The Philosophical Case for Robot Friendship

By John Danaher, NUI Galway

(pre-publication draft of a paper that is forthcoming in *Journal of Posthuman Studies*)

Abstract: Friendship is an important part of the good life. While many roboticists are eager to create friend-like robots, many philosophers and ethicists are concerned. They argue that robots cannot really be our friends. Robots can only fake the emotional and behavioural cues we associate with friendship. Consequently, we should resist the drive to create robot friends. In this article, I argue that the philosophical critics are wrong. Using the classic virtue-ideal of friendship, I argue that robots can plausibly be considered our virtue friends - that to do so is philosophically reasonable. Furthermore, I argue that even if you do not think that robots can be our virtue friends, they can fulfil other important friendship roles, and can complement and enhance the virtue friendships between human beings.

1. Introduction

Star Wars was a formative influence on my philosophical imagination. One of the things I liked most about it was its depiction of robots. I particularly liked R2D2, the quirky, high-spirited, garbage-can-lookalike companion to several of the human characters in the film series. I had no idea what R2D2 was saying — he/she/it spoke in a series of whistles and beeps — but the humans seemed to know. To them, R2D2 had a personality. 'He' was a valued companion and an invaluable assistant, always helping them out of close scrapes, without being completely at their beck and call. He was their friend just as much as he was their servant.

Fictional representations like R2D2 can provide useful inspiration for future technological developments. They can help us to imagine and plan for possibilities. But for some reason R2D2 doesn't seem to provide much inspiration to contemporary commentators on robotics. Our cultural conversation about robots seems to have taken on a much darker tone. Both the popular media and the academic literature is replete with people highlighting the risks associated with killer robots (Bhuta et al 2016; Sparrow 2007), sex robots (Danaher & McArthur 2017; Richardson 2015), care robots (Coeckelbergh 2015; Sharkey & Sharkey

2010; Sparrow & Sparrow 2006) and worker robots (Avent 2016; Ford 2016; Danaher 2017; Loi 2015). Robots are repeatedly viewed as a threat to cherished human values, possibly even an existential threat (Bostrom 2014).

This is not to say that everyone rejects the possibility of a more positive future with robots. Some people do, with some caution, see a role for robots as companions and friends (Gunkel 2018; Darling 2017; Elder 2015 & 2017; De Graaf 2016; Dumouchel and Damiano 2017). But the dystopian narrative tends to take precedence in popular conversations and thus limits our imaginative horizons. In this paper, I want to push back against the dystopian narrative and make a robust case for the view that robots can be — and perhaps should be — our friends. In defending this more optimistic outlook, I will not try to refute every argument presented by the doomsayers — the literature is far too vast to allow me to do so in the space of one article. My aims are more modest. I will simply argue that robots can plausibly be our friends (to conceive of them as such is within the bounds of philosophical reasonableness) and that robotic friendship can be a valuable social good. Consequently, we should not try to avoid creating robot friends — as some have argued (Bryson 2010; 2018) — we should instead actively pursue the most valuable opportunities for robot friendship.

I defend this argument in three phases. First, I situate my argument within the current literature, identifying the concerns one currently finds there and explaining exactly how my argument pushes back against those concerns. Second, I look at the concept of friendship, appealing to the classic virtue-model of friendship (which has its origins in the work of Aristotle), and arguing that it is philosophically reasonable to believe that robots can be our virtue friends. Admittedly, this is a strong thesis, so I follow this up by presenting an additional argument, which is that robot friendships can *complement* and possibly *enhance* human friendships.

2. Situating the Robot Friendship Thesis

I am defending what I call the 'robot friendship thesis'. This thesis has two parts:

Robot Friendship Thesis: Robots[1] can be considered our friends (to conceive of them as such is philosophically reasonable) *and* robotic friendship could be a social good.

I do not pretend to any radicalism with this thesis. If you have spent time in the company of roboticists and robot users, you will know that they can and do conceive of robots as their friends and companions (De Graaf 2016, Darling 2017, Darling and Breazal 2015, Dumouchel and Damiano 2017). This is true even when the robots themselves are not designed or intended to be companions. One clear illustration of this, derived from the work of Julie Carpenter (2016), is the emotional bond formed between soldiers and their bomb disposal robots. These bonds have resulted in elaborate battlefield funerals for 'fallen' robot comrades and a deep sense of loss among the soldiers when the robots are destroyed. On top of this, many roboticists clearly design robots to be friends and companions.[2] Look at the functionality of robots like Pepper (designed by Softbank) or even the digital assistants created by Apple, Google and Amazon. Their reassuring voices, playful tones, laughs and giggles, and fluttering eyelashes (in the case of Pepper) are all clearly intended to foster emotional attachment. So there is no shortage of people imagining and living with the idea of robot friendship.

I am also not the fist person to argue that robotic friendship is at least a possibility (De Graaf 2016, Darling 2017, Dumouchel and Damiano 2017, Elder 2015 & 2017; Emmeche 2014; Marti 2010). The radicalism of my thesis — such as it is — lies in my attempt to make a strong and unapologetic case for robotic friendship, to argue that robotic friendship is philosophically respectable, and to present it as a counterpoise to the philosophical and cultural criticism of robots that seems to be in the ascendancy. The academic literature on the social, legal and ethical implications of robots tends to highlight the ethical and social risks of robots. People are concerned about 'responsibility gaps' that will open up as robotic weapons systems and self-driving cars become widespread in society (Sparrow 2007; Matthias 2004; Bhuta et al 2016). People are concerned about robots stealing their jobs and leaving them destitute or disenhanced (Danaher 2017; Loi 2015**)**. People are even concerned about

[1] I am not going to offer a precise definition of 'robot' in this paper. As Gunkel (2018) notes, this may be

[2] To take but one illustration of this, Huang and his colleagues (2014) set themselves the challenge of designing 'friendliness' into a museum guide robot. They did so by modifying certain behavioural cues such as response time, approach speed, distance from user and attentiveness. There are many more examples of roboticists trying to work out the behavioural cues indicative and friendship and trying to design robots that can perform those cues e.g. Cañamero and Lewis 2016.

superintelligent robots using our bodies as resources to pursue their own, anti-humanistic, ends (Bostrom 2014).

These concerns have extended into a deep suspicion of robots with friendlike characteristics. This is most apparent in the literature on care robots (i.e. robots designed to provide care and companionship to the sick or diseased). That literature raises many important ethical issues, including the obvious safety and accountability issues that might arise from the mass deployment of care robots. But it also clearly raises issues associated with the nature and value of friendship (Elder 2015 & 2017). Several of the contributors to that literature worry about what will happen if our primary interactions are with robots — if we are starved of human contact — and if we are convinced that robot carers are our friends.

Sparrow and Sparrow (2006) painted a deeply dystopian picture:

> *"[Imagine] a future aged-care facility where robots reign supreme. In this facility people are washed by robots, fed by robots, monitored by robots, cared for and entertained by robots. Except for their family or community service workers, those within this facility never need to deal or talk with a human being who is not also a resident."* (2006, 152)

Others have followed suit. Their fears stem from the belief that any friendship or companionship provided by robots will be illusory (Elder 2015 & 2017; Turkle 2011). Robots will not be *true friends* (not philosophically proper or ethically valuable friends), though they may, through fancy machine learning tricks and clever engineering, con us into thinking they are. This will be terrible. We will view robotic contact as a substitute for human contact and we will lose out on important human and social goods.

Some suggest that we should, consequently, take steps to prevent robots from becoming thought of as our friends. Coeckelbergh, who otherwise adopts a social relational ontology that allows for the possibility of accepting robot 'others' into our communities (Coeckelbergh 2012,2014 & 2016), is very concerned about the risks of such acceptance. He argues that robots that 'appear to us' as human agents should be prohibited from the healthcare context (Coeckelbergh 2015). Bryson (2010 & 2018) has a more extreme position, arguing that roboticists should not design and market robots to function like human persons. To do so

would lead individual human beings to misallocate their cognitive resources (use up their friendship budgets) and misassign important ethical concepts (to commit what Nyholm (2015) calls 'evaluative category mistakes').

My goal here is to provide robust pushback against these fears. I want to argue that there is nothing illusory or unreal about robotic friendships. Robots can be our 'real' friends, at least under certain respectable and plausible conceptions of friendship, and that even if they fail to meet some philosophical ideal of friendship, their companionship can complement and enhance more ideal friendships with other human beings. This doesn't address all the dystopian fears one could have about robots — issues around responsibility, health and safety, existential risk, unemployment and so forth will remain — but it does provide a positive perspective that can be added to the mix when considering the future development of robotic technology.

3. Robots Can be Our Aristotelian Friends

I will defend the robot friendship thesis in two distinct ways. I start in this section by arguing that robots can plausibly be our virtue friends — that to conceive of them as such is philosophically reasonable, and that we should not condemn or discount the experiences of those who believe themselves to be in virtue friendships with robots. This is, admittedly, a strong claim, likely to get the backs up of many philosophers, so that's why I present a separate argument in the next section: even if they cannot be our virtue friends they can complement and enhance the virtue friendships we have with human beings. The second argument is, probably, more likely to win approval, at least in the short run, but I want to make the case for the first argument as being more important in the long run.

Some clarification is needed at the outset. What do I mean when I say that robots can be our virtue friends? The idea is widely debated and discussed, particularly in relation to the impact of technology on friendship (Elder 2014, 2015 & 2017; Kaliarnta 2016; Froding and Peterson 2012). It comes from books eight and nine of Aristotle's *Nicomachean Ethics* (Aristotle 2009; Costello 2015). There, Aristotle identifies three forms that friendship (*philia*) can take:

Utility form: A friendship that is pursued for instrumental gains to one or both parties.

Pleasure form: A friendship that is pursued because the interactions at the heart of it are pleasurable to one or both parties.

Virtue form: A friendship that is premised on mutual good will and well-wishing, and that is pursued out of mutual admiration and shared values on both sides.

Aristotle argues that the utility and pleasure forms of friendship are 'imperfect'. Although it is possible for them to be pursued on an egalitarian basis, they often involve asymmetries of power (one party gets most of the utility/pleasure) and they are easily dissolved when people stop deriving pleasure or utility from their interactions. Aristotle does not completely discount the value of such interactions, but suggests they are of a lesser type. The virtue form is different. It is much stronger, more meaningful, and an important part of the good life. For this very reason it also entails greater risk: one could try to attain virtue friendship with another and be betrayed or let down by the fact that they are only pursuing a pleasure/utility friendship. The sense of betrayal and loss here would be greater than if you knew it was only ever a pleasure/utility friendship (Margalit 2017**)**.

There have been many interpretations and applications of this virtue model of friendship over the years, including several attempts to identify the conditions that must (or likely should) be satisfied in order for them to exist (Costello 2015; Kaliarnta 2016; McFall 2012; Froding and Peterson 2012). Some of these accounts take us away from the original Aristotelian conception of that ideal, which was very much grounded in Aristotle's metaphysics and associated ethics. Nevertheless, they are inspired by and build upon his original conception and thus ought to understood as the direct descendants of his view.[3] These accounts tend to agree on the following conditions as being central to a virtue friendship: (a) *mutuality* (i.e. shared values, interests, admiration and well-wishing between the friends)*;* (b) *honesty/authenticity* (the friends must present themselves to each other as they truly are and not be selective or manipulative in their self-presentation); (c) *equality* (i.e. the parties must be on roughly equal footing, there cannot be a dominant or superior party)

[3] I am indebted to an anonymous reviewer for encouraging me to make this clarification.

and (d) *diversity of interactions* (i.e. the parties must interact with one another in many different ways/domains of life, not just one or two).

Given this understanding of friendship, and the conditions that need to be satisfied in order to pursue a truly valuable friendship, it seems like a tough sell to say that robots can be our virtue friends. Indeed, it seems like the virtue model can be used to argue that robots can never be our true virtue friends. In fact, some people, who are otherwise open to the idea of robotic friendship, have argued exactly that (e.g. Elder 2015 & 2017; and de Graaf 2016). The argument appears to work like this:

> (1) In order for someone to count as our virtue friend, certain conditions need to be met, including: (i) mutuality; (ii) authenticity; (iii) equality and (iv) diversity of interaction.
>
> (2) It is not possible for robots to satisfy conditions (i) - (iv).
>
> (3) Therefore, robots cannot be our virtue friends.

The critical premise here, of course, is (2). Prima facie, it looks like a strong argument can be made in its favour. First, it seems obvious that robots cannot meet the mutuality condition. After all, robots cannot have values and interests of their own: they only have the values with which they are programmed or that they acquire through, say, machine learning techniques. They cannot engage in mutual well-wishing and admiration. They don't (or won't for a very long time) have any inner mental life in which such states of mutuality are possible. Second, it seems obvious that robots cannot meet the authenticity condition. After all, the only way we could even begin to think of them as our virtue friends would be if they engaged in all the performative and behavioural acts we associate with virtue friendship. But this would entail a considerable act of deception: the robot would be going through the motions; they would not have any of the internal mental states that should accompany such outward performances in order for them to count as *authentic*. It would be like hiring an actor to be your friend (Elder 2015; Nyholm and Frank 2017). Third, it seems obvious that robots cannot meet the equality condition. After all, we are their masters and they are our creations. Until they achieve some greater-than-human powers, they will always be subservient to us. Fourth, and finally, it is difficult for robots to meet the diversity condition. For the time

being, robots will have narrow domains of competence. They will not be *general intelligences*, capable of interacting across a range of environments and sharing a rich panoply of experiences with us. They cannot really share a life with us.

That seems like a pretty powerful case for the prosecution. How can it be resisted? First, we need to clarify the nature of the impossibility claim that is being propounded in premise (2). Is it that it is not currently, *technically*, possible for robots to satisfy these conditions? Or is it that it is not *metaphysically* possible for robots to satisfy these conditions? If it is the former, then the argument is weaker than it first appears — it is at least possible that one day robots will become our virtue friends — though that day may be some distance off. If it is the latter, then the argument is more robust, but it is correspondingly much more difficult to prove. My suspicion is that many people in the debate favour the stronger, metaphysical impossibility claim, or at least a strong form of the technical impossibility claim — one which holds that while it may not be completely impossible for robots to satisfy all the conditions, the technical possibility is so remote that it is not worth considering (see, for example, Gunkel 2018 on the problem of 'infinite deferral' in debates about robot moral status).

Granting that there are these different ways of interpreting the impossibility claim, it then becomes important to distinguish between the impossibilities at stake in the four different conditions. For instance, it seems like conditions (iii) and (iv) (*equality* and *diversity*) could only ever really be construed as technical impossibilities, whereas as conditions (i) and (ii) (*mutuality* and *authenticity*) could more plausibly be construed as both technical and metaphysical impossibilities. Why is that? Presumably, equality is a function of one's powers and capacities and whether a robot is equal to a human with respect to its powers and capacities is going to be dependent on its physical and computational resources, both of which are subject to technical innovation. The same would seem to go for diversity of interaction. Whether a robot can interact with you across a diverse range of life experiences depends on its physical and computational dexterity (can it respond dynamically to different environments? can it move through them?) which is again subject to technical innovation. Contrariwise, conditions (i) and (ii) could, plausibly, be said to depend on more mysterious mental capacities (particularly the capacities for consciousness and self-consciousness) which many will argue are either metaphysically impossible for purely computational objects, or are so technically remote as to be not worth considering right now.

With these clarifications of premise (2) in place, we can build a case for the defence. It is, first of all, possible to resist the claim that robots cannot engage with us as equals or across diverse life experiences. We can do this by pointing out that the technical innovation needed to achieve this (enhanced intelligence and mobility) is well within our grasp. Indeed, it is possible to make a stronger claim: not only is it within our grasp, it is, in many instances, already here. To appreciate this point, we first need to think about the equality and diversity conditions in ordinary human friendships. The reality is that friends are rarely perfectly equal and rarely engage with each other in all domains of life. I have very different capacities and abilities when compared to some of my closest friends: some of them have far more physical dexterity than I do, and most are more sociable and extroverted. I also rarely engage with, meet, or interact with them across the full range of their lives. I meet with them in certain contexts, and follow certain habits and routines. I still think it is possible for to see these friendship as virtue friendships, despite the imperfect equality and diversity. But if this is right, then it should also be possible to achieve such virtue friendships with robots who are not our perfect equals or who do not engage with us across the full range of our lives. Imperfect, but close enough, equality and diversity will suffice.[4] Arguably, robots are already our imperfect equals (they are clearly better than us in some respects and inferior in others) and the degree of adaptability and mobility required for imperfect diversity is arguably already upon us (e.g. a drone robot companion could accompany us across pretty much any life experience) or not far away. Thus, it is not simply some technological dream to suggest that robots can (or will soon) satisfy the equality and diversity conditions.

The mutuality and authenticity conditions are more difficult. But we can, again, ask: what does it really mean to say that the mutuality and authenticity conditions are satisfied in ordinary human friendships? I would argue that all it means is that people engage in certain *consistent performances* (Goffmann 1959; de Graaf 2016**)** within the friendship. Thus, they say and do things that suggest that they share our interests and values and they rarely[5] do things that suggest they have other, unexpected or ulterior, interests and values. All we ever

[4] Aristotle himself may disagree and say that virtue friendship is incredibly rare and cannot exist without perfect equality etc. I make no attempt to reconcile my view Aristotle's. I would argue that longing for perfection is forlorn and that if it necessary it is deeply counterintuitive because it denies the experiences most of us have of our close friendships.

[5] I say rarely because, again, human friendships are often imperfect. We can occasionally feel betrayed by our friends or learn something about them that calls into question their honesty and authenticity. These occasional lapses are not fatal to friendship provided that they are rectified.

have to go on are these performances. We have no way of getting inside our friends heads to figure out their *true* interests and values. So the only grounds we have for believing that the mutuality and authenticity conditions are met in the case of ordinary human friendships are epistemically accessible grounds, in this case external behaviours and performances, not some deeper epistemically inaccessible, metaphysical attributes. But if that's all we have in the case of human friendships, then why can't these grounds provide similar justification for our belief in robotic friendships? More formally:

> (4) It is possible for the mutuality and authenticity conditions to be satisfied in our friendships with our fellow human beings (assumption).
>
> (5) The only grounds we have for thinking that the mutuality and authenticity conditions are satisfied in our friendships with our fellow human beings are the performative representations that they make to us, (i.e. these are the only epistemic grounds we have for believing in human virtue friendships).
>
> (6) These epistemic grounds for believing that the mutuality and authenticity conditions are satisfied in our virtue friendships with our fellow human beings can also be satisfied by robots (they can consistently perform mutuality and authenticity).
>
> (7) Therefore, it is (technically) possible for the mutuality and authenticity conditions to be satisfied in our friendships with robots.

This is an argument from analogy. It is not logically watertight. It is only as persuasive as we take the analogy to be. Some might challenge the analogy on the grounds that it is overly behaviouristic in its reasoning, but this is not quite right. The argument makes no claims about the ultimate metaphysical basis of the mind or intelligence. It only makes claims about the grounds upon which we justify our belief in our friendships. To defeat the argument you would need to argue that external performances are not all we have to go on when justifying our belief in our human friendships — that there are other epistemic grounds for that belief.

There are some possibilities in this regard. You could argue that we justify our belief in our human friendships because of our shared biological identity. In other words, we have first hand knowledge of the fact that we are conscious and self aware and that this is what allows

us to satisfy the mutuality and authenticity conditions. We have reason to suspect that our consciousness and awareness is linked to our biological properties (*i.e* our embodied nature and our sophisticated nervous systems). So we have reason to suspect that any creature that shares these biological properties will also be capable of satisfying the mutuality and authenticity conditions. We don't (and won't) share biological properties with robots, so we don't (and won't) have the same epistemic grounds for our belief in their capacity to satisfy the mutuality and authenticity conditions. On top of that, we will know things about the robots physical properties and ontological histories that will cast into doubt their ability to satisfy the relevant conditions. We will know that they have been engineered and programmed to be our friends — to perform in a certain way. This will undermine premise (6) of the argument. In this sense, knowing that your friend is a robot is akin to knowing that he/she is a hired actor (Elder 2015; Nyholm and Frank 2017).

This is an attractive line of thought — there is surely something about our shared ontological properties and histories that features in our justification for believing in human friendships — but it is less persuasive than it first appears. First of all, while the shared biological properties might give us *more* grounds for believing in our human friends it is not clear that these grounds are necessary or sufficient for believing in friendship. That they are not sufficient is apparent from the actor counterexample (the actor shares biological identity but is not a friend); that they are not necessary can be illustrated by another thought experiment. Imagine an alien race that is identical to human beings in all its outward appearances and behaviours, but has a different internal biology and evolutionary history. Could we form friendships with such beings? I see no non-question-begging reason to think not but in that case shared ontological properties and histories are not necessary for believing someone is your friend. Their consistent behavioural performances would given reason to discount the relevant of biology and ontology. On top of that, the claim that the programmed and engineered history of a robot should undermine our confidence in their friendship does not sit easily with the fact that many (including most philosophers) think that humans are engineered by evolutionary and developmental forces, and programmed by their genetic endowments and environmental histories. Engineering and programming does not differentiate humans from robots, particularly when you consider that modern robots are not programmed with specific top-down rules but with bottom-up learning algorithms. In this sense they are quite different from actors hired to be our friends.

I think we can push the point even further. I think that when it comes to the ethical foundation of our relationships with other beings, the only grounds we *should rely upon* are their consistent and coherent external performances and presentations. That is to say, I think we might be ethically obliged to normatively ground our relationships with others in how they consistently and coherently present themselves to us, not in what we may or may not know about their ontological histories or biological properties. Take a controversial example: transgender identity claims. Many people now advocate (and many legal regimes are beginning to recognise) the right for people to choose their gender identity. Accordingly, if a person chooses to (consistently and coherently) present themselves as a woman despite having the biological characteristics and ontological history of a man, we should respect that and rely upon that presentation in our interactions with them. I think this is broadly correct and that we are right to shift to this norm. This can be criticised. There are some reasons for thinking that people who consistently and coherently present themselves with a particular identity should not be treated in the same way as people who were raised with that identity. For example, treating the two groups equivalently may disrespect or trivialise a particular history of gender or racial oppression and inequality.[6] But those reasons are usually dependent on external political and social considerations, and about how people are recognised within political and social regimes, not on the ethics of interactions with the people themselves. When it comes to the intrinsic features of the interactions, the preferred norm is, I would argue, to ground the relationship in the external performances and not biological properties or ontological histories. If this is right, it gives us additional reason for endorsing premise (6). If robots consistently and coherently present themselves as our friends (appearing to satisfy the mutuality and authenticity grounds) then that is what we should base our beliefs in their friendship on.

Finally, I think there is another general argument for favouring the possibility of virtue friendship with robots. It is an argument from *epistemic humility* and *social tolerance*. People already form close emotional attachments to robots and other artifacts. To chastise or criticise such bonds on the grounds that they are not true virtue friendships is a form of social stigmatisation. Such individuals are treated as emotionally and socially defective. We should avoid such stigmatisation. This is not to say that all forms of stigmatisation or intolerance are imperssible. The stigmatisation of some alleged friendships can be justified (e.g. if your

[6] I'm not endorsing these arguments — I'm trying to skirt the ethical and political controversies of transgenderism and transracialism as much as possible.

teenage son has formed a close friendship with a group of Neo-Nazi white supremacists you would probably be justified in condemning and de-legitimating that friendship) but the epistemic standards that must be crossed before such stigmatisation is justified are high. If there is reasonable doubt about the legitimacy of a claim to friendship, it should be tolerated unless it is doing some clear and unambiguous harm. And so, if the performative account of friendship that I have defended falls within the scope of reasonability — i.e. if people can reasonably disagree about its ability to justify claims to virtue friendship — I think claims to virtue friendship that are justified on that basis should be tolerated.

In conclusion then, I think that robots could, plausibly, be viewed as our virtue friends. To do so is within the bounds of philosophical reasonableness, and entails no error sufficient to warrant philosophical and social condemnation. Indeed, if my argument is correct, the reasons for thinking that humans can be virtue friends can apply to robot friendships too. Since virtue friendship is widely agreed to be an important element of the good life it follows that robotic friendship can be an important element of the good life too.

4. Robots Can Complement and Enhance Human Friendships

Although I think the preceding argument is persuasive, I am aware that some people will continue to insist that it is wrongheaded. Fortunately, to them I can issue another response: so what? If robots cannot be our virtue friends, they can still be our utility or pleasure friends. We can derive instrumental gains and intrinsic pleasures from our interactions with them and so they can be, on net, a social benefit.

Very few people deny that robots can be our utility or pleasure friends. Even some of the critics of the idea of robotic virtue friendship agree that robots can be our utility or pleasure friends (e.g. Elder 2015 & 2017; De Graaf 2016). Nevertheless, there are concerns that having robots as utility and pleasure friends will have negative consequences for human virtue friendships. There seem to be at least two mechanisms that could be at play that lead to this opinion:

> **Forced Replacement**: As robot companions proliferate, we will be forced/compelled to interact with them instead of with human beings. In other words, our opportunities for human contact — and hence proper virtue friendship — will be curtailed.

Corrosion: As utility and pleasure friendships with robots become more widely available they will corrode or undermine the value/attractiveness of virtue friendship with human beings.

The forced replacement argument might be animating some of the arguments against the rise of carebots. We see it, to some extent, in Sparrow and Sparrow's (2006) dystopian description of the future-aged care facility, in Mark Coeckelbergh's (2015) concerns about robots replacing humans in the performance of care-related tasks and in Turkle's (2011) concerns about the impact of digital technologies on human sociability. The 'force' implicated in this argument is not some crude form of coercion. Nobody thinks that someone is going to put a gun to our heads and compel us to only interact with robots; the compulsion will be more subtle. The fear is that we will be starved of human contact because of the alleged efficiencies (economic and otherwise) of robots in workplaces and care facilities.[7] There will be less opportunity for human friendship as a result.

I do not think this argument is worth taking seriously in and of itself. It is important to bear in mind the more general effects of substituting humans for robots (outside of the friendship context). The effects of replacing humans with robots in other contexts could cut both ways when it comes to friendship. For example, the replacement of humans by robots in the workplace could arguably be a boon to friendship. It will increase the amount of leisure time available to the displaced human beings and they can use that leisure time to pursue virtue friendships with humans. In other words, if no one is coercing us into solely robotic friendships, and other human beings continue to exist and be available as potential virtue friends, we will continue to have the opportunity of forming virtue friendships with them, and possibly have more opportunities for this thanks to the proliferation of robots.

But what if we are not exercising those opportunities? What if we stop seeking out human contact? That's where the corrosion mechanism comes in. It suggests that we will stop exercising the option of human friendship because robotic friendships will be addictive: their benefits will be too immediate (too pleasurable and utilitarian) and this will make the hard

[7] An extreme illustration of this possibility can be found in Isaac Asimov's Robot series novel *The Naked Sun*, which imagines a future society where humans live with teams of robot servants and rarely interact with one another.

work of forming an virtue friendship with our fellow human beings less attractive. Is there anything to be said in favour of this argument?

The debate about the corrosive impact of the internet on virtue friendship is instructive in this regard. A number of people have criticised online friendships (e.g the friendships you acquire through social media, vlogging, blogging and chatrooms) for their shallow, utilitarian and pleasure-seeking nature (Froding and Peterson 2012; McFall 2012). They argue that the 'friends' we gain from our online interactions are not true virtue friends because we can only interact with them in a shallow way (text messages, emoticons, likes and retweets), and because the style of interaction encouraged by social media platforms is deliberately utilitarian and pleasure-seeking, and hijacks other motivations for friendship.

There is undoubtedly something to this. Online friendships can be addictive. I have felt this myself. I have lost countless hours to amassing several thousand twitter followers and facebook friends. I have lost countless hours to the desire to see my social media posts retweeted and shared. I get a nice feeling every time my numbers go up. I want to experience that momentary 'high' over and over again. And so I go on doing it, constantly growing my social network well beyond its alleged evolutionary limits (Dunbar's number of 150).

The same concern could apply to robotic friendships. Perhaps the immediate rewards of robotic interactions will prove to be addictive and will monopolise our attention. Perhaps we will spend all our time getting our robot friends to laugh at our jokes, promote our profiles, cook our meals, play games with us and so on. The robots will never get jealous of our successes, laugh at our misfortunes, gossip about us behind our back, or fail to turn up for an appointment. Who would want to bother with the messy complexities of human friendships after that?

This is a speculative argument. We don't yet know whether robotic friendships will prove corrosive of human friendship. We don't have a sufficiently large sample yet to prove it one way or the other. There is some plausibility to the speculations just outlined — and there are groups of people who claim to prefer non-human companions (Beck 2013**)** — but there are also some counterbalancing speculations that I think should be factored into the debate. In particular, I think the addictive and monopolising potential of robotic friendships should be counterbalanced against their potential to *complement* and *enhance* human friendships.

What do I mean by this? Again, the debate about internet friendship is helpful. The claim that the internet corrodes friendship by causing us to become addicted to shallow, utilitarian and pleasure-seeking interactions has been challenged. As both Sofia Kaliarnta (2016) and Alexis Elder (2014) have argued, the number and style of interactions that we have online are not limited in the ways that critics allege. Thanks to immersive online environments (such role-playing video games or Second Life) and the possibility of haptic, video, text, and audio-based communication, we can have rich and diverse interactions with our online companions. And while it is true that certain social media platforms encourage shallow forms of interaction, evidence suggests that people appreciate these interactions for what they are and do not view them as substitutes for richer bonds with 'offline' friends. This is encouraging for the defender of robotic friendships. It suggests that people will be able to have rewarding experiences with robot friends while at the same time retaining richer bonds with human friends.

More important than that, however, is what Kaliarnta has to say about the ways in which web-based communications can actually complement and facilitate virtue friendships. One thing that often prevents people from forming virtue friendships in their offline interactions is the interference of various biases and prejudices with the satisfaction of the virtue conditions. Virtue friendship requires rough equality between the friends, but in real life people often interact with presumptions of inequality. Aristotle himself was notoriously guilty of this, arguing that slaves and women were necessarily inferior to propertied men, and hence friendship was impossible between these groups. Such prejudices (perhaps in less extreme forms) still impact upon friendship to this day. We see people through the prism of their gender, skin-colour, disability, mental illness, social class, accent and so on. All of these things can form a barrier to full virtue friendship. One of the advantages of internet-based interactions is that they can *filter* out some of these biasing factors, enabling us to move beyond our own prejudices (Bulow and Felix 2014). When our interactions are forced through a narrow channel (e.g. text only), they are, in some ways, purified. Thus, when I interact with someone via a chatroom, my interpretation of them can be based solely on the quality and content of what they say, and not contaminated by irrelevant factors such as their

race, accent and social class.[8] This filtering effect of web-based communication can be a great boon to virtue friendship.

Now, I don't imagine that robots will play a similar filtering role. But I do think they can complement and enhance virtue friendships in broadly analogous ways. In particular, I think that robotic friendships could help ensure greater equality between human friends, thereby complementing and enhancing the bond between them. We often think of friendships as two-way interactions (there's *you* and *your friend*), even though we accept that one individual can have many friends. For the purposes of this argument, we need to imagine some friendships as three-way interactions: you, your robot friend and your human friend. Once we have that model in mind, we can start to see how certain interactions with the robot friend could complement and enhance the interactions between the human friend. Two modes of interaction are of particular importance:

> **Avatar interactions**: This arises where one of the human friends interacts with the other human friend via a robotic intermediary.
>
> **Outsourcing interactions**: This is where one of the human friends outsources some of their emotional, cognitive, physical or other friendship needs to the robot.

Avatar interactions can complement human friendships where one of the human partners is physically (or otherwise) incapable of engaging in certain activities with their human friends. Perhaps they are bedridden or otherwise physically disabled. This prevents them from sharing the rich diversity of interactions that the virtue mode of friendship envisages (i.e. prevents them from truly sharing a life with another). But if they could interact with their friends via a robotic intermediary (one that has the physical dexterity they lack) at least some of those interactions could be opened up to them.

This is could be a practically significant case, but it is not philosophically significant. As described, the robotic avatar sounds like it would not really be an independent, autonomous agent in its own right. Also, if we wanted to facilitate more diverse interactions between human friends, advances in human body-prosthetics seem like a more fruitful avenue to go

[8] The removal of such biasing factors was the motivation for the original set-up of the Turing Test.

down. That said, there are similar cases where an independent, semi-autonomous, robotic companion could make up for a deficit or disability that prevents one of the humans from engaging more fully in the activities of friendship. Guide dogs are companions to blind people that enable them to engage more fully with life. We could easily imagine the robotic equivalent of the guide dog providing assistance to one of the human partners (Sullins 2006).

The more interesting case is that of outsourcing interactions. There, the robot is not simply a tool used to facilitate an interaction with another human being; the robot is a distinct entity that fulfils certain friendship roles that would be too demanding or exhausting for the other human partner to fulfil and which are proving to be an impediment to virtue friendship. Perhaps your human friend likes almost everything about you except your constant demand to play tennis. Your demands to play it become so incessant that they end up avoiding you as a result, knowing that if you bump into them you will invariably demand a tennis match. Here, one of your utility/pleasure seeking demands is proving to be a barrier to virtue friendship. If you could outsource that demand to another companion, you could remove that barrier. We undoubtedly engage in this kind of demand outsourcing all the time with our human friends, but humans have their limitations in patience, kindness and enthusiasm. The attractive thing about robots is that the same basic logic can apply to our interactions with them — we can outsource some of our friendship demands to them — and they can satisfy those demands in an endlessly patient and enthusiastic manner.

This is not a purely hypothetical example. There is already some evidence to suggest that robot friends can perform this outsourcing function. Consider, the journalist Judith Newman's (2014) story of the friendship between her autistic son and Apple's Siri. As she recounts in her article 'To Siri, With Love', her son was obsessed with different sets of facts, e.g. weather formations, and had a tendency to repeatedly ask questions about them. She was unable to answer them all and sometimes grew weary with the incessant barrage of questions. Siri proved to be a godsend. She[9] was able to answer all her son's questions in a predictably kind and courteous manner, and encourage him to engage in some conversational back-and-forth in the process. This, in turn, seemed to have benefits for the interactions between mother and son:

[9] Siri's voice can be either male or female but was set to be female by Newman and her son.

"For most of us, Siri is merely a momentary diversion. But for some, it's more. My son's practice conversation with Siri is translating into more facility with actual humans. Yesterday I had the longest conversation with him that I've ever had. Admittedly, it was about different species of turtles and whether I preferred the red-eared slider to the diamond-backed terrapin. This might not have been my choice of topic, but it was back and forth, and it followed a logical trajectory. I can promise you that for most of my beautiful son's 13 years of existence, that has not been the case."

(Newman 2014)

This is admittedly just one illustration, but it is an important proof of concept that is backed up by other initial studies (Elder 2017). Robots can perform the outsourcing function and this must be weighed against the allegedly corrosive effects that they might have. And if they can perform this outsourcing function, two things follow. The first is that robot friendship could be an important social good because it could enable more people to approach the ideal of virtue friendship in their interactions with human beings; the second is that we should encourage the development of robots that can fulfil this outsourcing role. Doing so does not mean that we have to be blind to the darker side of social robots; it just means we can be more open to their lighter side.

5. Conclusion

And so we return to R2D2. I noted at the outset how he provided fuel for my early philosophical imagination, pointing to the possibility of a future filled with robot friends. What I hope to have provided in this article is some reason to think that the model represented by R2D2 is much stronger and more respectable than is often supposed. Virtue friendships with robots are technically possible. To suppose that we could form such a bond with a robotic agent is not philosophical unreasonable. The same grounds we have for believing in virtue friendships with other human beings carry over to robots. On top of this, even if robots cannot be our virtue friends, we can still form other valuable friendships with them. Doing so need not corrode or undermine our friendships with other human beings. Indeed, there is every reason to think that robots could complement and enhance the friendships between human beings by removing some of the barriers to the ideal of virtue friendships that are present in human-human interactions.

Conflicts of interest: None.

Acknowledgements: I would like to thank David Gunkel and Alexis Elder for conversations on human-robot relations that inspired some of the content of this paper. I would also like to thank two anonymous reviewers for feedback on an earlier draft of this paper.

Bibliography

Avent, R. (2016). *The Wealth of Humans*. New York: St. Martin's Press.

Aristotle (2009). *Nicomachean Ethics* (translated by WD Ross). Oxford: OUP.

Beck, J. (2013). Married to a doll: Why one man advocates synthetic love. *The Atlantic* 6th September 2013.

Bhuta, N., Beck, S., Geis, R., Liu, Hin-Yan and Kres, C. (2016). *Autonomous Weapons Systems*. Cambridge: Cambridge University Press.

Bostrom, N. (2014). *Superintelligence*. Oxford: OUP.

Bryson, J. (2010). Robots Should Be Slaves. In Yorick Wilks (ed) *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*. Amsterdam: John Benjamins.

Bryson, J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*. doi:10.1007/s10676-018-9448-6

Bulow, W. and Felix, C. (2014). On friendship between online equals. *Philosophy and Technology* 29(1): 21-34.

Cañamero, L. and Lewis, M. (2016). Making New "New AI" Friends: Designing a Social Robot for Diabetic Children from an Embodied AI Perspective. *International Journal of Social Robotics* 8(4): 523-537.

Carpenter, J. (2016). Culture and Human-Robot Interactions in Militarized Spaces: A War Story. London: Routledge.

Coeckelbergh, Mark. 2012. *Growing Moral Relations: Critique of Moral Status Ascription*. New York: Palgrave MacMillan.

Coeckelbergh, Mark. 2014. Robotic appearance and forms of life: A phenomenological-hermeneutical approach to the relation between robotics and culture. In *Robotics in Germany and Japan: Philosophical and Technical Perspectives*, ed. Michael Funk and Bernhard Irrgang, 59–68. Frankfurt am Main: Peter Lang.

Coeckelbergh, M. (2015). Artificial agents, good care, and modernity. *Theoretical Medical Bioethics* 36: 265-277.

Coeckelbergh, Mark. 2016. Is it wrong to kick a robot? Towards a relational and critical robot ethics and beyond. In *What Social Robots Can and Should Do*, ed. Johanna Seibt, Marco Nørskov, and Søren Schack Andersen, 7–8. Amsterdam: IOS Press.

Costello, S. (2015) *What is Friendship? Conversations with the Great Philosophers*. Dublin: The Liffey Press.

Danaher, J. (2017). Will life be worth living in a world without work? Technological unemployment and the meaning of life 23(1): 41-64.

Danaher, J. and McArthur, N. (2017). *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press.

Darling, Kate. 2017. "Who's Johnny?" Anthropomorphic framing in human-robot interaction, integration, and policy. In *Robot Ethics 2.0: From Autonomous Cars to Artificial*

Intelligence, ed. Patrick Lin, Ryan Jenkins, and Keith Abney, 173–191. New York: Oxford University Press.

Darling, Kate, Palash Nandy, and Cynthia Breazeal. 2015. Empathic concern and the effect of stories in human-robot interaction. *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, ed. IEEE, 770–775. doi: 10.1109/ROMAN.2015.7333675. Also available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2639689.

De Graaf, M.A. (2016). An ethical evaluation of human-robot relationships. *International Journal of Social Robotics* 8: 589-598.

Dumouchel, P, and Damiano, L (2017). *Living with Robots* (trans. Malcolm DeBoivse). Cambridge, MA: Harvard University Press.

Elder, A. (2014). Excellent Online Friendships: An Aristotelian Defense of Social Media. *Ethics and Information Technology* 16 (4):287-297 (2014)

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

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

Emmeche, C. (2014). Robot Friendship: Can a Robot be a Friend?. *International Journal of Signs and Semiotic Systems* (IJSSS) 3(2)

Ford, M. (2015). *The Rise of the Robots*. New York: Basic Books.

Froding, B. and Peterson, M. (2012). Why virtual friendship is no genuine friendship. *Ethics and Information Technology* 16: 287-297.

Goffman, E. (1959). *The presentation of the self in everyday life*. New York: Random House.

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

Huang, C.-M., Iio, T., Satake, S., and Kanda, T. (2014). Modeling and Controlling Friendliness for an Interactive Museum Robot. *In Proceedings of the 2014 Robotics: Science and Systems Conference (RSS'14)*

Kaliarnta, S. (2016). Using Aristotle's theory of friendship to classify online friendships: a critical counterview. *Ethics and Information Technology* 18: 65-79.

Loi, M. (2015). Technological unemployment and human disenhancement. *Ethics and Information Technology* 17(3): 201-210.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3): 175–183.

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

Marti., P. (2010) "Robot companions: Towards a new concept of friendship?" *Interaction Studies* 11(2): 220-226

McFall, MT (2012). Real character-friends: Aristotelian friendship, living together, and technology. *Ethics and Information Technology* 14: 221-230.

Newman, J. (2014). To Siri, With Love. *New York Times* 17 October 2014.

Nyholm, S. and Frank, L.E. (2017). From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible? In Danaher, J. and McArthur, N. (eds). *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press.

Nyholm, S. (2015). The medicalization of love and broad and narrow conceptions of well-being. *Cambridge Quarterly of Healthcare Ethics* 24(3):337-46

Richardson, K. (2015). The asymmetrical 'relationship': parallels between prostitution and the development of sex robots. *ACM SIGCAS Computers and Society - Special Issue on Ethicomp* 45(3): 290-293.

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

Sparrow, R. and Sparrow, L. (2006). In the hands of the machines? The future of aged care. *Minds and Machines* 16: 141-161

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy* 24(1): 62-77

Sullins, J. (2006). When is a robot a moral agent? *International Review of Information Ethics* 6(12): 23-30.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.