*. https://www.bbc.com/news/technology-43653649
Walker, M. J., & Mackenzie, C. (2020). Neurotechnologies, Relational autonomy, and authenticity. *International Journal of Feminist Approaches to Bioethics, 13*(1), 98–119.
Walsh, T., Levy, N., Bell, G., Elliott, A., Maclaurin, J., Mareels, I., & Wood, F. (2019). *The Effective and ethical development of Artificial Intelligence* (p. 250). ACOLA. 10
Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence, 10*(2), 1–37.
Watson, G. (1975). Free agency. *Journal of Philosophy, 72*(8), 205–220.
Woiceshyn, L., Wang, Y., Nejat, G., & Benhabib, B. (2017). Personalized clothing recommendation by a social robot. *IEEE International Symposium on Robotics and Intelligent Sensors (IRIS), 2017*, 179–185.
Wolf, S. (1990). *Freedom within reason*. Oxford University Press.