Ethics and Information Technology (2019) 21:241–251
https://doi.org/10.1007/s10676-019-09505-2

ORIGINAL PAPER

Objections to Simpson's argument in 'Robots, Trust and War'

Carol Lord[1]

Published online: 9 May 2019

Abstract

In "*Robots, Trust and War*" Simpson claims that victory in counter-insurgency conflicts requires that military forces and their governing body win the 'hearts and minds' of civilians. Consequently, forces made up primarily of autonomous robots would be ineffective in these conflicts for two reasons. Firstly, because civilians cannot rationally trust them because they cannot act from a motive based on good character. If they ever did develop this capacity then the purpose of sending them to war in our stead would be lost because there would be no moral saving. Secondly, because if robot forces did offer a moral saving then this would signal that the deploying government could not be trusted to be committed to the conflict. I disagree with both claims. I argue firstly that there are less demanding grounds that could allow robot forces to be trusted sufficiently to be effective whilst still achieving a moral saving over the deployment of human ones. Secondly, that this moral saving would not necessarily signal that the deploying body lacked commitment because its interpretation would be highly context-dependent. I conclude therefore, contra-Simpson, that robot forces could plausibly be effective in counter-insurgency engagements in the foreseeable future. I suggest therefore that there may be a case for developing a more finely grained understanding of the opportunities for, and challenges of, their use.

Keywords Trust · Trustworthiness · War · Robots · Counter-insurgency

Introduction

In "*Robots, Trust and War*", Simpson (2011) highlights the rise of asymmetric or counter-insurgency wars where non-state actors with few resources hide amongst the civilian population and use the advantage of tactical surprise to challenge opposing military forces who are better armed. He claims that winning such wars requires that both the forces that are mobilised against such actors, and their governing body, win the 'hearts and minds' of the civilian population i.e. that they gain their normative trust and that this trust is rational. However, he argues that a military force made up of mainly or entirely of autonomous robots would be unable to do so for two reasons. Firstly, because it would be irrational to trust them because they cannot act from a motive based on good character. Secondly because the use of such an army would signal to civilians that that the forces mobilising them could not be trusted to be committed to see the project through.

In this paper I challenge both claims. In doing so I make six main assumptions. The first two I adopt for sake of argument i.e. firstly, that winning 'hearts and minds' is a militarily valuable approach to counter-insurgency wars—even though it has been claimed that at a strategic level there is a 'staggering lack of empirical evidence to support it' (Egnell 2010 p 292). Secondly, that autonomous military robots are physical entities operating amongst the civilian population. I will refer to them in the text as 'robots' or 'robot forces' by way of shorthand.

A further two assumptions are simply those that Simpson makes as the basis for his argument. Firstly, that the ability of a robot to act from a motive based on good character marks the watershed of where it would be considered as morally significant as a human soldier. Secondly, that the relevant form of trust is inter-personal and based on a three—part model i.e. that P trusts Q to X—where P is the party who trusts i.e. the trustor, Q is the party who is trusted i.e. the trustee, and X is the action or set of actions that they are trusted to do. This model is commonly assumed in much of the literature although I also acknowledge the argument by Faulkner (2017) that it may not be the most fundamental form.

✉ Carol Lord
c.lord@cyberservices.com

[1] Surrey, UK

My fifth assumption is based on one that Simpson makes, though I extend upon it. He assumes that acting from the right motive is most commonly based on an individual being of good character. I assume that because his description of trust refers to *normative* trust, and the trustee having the *right* motive, and *good* character, it is not just any stable set of values that qualifies. What might constitute the right and the good has long been the subject of philosophical debate, and Simpson does not address this. However, I note that how this is interpreted will have a material influence on whether—and how far—robots could be capable of it.

My sixth and final assumption is that the expeditionary forces that Simpson describes as being made up 'entirely or very substantially of robot warriors (p332) do not include a human element that is sufficient alone to develop trusting relations with a host population. As a consequence, the focus of the question remains on whether robot forces could themselves be trusted—without relying on the human element of an integrated force.

I argue, contrary to Simpson's first claim, that rational normative trust based on good character is unlikely to be the most common form, and furthermore, is not feasible, necessary or even desirable in the context of counter-insurgency engagements. Robots are more likely to be rationally trusted on other grounds that are not as demanding of their capabilities and are therefore more common. Furthermore, trust may not necessarily require that robots have motives, or even any particular capacities at all. They could therefore potentially be trusted whilst at the same time offering a moral saving over the use of a human force.

Contrary to Simpson's second claim I argue that the moral saving of using a robotic force would not necessarily signal that the deploying body could not be trusted to see the project through. This is because how this act is perceived will depend on many factors including civilian beliefs, the capacities of the robots and circumstances of the engagement.

I conclude therefore, contra-Simpson, that it is plausible that robot forces could be trusted whilst at the same time providing an acceptable moral saving over the use of a human force—and therefore may be effective in counter-insurgency engagements in the future.[1]

In the following sections I outline Simpson's case and set out his explanation of rational normative trust based on good character. I then develop my case against his two main claims. Firstly, I explain why trust based on good character is unlikely to be either the most common type of trust, or feasible, necessary or desirable in the context of a counter-insurgency robot force. I describe alternative grounds for rational normative trust that are likely to be both more widespread, and pertinent to robot counter-insurgency forces, and claim that neither motives, nor even any evidence of the capacities of trustees are strictly necessary for it. On this basis I conclude that, contra-Simpson, forces made up primarily of autonomous robots could potentially be trusted whilst still offering a moral saving over the use of human ones. Next, I consider, but then reject Simpson's second claim that the deployment of robot forces would necessarily signal a lack of commitment to the conflict—on the grounds that perceptions of this would be highly contingent and would not necessarily undermine trust in the deploying body. I conclude that Simpson's two main arguments are therefore unconvincing and that in the future robots may have the potential to form effective counter-insurgency forces.

Simpson's case in outline

In "*Robots, Trust and War*", Simpson observes that success in wars amongst the people requires that counter-insurgency forces win their 'hearts and minds' and cannot just depend on technical superiority on the battlefield. He argues that this would not be possible with armies made up solely or primarily of autonomous robots for two reasons—both of which are based on certain psychological claims about how robot forces may be perceived.[2]

Firstly, because they are unable to win the rational, normative trust of the people. Furthermore, if they did develop this capacity then deploying them would no longer be a moral saving. He argues that this is because the 'more important and wide-ranging forms of trust' are based on an expectation of the trustworthiness of the trusted i.e. most commonly expecting trustees to 'have good character such that they act from a motive' (p332). Simpson claims that robots do not have the capacity to act in this way and would instead be prone to perverse behaviours from rule-following which would be considered failures of judgement in people. Furthermore, he claims that if robots ever *could* act from a motive, then this would be the point at which people could 'begin to form relationships with them, empathise and care for them, and view them as morally responsible' (p333). In this event, there would be no moral saving in sending them to war instead of humans.

Secondly, Simpson argues that a body that deployed such a force in a counter-insurgency engagement would not be trusted because it would signal that they were not committed to seeing the project through. This is because robots are

[1] I make no claims as to whether it is morally acceptable to use them in such contexts however.

[2] I am grateful to an anonymous reviewer for highlighting this, which I make further reference to the penultimate section of this paper.

expendable in a way that human soldiers are not, so that by using robots, governments would therefore be avoiding the moral costs of conflict—making it easier to not only enter into war, but also to walk away at any point as there would be fewer sunk costs.

As a consequence of both factors Simpson concludes that a primarily robotic force would be unable to be trusted and would therefore be ineffective in counter-insurgency engagements.

In the following paragraph I set out in more detail the account of trust that Simpson adopts as the basis for his claims.

Simpson's account of trust

Simpson acknowledges that trust is a highly contested concept and may even elude precise definition. However, he does provide an account of what he considers to be the most common type of trust. He follows Hollis (1998) in distinguishing between predictive trust—which he regards as merely reliance—and normative trust which 'occurs when I rely on an agent to take me into account in the way that they act' (p327). After Pettit (1995) Simpson claims that trust is distinguished from reliance by being 'dynamically interactive' (p327) i.e. that the trusted is aware of being trusted and responds to this. I assume that his use of the term 'normative' implies not just that trustors expect that trustees *will* take them into account, but that there is at least a prima facie reason for them to do so. Finally, Simpson claims both a vulnerability to betrayal and a liability to reactive sentiments as 'touchstones' of normative trust.

He considers trust to be either irrational or rational, and his claims in the rest of the paper are based on rational trust alone. He considers that 'the primary dimension in which trust is assessed for rationality is that of *practical* rationality' because it is a decision about action. Practical rationality requires that the trustor seek good evidence that the trusted has the right attitude because trusting where this is not available would be foolish. However, he emphasises that he does not claim that it is *always* the case that trust is rational only if there is good evidence for trustworthiness.

Simpson also claims that 'the most important and wide-ranging' forms of trust involve the trustor expecting the trustee to be motivated to act based on good character—this being 'not less than having the stable disposition to think certain things to be important and valuable, and being competent to act sensibly in the light of these' (p328). He adds that 'the value of some things, at least, is the subject of relatively invariant agreement' (p328/9). That is not to deny that there are other grounds for trust—as he claims that trustees might also be trustworthy 'because they think it is in their long-term interest to be so, or because they love, care for or have goodwill towards the person who trusts them, or because they regard themselves as owing it to the other (respectively: Hardin 2002; Baier 1994 and Jones 1996; Holton 1994)' (p328).[3]

Finally, he notes that trust also incorporates an assumption about the competence of the trustee to perform as expected i.e. in the context of a counter-insurgency engagement, civilians 'must trust that you will defeat the insurgents' (p330).

By way of summary, Simpson's account of rational normative trust therefore requires, either explicitly or implicitly, that the trustor:

- Believes that the trustee is competent to do as expected as the basis for the trustor to rely on them.
- Believes that the trustee can influence what they are being trusted to do and can freely choose how to behave with respect to it i.e. making the trustor potentially vulnerable to betrayal.
- Anticipates that the trustee will take them into account in the way that they act, based on appropriate evidence that they will do so i.e. most commonly based on evidence that the trustee will act from a motive based on good character.
- Cares that the trustee acts as the trustor expects them to—and feels betrayed by the trustee if they do not.

In the following section I argue that trust based on good character is however unlikely to be either the most common form as he claims, or even necessarily required of a military counter-insurgency force.

Why trust based on good character may not be common, feasible, necessary or desirable

Simpson acknowledges that there may be multiple forms of trust. Therefore, rather than debating the strengths and weaknesses of his particular account I examine his claim that trust based on the trustworthiness of the trustee and grounded in their good character is the 'more important and wide-ranging' form (p332), and his assumption that it is also what is required in the particular context of counter-insurgency engagements.

I argue firstly that rational normative trust based on good character is unlikely to be the most common form, and

[3] I note that not all of these motivations would commonly be recognised as morally or socially desirable—so he appears to use the term trustworthiness simply to refer to the capacity of a trustee to do as they are expected. As such, it would not necessarily be a virtuous trait (Wright 2010).

secondly that it is not obvious that such trust is feasible, necessary or even a desirable feature of counter-insurgency forces. Consequently, trust based on the good character of either human or robot counter-insurgency forces is unlikely.

Normative trust based on good character is unlikely to be the most common form

Simpson does not justify his claim that trust is most commonly based on the good character of the trustee, and I argue that this is unlikely in general because it would be challenging for a trustee to be of good character, or for a trustor to recognise them as such. I support my claim with examples that show that this claim is also unlikely to hold true in the particular context of counter-insurgency engagements.

That is, to be trustworthy based on good character according to Simpson's account, a trustee needs to consider how to behave, taking account of the trust placed in them by the trustor as just one factor. This is quite demanding as it requires the trustee to be capable of both moral reasoning and of consistently acting on those reasons.

Furthermore, for that trust to be rational as Simpson presumes it should be, he claims that it needs to be based on good evidence that it is likely to be well-placed. However, as Hardin (2002) has already argued, getting evidence of good character is challenging as it would require significant familiarity with the trustee. To elaborate on this, it would require not only evidence of a potential trustee's actions in a variety of contexts, and over time but also an appreciation of the motivation behind them—because only some types of motivation would be consistent with good character. For example, a trustee would not usually be regarded as being of good character if they complied with a trustor's expectations merely for reasons of self-interest.

Evidence of a trustee's real motives are also likely to be concealed where there are significant levels of external controls over their actions (Wright 2010). For example, in the particular context of a counter-insurgency force, the governing body specifies objectives and there are rules and values that soldiers are expected to comply with. Their behaviour is scrutinised by colleagues, their command, other participants and civilians, and those deviating from what is required are likely to be subject to punishment. As a consequence, it could be rational for civilians to trust individual soldiers on the basis of their self-interest as it would appear to be in their interests to conform.[4] However, this would make good character difficult for civilians to discern as it would not be clear whether behaviours were driven by the good character of a soldier or these external controls. Trust based on the good character of soldiers would therefore be unlikely, even if there was the potential for it.

Simpson (2013) has disputed Hardin's claim that getting evidence of good character is challenging on the grounds that reliance can be placed instead on the judgement of others, or of institutions to which the potential trustee belongs. In the context of counter-insurgency engagements presumably reliance might be placed on participation in bodies such as the UN and NATO, and commitment to the Geneva Convention and an ethical foreign policy. However, I argue that this begs the question because the trustor would still need to satisfy themselves that those individuals and/or institutions could in turn be trusted to provide such broad-ranging assurance. Most likely they would provide assurance in only limited respects, so multiple sources would have to be sought and the trustor would have to seek sufficient grounds for trusting each. Furthermore, some institutions may establish only principles to guide behaviour that are subject to interpretation, and therefore may not provide the level of assurance that is required. In the circumstances of counter-insurgency engagements finding a sufficient basis for such assurance may be particularly challenging, and essentially impractical for civilians.

In the same paper Simpson also argues that 'a decent moral character may be permissibly assumed in situations where it does not matter too much if the assessment is wrong and decency is sufficiently prevalent across a population' (2013 p554). However, in the particular context of a counter-insurgency war it is unlikely that either condition is met. Not only are the very lives of civilians often at stake, but counter-insurgency forces may not even be recognised as part of the civilian population—for example if they are foreign or robotic—making it implausible that civilians would assume that they had a level of moral decency that matched their own.

Therefore, contrary to Simpson's claim, normative trust based on the good character of the trustee seems unlikely to be the most widespread form of trust—both in general and in the particular context of a counter-insurgency engagement—because it is more difficult for a trustee to achieve, or for a trustor to establish. I next argue that this form of trust

[4] Hardin (2002) has already noted that most trusting relationships occur within the context of a network of relationships—which can provide a sufficient basis to expect that an individual might be trustworthy on the basis of self-interest. Simpson (2013) notes that this requires only that the interests of the trusted are better served by being in a group of others, and that it is likely that untrustworthy

Footnote 4 (continued)
behaviour is sanctioned. He claims that this would not be a strong basis for trust under certain conditions however i.e. if membership of the group was not advantageous or was about to end, if betrayal was unlikely to be sanctioned, if the group was about to cease to exist, or if there was a significant single trade-off that would provide the trusted with a greater gain than group membership. Consideration of these conditions is beyond the scope of this paper however.

is also unlikely to be appropriate in the context of counter-insurgency engagements.

Normative trust based on good character may not be feasible, necessary or desirable in counter-insurgency forces

I claim that interpersonal normative trust based on the good character of the trustee is also unlikely to apply in the particular context of a counter-insurgency engagement because it is not feasible, necessary, or desirable. I argue that it is not feasible because the nature of these conflicts means that it is questionable whether soldiers in counterinsurgency forces could *ever* be perceived as being of good character by civilians because of significant mismatches between the values of the respective cultures. For example, Egnell (2010) claims that 'traditional sources of legitimacy in Afghanistan, (are) often based on identity and cultural affinity', and that the international coalition are foreigners with values that are 'oceans apart' from local civilians (p299). It is therefore not simply that counter-insurgency forces could not be rationally trusted because they are unable to adhere to moral values, it is that even if they did so, those moral values could be very different to those of the civilians. This is consistent with an observation made by Lewicki et al. (2006) that initial distrust between parties may arise from cultural/social factors, an untrustworthy reputation or other contextual factors, which Kramer (1999) has suggested may form psychological barriers to the formation of trust. Efforts to win civilian 'hearts and minds' based on making robot counter-insurgency forces trustworthy—as measured by the values of their developers—may therefore ultimately be misguided.

It might be argued that Simpson's account of good character is sufficiently thin to allow overlap between civilians and counter-insurgency forces.[5] However, I claim that even if this is the case, areas of divergence could plausibly prove more influential than areas of overlap because trust is usually characterised as being easier to lose than to gain.

Furthermore, if Simpson's account of good character is taken to be very thin then it will be insufficient to support his other claims i.e. that it is the basis for winning 'civilian hearts and minds' and excludes robots from being trusted. That is, if good character is interpreted in the thinnest sense that Simpson describes as 'having the stable disposition to think certain things to be important and valuable, and being competent to act sensibly in light of these' (p328) then robots could be capable of it already because they can be programmed with some limited set of values and are arguably competent to act 'sensibly'—depending on how this is interpreted. For example, they might be considered to act 'sensibly' on such values under most circumstances if they could at least prioritise them—even if unable to do so in a more sophisticated way e.g. by resolving conflicts. Furthermore, winning 'hearts and minds' requires that soldiers take the interests of civilians into account—so they would need to respect not just *any* values, but those *particular* values that are pertinent to civilians. If they did not then their actions would not provide sufficient evidence of good character in civilian terms for them to be rationally trusted.

Even if certain values are the subject of 'invariant agreement' as Simpson suggests (p329) and are therefore shared, their interpretation and/or application may still differ between civilians and soldiers resulting in a loss of trust—and this may be exacerbated by membership of the counter-insurgency force. By way of example, it is likely that there will always be tension between the demands of soldiers as members of a counter-insurgency force and the requirement to be of good character as a civilian. Both civilians and forces may value integrity, duty, and loyalty—but as soldiers Simpson observes that "the most fundamental and deeply held value is the successful completion of the given task" (p326). As a consequence, other values which may be more important both to civilians, and even to soldiers as individuals, may be given a lower priority when ensuring that a particular end is achieved.

It may also be unnecessary for robot forces to win hearts and minds because they may not materially contribute to achieving this objective for two reasons. Firstly, because perceptions of the force as a whole may be more influential in establishing trust than the actions of individual soldiers.[6] This is because soldiers are not autonomous in many ways as they are subject to the rules of the force (i.e. including its governing body) who define the objectives of the engagement, as well as the rules and values that should be adopted in pursuing them. Civilians therefore remain vulnerable to betrayal by the choices of another as Simpson's account of trust specifies—but this choice is largely in the hands of the force and not individual soldiers.

Secondly, there are alternative interpretations of, and approaches to, achieving 'hearts and minds' which involve different tactical activities in which military forces do not, and perhaps should not participate. That is, Simpson's account claims that soldiers would need to 'simultaneously conciliate and build relational bridges at the same time as defending themselves' as the basis for engendering civilian trust (p 331)—which suggests an assumption of a generally

[5] I am grateful to an anonymous reviewer for raising this possible objection.

[6] I am grateful to an anonymous reviewer for highlighting the question of how far interpersonal trust might be reflected in trust of the governing body. This raises interesting questions about the respective roles of group and interpersonal trust which are outside the scope of this paper but merit additional consideration.

less coercive approach to an engagement. However, Egnell (2010) claims two alternative approaches involve undertaking humanitarian and development activities and specialist information operations. He suggests that these may not require, and possibly should not involve military forces at all because it could undermine civilian trust in them. For example, if military forces are less well equipped than others to undertake humanitarian activities then they are unlikely to be trusted because not only might they fail to deliver as promised, but their involvement may confuse civilian expectations of them.

It might be argued that it would still be necessary for soldiers to be trustworthy—or at least not untrustworthy—whatever their involvement in a counter-insurgency engagement and however extensive their role because distrust between civilians and soldiers on the ground could stymie attempts to create trust through these other means. However, I claim that this could be achieved without requiring them to be trustworthy based on good character[7] as I explain in the following section, and it may even be sufficient that they are merely trusted predictively.

Finally, I argue that good character on Simpson's account may not even be desirable in counter-insurgency forces because it may not support what civilians and counter-insurgency leadership require them to do. As Wright (2010) has argued, if trustworthiness is a moral characteristic, then it does not necessarily entail doing as a trustor expects. A trustee acting on the basis of good character may take a trustor into account yet decide not to do as they expect if they consider the action morally wrong—or less morally valuable than some other action external to that relationship. The good character of a trustee is therefore only a good reason for a trustor to trust them if what they are trusted to do is something that the trustee is also likely to regard as morally correct. For example, a civilian might rely on counter-insurgency forces to shoot insurgents on sight if that is what they expected of a protective force. However, if the force was trustworthy in a moral sense, it might hesitate to kill insurgents rather than capture them. For similar reasons it may be undesirable for the leadership of counter-insurgency forces to deploy forces that are trustworthy based on good character if this leads to them questioning the morality of the war, its conduct and operations.

I note however that even if trust based on good character is neither the most common form, nor what is required of counter-insurgency forces, it may still be the most important in general—particularly if it helps to maintain trust in society as a whole (e.g. McGeer and Pettit 2017) though consideration of this matter is beyond the scope of this paper.

In conclusion, I have argued that rational, normative trust based on the good character of the trustee is unlikely to be a common form of trust, and is not feasible, necessary or even desirable in the context of a counter-insurgency force. In the following section I suggest alternative grounds for rational normative trust in soldiers that are arguably more plausible.

Robots may be more easily trusted on other grounds

Simpson does acknowledge that there are grounds for rational normative trust other than the trustee having motives based on good character and I next describe an alternative account that is considerably less demanding of the trustee, and is therefore likely to be more widespread. I also claim contra-Simpson that it is plausible that neither motives, nor even any particular attributes are necessarily required of a trustee as the basis for rational normative trust. As a consequence, it is plausible that civilians could trust robots—and more easily than Simpson anticipates–even whilst they offer a moral saving over the use of a human force.

I note however that none of the grounds for trust that are discussed can provide a reason why a counter-insurgency force made up primarily of autonomous robots would be generally more (or less) likely to be trusted.

Robots could be rationally trusted on grounds other than good character

As Simpson notes, rational normative trust based on the good character of the trustee provides assurance about their behaviour even in circumstances where what is expected of them cannot be specified in advance. Although commonsense dictates that this would usually be desirable, I argue that it may be sufficient for a trustee to be trusted on more limited grounds—and that this is likely to be more common because it is less demanding.

For example, Cogley (2012) suggests alternative grounds for trust which seem particularly plausible as they explain feelings of betrayal—which is widely acknowledged to be a core feature of trust. Rather than grounding trust entirely in features of the trustee i.e. on their good character, he claims instead that trust may involve the trustor feeling entitled to the consideration of a trustee because they are party to a normatively characterised relationship with them. The trustor would feel betrayed if the trustee breached their trust in this case because not only would they lose the expected benefits of that trust, but also because it would adversely affect the relationship between them. On such grounds the trustee would therefore be trusted only within the particular domain

[7] I acknowledge that there may still be other reasons why robots might nonetheless need to be capable of being trustworthy based on good character—for example to meet the needs of the human soldiers in a primarily robotic force, or the expectations of the population in their country of origin.

of that relationship, and this trust would be rational if the trustor had evidence that the trustee was likely to conform with their, and perhaps societal expectations of that role.[8] On such grounds the trustee could even be trusted to do something that was immoral if that was what was expected within the scope of that relationship. For example, a criminal could trust a colleague to help rob a bank.

By way of example, I could trust my regular mechanic to fix my car–not because he is trustworthy as matter of good character but because he is competent to do as I expect, and has good reason to fulfil his commitment to me as a tradesman. This is not simply predictive trust as it has the characteristics of interactivity that Simpson claims distinguish predictive from normative trust as set out in my summary of Simpson's account of trust i.e. I trust the mechanic because I rely on him to fix my car, he is aware of this and we both believe that he ought to take my trust into account because we have an implied relationship in virtue of his role. If he did a bad job nonetheless then I would feel betrayed because not only would my car not have been mended but our relationship would also have been undermined.

I claim that this form of trust addresses the two practical difficulties of trust based on good character that I identified in the previous section, and is therefore plausibly a more common form of rational, normative interpersonal trust. That is, it is less demanding of the trustee as they need only to conform to the trustor's expectations of them. Furthermore, the trustor would require less, and more easily available evidence for this trust to be rational. This could include anything that increased the likelihood that the trustee would behave as they ought to, but only within the specific domain of the relationship. It would not even necessarily require evidence of trustee motivation—unless this was constitutive of that relationship. For example, caring for the other is usually constitutive of a close friendship—so in this case, there would need to be evidence that the friend was motivated to look after you in various contexts simply because they were your friend and not, for example, because they believed that it was in their self-interest to do so. However, in less demanding relationships such as the mechanic example, I may trust them regardless of why they are fulfilling their role—as long as there is evidence that they are likely to continue to do so.

This account also better explains how in practice we typically trust others to behave in a certain way in some circumstances but not in others, and about some things but not others—as this is dependent only on the scope and nature of the trustor's relationship with the one who is trusted. For example, I may trust a friend to be honest with me without any expectation that he will also tell his mother the truth. In contrast, Simpson's account implies that to be trusted on the basis of 'good character' would require the behaviour of a trustee to be consistent on a much broader basis—though I acknowledge that the extent and nature of this is likely to depend on exactly how his account of the 'good' in 'good character' is interpreted. A consideration of the possible alternatives is beyond the scope of this paper however.

The immediately obvious objection to applying these grounds for trust in respect of robot forces is that trusting robots would arguably be little more than reliance (or predictive trust) given that relationships between robots and civilians are not properly interactive, and we do not usually feel betrayed by technology when it fails to perform. This is only the current state of affairs however, and in the future, this will depend on how our relationships with robots, and their capacities, develop. If robot functionality remains limited then they may continue to be considered simply as tools and trusted predictively. In this case only their originators, rather than the robots themselves, would be thought to be responsible and accountable for their actions. On the other hand, if they became more functionally advanced civilians might experience some sort of interactive relationship with them and might indeed trust them and feel betrayed in the event that their trust proved misplaced.[9] This is not as outlandish as it seems, for reasons that I address in the section below.

There are therefore alternative grounds for rational normative trust that are less demanding of both trustees and trustors, and are consequently plausibly more common as the basis for civilian trust in robot forces. In the following sections I argue, also contra-Simpson, that not only are motives unnecessary for rational normative trust, but that robots might even be trusted on this basis without having any particular capabilities at all.

Robots may not need motives to be trusted

According to Simpson's account, robot forces could not be rationally trusted because they cannot act from a motive based on good character. They would instead display failures of judgement because their 'rule-following behaviour'[10]

[8] This is consistent with Hollis (1998)—who Simpson references in distinguishing between predictive and normative trust. He claims that normative trust has merely a 'moral flavour' that 'hovers uneasily between moral obligations and the local requirements of a particular society' (p11).

[9] I acknowledge that there is a wider debate about whether, and how far robots might be considered effectively morally responsible, but this is outside the scope of this paper.

[10] I assume that he interprets 'rules' very broadly here as some forms of artificial intelligence are already capable of considerably more than literal rule-following. Machine-learning systems of various types are able to create their own rules on a probabilistic basis, and as such might already be said to reason—in the sense that they can develop their own basis for favouring a particular action.

would lead to action with perverse outcomes' (p332). In this section I argue that trust based on the motives of a trustee is not necessarily rational, and that motives may not even be necessary for values-based behaviours. Furthermore, robots might still be rationally trusted even when there are 'perverse outcomes' due to rule-following behaviour.

That is, Simpson claims that robots could not be rationally trusted without having a motive to act. Convention is that normative and/or moral motivation is based on the reasons that an individual has to act (Rosati 2016), although the relationship between moral/normative reasons and motivation is not well understood. Whilst some believe that reasons are directly motivating, others require that both reasons and desires are involved. If the latter is the case then arguably it would be irrational to trust robots if they *could* act from a motive—because just like humans their own desires might then overwhelm whatever reasons they had to act.

By maintaining that motives are necessary for trust, Simpson also assumes that the processes involved in values-based reasoning and decision-making must necessarily operate in the same way for both robots and humans. It is not obvious why this should be the case however—and values-based behaviours could perhaps be achieved in robots through other means whilst still supporting the kind of trust that I describe in the previous section.

Even with the possibility of perverse outcomes due to rule-following behaviour I claim that it is not obvious that it would be irrational to trust robots however, because the occurrence and severity of these may eventually be reduced *sufficiently* for trust still to be rational. For example, with technological advances, unexpected robot behaviours may become less common[11]—particularly if robots are limited in the scope of their functionality. At the same time, the risks of poor outcomes may be compensated for by the benefits that the enhanced capabilities of robots can bring e.g. being more physically robust, they may able to immobilise rather than kill insurgents (Lokhorst and van den Hoven 2014), and be able to take more measured decisions in the absence of emotional responses. Consequently, in practice civilians may have limited evidence of reasons to distrust a robot force but good reasons to trust them on other grounds e.g. if their lives are at risk.[12] Arguably this could make it sufficiently rational for a civilian to trust a robot force—or at least not to distrust it.

Whether this is actually rational according to Simpson's account depends on what is conceived to be valid evidence for justified belief e.g. whether evidence needs to reflect an external physical truth or not (e.g. Kelly 2016). In practice the technical sophistication and opacity of such systems—particularly in the context of military technology—means that humans are increasingly unlikely to be aware of and/or understand the real risks of them. As a consequence, they might believe a robot to be capable of being trusted if there was no evidence of perverse outcomes, even if the impact should one happen is extremely high. That is, it may be theoretically rational to trust robots based on the evidence that civilians are likely to have, when in reality, it may not be advisable to do so.

Robots could be trusted regardless of their attributes

Finally, I claim that it is also plausible that humans may trust robots regardless of their actual attributes or behaviours because it may be sufficient for them to simply *appear* to have the necessary attributes of a trustee.

For example, in the context of his account, Cogley (2012) claims that it may be sufficient for a party to merely believe that they are party to a normative relationship with a trustee as the basis for trust—whether or not they actually are. That is, trust may not be dependent on attributes of a trustee that have an objective ontological status as Simpson appears to suggest, but are instead epistemically subjective (Kirkpatrick et al. 2017).[13] Such expectations might be unrealistic, but could be the basis of trust nonetheless. In similar vein, Coeckelburgh (2009) claims that trust may not be based on the trustor's assessment of the plausibility of the trustee but simply on perceptions of emotional bonds between them. Trust on these grounds would therefore be an emergent property of social relationships and whether a robot force could be trusted would therefore depend on whether they were perceived to be part of those interactions—which would be possible simply on the basis of them *appearing* to have the characteristics of a social entity regardless of whether they actually had them. He argues that this is consistent with the way that we treat other humans—as in practice we do not have evidence that even they have the capacities to reason, have motives, or be autonomous because all we see is their outward behaviours—but we trust them nonetheless. Whether such trust may be considered rational depends again on exactly what is conceived of as valid evidence for justified belief—though again, this debate is beyond the scope of this paper.

[11] Though as a general observation, with increasing complexity comes an increasing risk of unanticipated error—so this will be by no means an easy task.

[12] This presupposes that we can decide to trust—which is controversial, though Simpson's emphasis on the rational nature of trust appears to assume it. Furthermore, trust usually requires that this decision is freely made—which in a counter-insurgency context may be debatable.

[13] I am grateful to an anonymous reviewer for bringing this article to my attention.

There is certainly empirical evidence that in practice humans appear to respond to, and potentially trust robots despite a lack of evidence that they actually have the necessary capabilities to be plausible trustees—and therefore whether or not this is 'foolish' as Simpson claims. Whether these responses are based on affective responses, irrational or rational beliefs, irrational or rational responses to evidence (or lack of it), other causes, or a combination of factors is not always distinguished however.

For example, Nass et al. (2000) found that that when we encounter an entity that behaves like a human our brains default to treating it as such, 'mindlessly' applying social rules and social behaviours such as politeness and reciprocity without even being aware of doing so, and even responding to indicators of 'personality'. Surprising as this seems, these phenomena have been observed even when participants deny that the technology that they were dealing with had any human traits or characteristics. It is possible that there are alternative explanations of these observations e.g. that participants in their study were thinking of the programmer rather than the computer when responding—though Nass and Moon deny this.

There is also evidence that Artificial intelligence (AI) does not even need to closely mimic humans—either physically or in terms of behaviour—to elicit such responses. For example, Sung et al. (2007) reported that the autonomous Roomba vacuum cleaner, which has neither human-like appearance or behaviours, can trigger social emotions, and Breazeal (2003, p. 151) noted that when a robot can physically mimic even very basic human expressions, humans interacting with it see its actions as "the product of intents, beliefs, desires, and feelings" and respond accordingly.

Such responses may even lead to trust. For example, the ELIZA program was an early form of chatbot that was built to mimic a Rogerian psychotherapist. It did no more than decompose phrases in the input and present them back to the user in a way that would sustain conversation. However, despite this limited functionality, users believed that the program had real understanding of what they were saying—and the creator of the program even found his secretary confiding sensitive personal information to it (Weizenbaum 1984).

In conclusion, the grounds for trust that I have described in this section would allow a trustee to be trusted without necessarily having the capacities that would give them moral status equivalent to that of a human. If trustees can be trusted without any evidence of their attributes i.e. on a purely subjective basis, then this is clearly the case. In the case of trust that is based on some evidence of an actual relationship between trustor and trustee this requires more explanation. That is, in this case a trustee would need to understand the requirements of their relationship with the trustor and respond accordingly i.e. they need to have some capacity for loyalty and hence of differentially valuing actions to at least some extent– but do not require a full capacity for moral reasoning. If moral status is scalar (Jaworska and Tannenbaum 2018) then as Arneson (1999) has suggested, it is not just the capacity to value that makes a difference to moral status, but also the level of this capacity. As such, a trustee with a more limited ability to prioritise actions within the scope of their relationship with a trustor would be accorded less moral status than a trustee who was of good character more broadly and could therefore value a greater range of things, and more fully.

As a consequence, contrary to Simpson's first claim, it is plausible that alternative accounts of rational normative trust based at least partially on subjective grounds are likely to be more common in the context of a counter-insurgency engagement and could allow robot forces to be trusted whilst still offering a moral saving over the use of human ones. Having established this as a possibility, I next make my case against his second main claim i.e. that such forces would nonetheless be ineffective because their use would undermine civilian trust in those that deployed them.

However, before proceeding, I note that none of the accounts of trust that have been considered here can provide a reason why a force made up primarily of autonomous robots would be *generally* more or less likely to be trusted. According to Simpson's account, trust depends on the trustee having certain attributes, but different types of robots have very different capabilities and therefore may be more or less credible as subjects of trust. Alternatively, if trust is based on the attitude of the trustor then commonly held beliefs about technology could make normative trust of robots as a group more (or less) likely regardless of their actual attributes—but such beliefs may not be consistent across different individuals and cultures making it again difficult to generalise.

Nonetheless, evaluating the potential for different forms of trust, and under different conditions, is still a useful exercise as the basis for scenario planning. That is, it allows the implications of these accounts to be considered in advance in case the conditions for them come to be met in the future e.g. through advances in robot capabilities, or engagement in counter-insurgencies amongst civilians with more robo-friendly beliefs. Consideration of such scenarios—even if speculative—allows the potential ethical consequences of such trust to be considered, as well as the structures and measures that might be required to address them. It also provides an opportunity to re-direct the development of the technology if there are unacceptable risks involved.

Moral savings do not undermine trust in the deploying entity

Simpson's second and related claim is that robot forces would be ineffective even if they could be trusted whilst offering a moral saving, because their use would signal to civilians that the deploying entity could not be trusted to be committed to the conflict. I argue that this claim is also unconvincing because it makes two unjustified assumptions. Firstly, it assumes that robot forces will always be distinguishable from human ones. However, if they are not e.g. if they are heavily armoured, operated remotely but by other robots, or eventually resemble humans—and have sufficiently human-like behaviours—then they will not send signals to civilians that are any different to human forces.

Secondly, it assumes how civilians would interpret the use of robot forces—though this is speculative and likely to be confirmable only through empirical methods. Without further justification it is equally likely that the use of robots could indicate a commitment to a long-term presence in the country because a lack of human fatalities would reduce political pressure in the home country for withdrawal. Furthermore, if robot forces were, or were perceived to be, more competent than human equivalents in at least some aspects of their role, then civilians might feel that *failing* to deploy them would be an indication that the government did not have their interests at heart. In fact, it seems likely that the effect of the deployment of robots on trust in the governing body will depend on a whole range of factors. For example, Ferrin (2003) has identified as many as fifty determinants of trust levels and covariants of trust which could plausibly be contributory. These include the qualities of the trustor (such as the disposition to trust and familiarity with robots), qualities of the trustee (such as ability, integrity, reputation), characteristics of past relationships with the entities involved, characteristics of their communication (such as threats, promises, and openness as regards intentions), and structural parameters surrounding the relationship (such as governance arrangements).

Conclusions

To conclude, in "*Robots and War*", Simpson argues that wars amongst the people require counter-insurgency forces to win their 'hearts and minds'. He argues that this is not possible with armies made up solely of autonomous robots for two reasons. Firstly, because robot warriors cannot be rationally trusted because they cannot act from a motive based on good character. Secondly because the use of such a force would signal that the deploying body was not committed to resolving the conflict. I have disputed both claims.

Regarding the first, I have argued that being of good character is not necessarily material to whether we could trust robot warriors because it is unlikely to be either the most common basis for trust, or even feasible, necessary or desirable. They may instead be trusted on other grounds, which are less demanding than Simpson supposes, and do not require that they have attributes that would make them of equal moral status to humans i.e. such that their use could still offer a moral saving over the use of a human force. Regarding his second claim, I have argued furthermore that this moral saving would not necessarily undermine trust in the deploying body because how it is perceived would be influenced by a number of factors—such as the relationship between civilians and the deploying body, and levels of familiarity with the use of robot forces.

As such, robot forces may yet be trusted whilst offering an acceptable moral saving over the use of human ones, and could therefore potentially be effective in counter-insurgency engagements. Consequently, there may be a case for developing a more finely grained understanding of the opportunities for, and challenges of, their use. This should consider the wide range of factors that may be involved including the nature of the technology, the evolution of robot relationships with humans, and the variation between different counter-insurgency engagements. It seems unlikely that there will be one right answer.

Acknowledgements I would like to extend my thanks to Dr Kevin Macnish (University of Twente) for his advice, guidance and encouragement whilst writing this paper.

References

Arneson, R. J. (1999). What, if anything, renders all humans morally equal? In D. Jamieson (Ed.), *Singer and his critics* (pp. 103–127). Oxford: Blackwell.

Baier, A. (1994). *Moral prejudices: Essays on ethics*. Cambridge, Mass: Harvard University Press.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *The International Journal of Human-Computer Studies, 59*(1-2), 119–155.

Coeckelburgh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception and performance of artificial agents. *AI & Society, 24*(2), 181–189.

Cogley, Z. (2012). Trust and the trickster problem. *Analytic Philosophy, 53*(1), 30–47.

Egnell, R. (2010). Winning 'hearts and minds'? A critical analysis of counter-insurgency operations in Afghanistan. *Civil Wars, 12*(3), 283–303.

Faulkner, P. (2017). The problem of trust. In P. Faulkner & T. Simpson (Eds.), *The philosophy of trust*. Oxford: Oxford University Press.

Ferrin, D. (2003). Definitions of trust. In *Presentation to a workshop. Building and rebuilding trust: State of the science, research direction, managerial interventions*. Workshop organised by Kurt Dirks & Don Ferrin, Academy of Management annual meeting Seattle, WA.

Hardin, R. (2002). *Trust and trustworthiness*. New York: Sage.

Hollis, M. (1998). *Trust within reason*. Cambridge: Cambridge University Press.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72,* 63–76.

Jaworska, A., & Tannenbaum, J. (2018). The grounds of moral status. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Available from: https://plato.stanford.edu/archives/spr2018/entries/grounds-moral-status/. Accessed 23 August 2018.

Jones, K. (1996). Trust as an affective attitude. *Ethics, 107,* 4–25.

Kelly, T. (2016). Evidence. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Available from. https://plato.stanford.edu/archives/win2016/entries/evidence/. Accessed 12 June 2018.

Kirkpatrick, J., Hahn, E., & Haufler, A. (2017). Trust in human–robot interactions. In P. Lin, G. Bekey, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0*. Oxford: Oxford University Press.

Kramer, R. M. (1999). Stalking the sinister attribution error: Paranoia inside the lab and out. *Research on Negotiation in Organisations, 7,* 59–91.

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence and future directions. *Journal of Management, 32*(6), 991–1022.

Lokhorst, G., & van den Hoven, J. (2014). Responsibility for military robots. In P. Lin, K. Abrey, & G. A. Bekey (Eds.), *Robot ethics*. Massachusetts: MIT Press.

McGeer, V., & Pettit, P. (2017). The empowering theory of trust. In P. Faulkner & T. Simpson (Eds.), *The philosophy of trust*. Oxford: Oxford University Press.

Nass, C., Moon, Y., et al. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Pettit, P. (1995). The cunning of trust. *Philosophy & Public Affairs, 24*(3), 202–225.

Rosati, C. S. (2016). Moral Motivation In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Available from: https://plato.stanford.edu/archives/win2016/entries/moral-motivation/. Accessed 21 January 2018.

Simpson, T. W. (2011). Robots, trust and war. *Philosophy and Technology, 24,* 325–337.

Simpson, T. W. (2013). Trustworthiness and moral character. *Ethical Theory and Moral Practice, 16*(3), 543–557.

Sung, J.Y., Guo, L., Grinter, R.E., & Christensen, H. I. (2007). My Roomba is Rambo: Intimate home appliances. In: J. Krumm, G. D. Abowd, A. Seneviratne, & T. Strang (Eds.), *Proceedings of UbiComp 2007*. Innsbruck, Austria: Springer.

Weizenbaum, J. (1984). *Computer power and human reason: From judgement to calculation*. Harmondsworth, UK: Penguin.

Wright, S. (2010). Trust and trustworthiness. *Philosophia, 38*(3), 615–627.