«I Don't Trust You, You Faker!» On Trust, Reliance, and Artificial Agency

Fabio Fossa

Introduction

In Asimov's 1975 story *A Boy's Best Friend* Jimmy, a young dweller of Lunar City, instructs its robotic dog Robutt not to get out of his sight and exclaims: «I don't trust you, you faker!»[1]. Trust and distrust in robots and computers is indeed a recurring theme in many of Asimov's stories. In *Robbie*, for instance, Grace struggles to accept leaving her little daughter to the care of a robotic nanny, of which however her pupil Gloria becomes quickly fond. While discussing the matter with her husband, Grace exclaims: «I won't have my daughter entrusted to a machine – and I don't care how clever it is. It has no soul, and no one knows what it may be thinking»[2]. In *Reason*[3], Powell and Donovan wonders whether Dave, a robot that has concocted an absurd interpretation of its own condition, is to be held trustworthy. In *Point of View*[4], a smart child asks whether the Multivac, a supercomputer his father is working on, can be trusted even though sometimes it makes trivial mistakes.

As Asimov did not miss to notice, the impact of information technologies on trust relationships is deep and multifarious. The more human beings rely on technological products to accomplish their aims, the more the mediation provided by such technologies affects trust relationships and modifies their characters. Information technologies impinge on trust in at

[1] I. Asimov, *The Complete Robot*, Doubleday, New York 1982, p. 4.

[2] *Ivi*, p. 138.

[3] *Ivi*, pp. 227-244.

[4] *Ivi*, pp. 37-40.

DOI: 10.4454/teoria.v39i1.57

least two different situations. The first situation, that of *e-trust* or *online trust*, occurs when trustors and trustees are Human Agents (HAs) who get in touch through digital platforms, mostly the internet. In general, scholars working in this field of inquiry try to shed light on «the *definition* and the *management*» of e-trust[5]. Specific problems are, for instance, how trust can be secured in digital environment[6] and what connection exists, in online contexts, between trust and reputation[7] or knowledge[8].

The second situation, which may be labelled *robotrust*[9], occurs when human trustors put trust in artificial trustees. In this case, trust relationships are not supposed to concern human actors exclusively, but to occur between human users and technological products as well. In particular, similar relationships are thought to emerge when human beings delegate tasks to autonomous technologies, such as AI systems and robots. The basic idea underlying these inquiries is that, since relationships between human beings and Artificial Agents (AAs) happen to arouse expectations of trust, it is necessary to "update" our conception thereof in order to include AAs as possible trustees. Finally, some scholars claim that trust frameworks should also be applied to the study of mutual relationships between AAs in Multi-Agent Systems (MAS)[10]. We may name this last domain *artificial trust*.

The focus of this paper is on *robotrust*, i.e., on trust relationships between human and artificial agents (HA→AA). My aim is to clarify the extent to which such relationships can be framed in terms of trust. Usually, relationships between human beings and artefacts are not supposed to imply trust, but reliance. The situation, nonetheless, appears to be opposite

[5] M. Taddeo, *Modelling Trust in Artificial Agents, a First Step Toward the Analysis of E-Trust*, in «Minds and Machines», 20 (2010), pp. 243-257, p. 244.

[6] H. Nissenbaum, *Securing Trust Online: Wisdom or Oxymoron?*, in «Boston University Law Review», 81 (2001), n. 3, pp. 635-664.

[7] T. Simpson, *E-trust and reputation*, in «Ethics and Information Technology», 13 (2011), pp. 29-38.

[8] J. Simon, *The entanglement of trust and knowledge on the web*, in «Ethics and Information Technology», 12 (2011), pp. 343-355.

[9] U. Pagallo, *Robotrust and Legal Responsibility*, in «Knowledge, Technology and Policy», 23 (2010), pp. 367-379.

[10] M. Taddeo, *Modelling Trust*, cit.; J. Buechner, H.T. Tavani, *Trust and multi-agent systems: applying the "diffuse, default model" of trust to experiments involving artificial agents*, in «Ethics and Information Technology», 13 (2011), n. 1, pp. 39-51; F.S. Grodzinsky, K.W. Miller, M.J. Wolf, *Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?"*, in «Ethics and Information Technology», 13 (2011), pp. 17-27.

when AAs come under scrutiny. As shown in the following section, HA→AA relationships are often assumed to imply trust. I disagree with this assumption and suggest that it would be more accurate to frame *direct* HA→AA relationships in terms of reliance instead. In a word, I argue that the relationship between us and our artefacts should be interpreted in terms of reliance even when AAs are involved.

To some extent, confusion on this matter may arise since trust characterises the social milieu in which relationships between human beings and autonomous technologies occur. AAs, in fact, can also be conceived of as mediums of human actions, even though in a different way compared to how technological platforms mediate human activities in e-trust scenarios. As entities to which tasks are delegated, AAs *indirectly* mediate trust between users and other social actors involved in their design, manufacture, commercialisation, and deployment. In this sense, AAs mediate *social* trust. However, the fact that AAs mediate trust relationships between social actors does not imply that direct HA→AA relationships can or should be understood by reference to trust. The two relationships are different from each other and must not be confused.

The rest of the paper is structured as follows. Section 2 shows in what sense trust relationships between HAs and AAs have been perceived as requiring to be acknowledged rather than proved. This presupposition, however, is problematic. As argued in Section 3, in fact, trust is seldom distinguished from reliance when HA→AA relationships are discussed. Yet, since relationships between human beings and technological products are normally framed in terms of reliance, interpreting HA→AA relationships as trust relationships implies asserting that the concept of reliance does not suffice here. This, in turn, is a questionable assumption. Section 4, then, focuses on task delegation to show that AAs can only arouse expectations of reliability, i.e., not the sort of expectations that require trust to be adequately met. Therefore, placing trust in AAs appears as a form of anthropomorphism, which may result in deception and social harms. Finally, Section 5 tries to determine the extent to which AAs mediate social trust between human actors such as designers, engineers, programmers, companies, and end-users. Addressing issues related to this kind of technologically mediated trust will probably be one of the most compelling future challenges for policy makers and social institutions.

1. *A matter of fact?*

At first glance, it may seem obvious that we necessarily entertain trust relationships with AAs as soon as they enter the social stage[11]. Indeed, the social pervasiveness of trust has been strongly underscored by Luhmann[12] and repeatedly stressed ever since. Unlike traditional tools, AAs are capable of executing complex functions without supervision. This ability invests them with an ambiguous social status which falls somewhere in between that proper to things and that proper to people. Since AAs take an active part in the social organisation of work, as humans do, it is easy to see the reason why trust may seem to be required. Trust is widely recognised as one of the most fundamental elements in the organisation of complex activities. Enabling task allocation and coordination, trust allows sparing time and resources to be reassigned to new undertakings. Placing trust in others to carry out tasks aligned to a final purpose is likely to be the most effective way to face complexity and cope with multiple challenges successfully. Society cannot do without delegation, and delegation seems to require trust.

Even if the relationships between human beings and artefacts has been usually framed in terms of reliance, many authors bring trust into play when it comes to AAs. As information technologies become more fine-tuned and versatile, AAs naturally appear as suitable substitutes for human trustees. Technological products that can carry out tasks without requiring constant human oversight or intervention are great candidates for delegation. After all, robots have always been envisioned as possible substitute for human delegatees[13]. As of now, delegation of tasks to AAs is already a well-established practice in contexts as different as producing goods in factories, running driverless train systems, or providing basic costumer support. Most likely, this trend will not reverse itself soon; and the more we cooperate with AAs or let AAs operate in our place, the more it may seem sensible to think of them as trustees and of ourselves as trustors.

Besides, the rise of automation has also caused the pairing of artefacts and reliance to be surprisingly challenged. From this controversial perspective, trust is considered to be a constitutive element in relations

[11] B. Kuipers, *How can we trust a robot?*, in «Communication of the ACM», 61 (2018), n. 3, pp. 86-95.

[12] N. Luhmann, *Trust and Power*, John Wiley and Sons, Chichester 1979.

[13] N. Wiener, *The Human Use of Human Beings*, Houghton Mifflin Company, Boston 1950.

between human users and any artefact to which tasks are delegated[14]. For instance, Taddeo writes that we may trust elevators to lift us safely to our floor[15] just as, I might add, we may trust thermostats to keep the temperature constant and washing machines to wash our clothes. According to this viewpoint, delegating tasks to artefacts implies placing trust in them regardless their degree of complexity. In the context of delegation, then, trust should be primarily understood as a delegators’ attitude towards any entity, be it a subject or an object, capable of carrying out specified tasks. In this sense, trust is «a *property of relations*», and thus also a property of delegation, that indicates the minimisation of «effort and commitment for the achievement» of the trustor’s goal[16]. As such, trust may involve any kind of delegatee, technological products included: «As digital technologies evolve and become more refined and effective, our expectation has become an expectation to *trust* (by delegating and not supervising) them with important tasks»[17].

Other authors conceive trust as a relational dimension that may include technological products. In Coeckelbergh’s opinion, for example, trust is placed on AAs mostly in light of the peculiar position they occupy in human society. «If a human-robot relation grows as a social relation», Coeckelbergh writes, «then trust is already there as a ‘default’ in the social relation»[18]. From this perspective, trust appears as «an emergent and/or embedded property»[19] that belongs more to delegation as a social relationship rather than to the minds of the subjects who set purposes, delegate tasks, and choose to trust. In the end, the connection between delegation and trust must be traced back to the correlation of social relationality in general – of which delegation is a case – and trust. Therefore, «in so far as robots are already part of the social and part of us, we trust them as we are

[14] B. Latour, *Where are the missing masses? The sociology of a few mundane artifacts*, in W.E. Bijker, J. Law (eds), *Shaping Technology-Building Society. Studies in Sociotechnical Change*, MIT Press, Cambridge 1982, pp. 151-180.

[15] M. Taddeo, *Defining Trust and E-trust: From Old Theories to New Problems*, in A. Mesquita (ed.), *Sociological and Philosophical Aspects of Human Interaction with Technology*, Information Science References, Hershey 2011, p. 24.

[16] M. Taddeo, *Trusting Digital Technologies Correctly*, in «Minds and Machines», 27 (2017), n. 4, pp. 565-568, p. 565.

[17] *Ivi*, p. 566. For similar considerations see also M. Taddeo, L. Floridi, *How AI can be a force for good*, in «Science», 361 (2018), n. 6404, pp. 751-752.

[18] M. Coeckelbergh, *Can We Trust Robots?*, in «Ethics and Information Technology», 14 (2012), pp. 53-60, p. 58.

[19] *Ivi*, p. 56.

already related to them»[20]. Trusting AAs does not seem to be entirely a matter of choice; rather, it appears to represent a matter of fact that necessarily follows from delegation.

Herman Tavani tackles the issue of trusting AAs in a similar vein. Building on Walker's notion of zone of default trust[21], Tavani proposes to focus on the practical contexts in which agents «come to know 'what to expect' from others and 'whom to trust'»[22]. Inside a zone of default trust, normative expectations concerning the way in which delegatees should behave emerge by disposition and, rather than concentrating on specific individuals, may diffuse on more or less undefined entities. Such expectations arise in the delegating subjects in virtue of the social situation itself, which is intrinsically determined by trust. Therefore, as long as AAs can successfully occupy the place usually reserved for human trustees, they are already «capable of being in trust relationships with human beings»[23]. From this perspective, in fact, «we can now speak of cases of various kinds that are intrinsically different, but whose common feature is that they involve a zone of default trust», so that «the concept of a zone of trust can do much of the work in assimilating a wide range of disparate cases»[24]. Trust, consequently, appears to be both a subjective disposition of a normative nature, which enables and supports task delegation, and a correlative dimensional property, which defines the features of a relational 'zone'. In sum, «HAs can enter into trust relationships with several different kinds of AAs, simply in virtue of the nature of the default and the diffuse-default zones (of trust) involved»[25]. Again, trust in AAs seems to be a simple matter of fact.

These approaches to *robotrust*, however thought-provoking they might be, are still too coarse-grained and risk masking important differences between delegation to AAs and delegation to HAs. It is certainly correct to state that AAs to which tasks are delegated occupy the social position normally reserved to human trustees. Nevertheless, once the substitution of HAs with AAs occurs, the general relationship between delegator(s) and

[20] *Ivi*, p. 59.

[21] M.U. Walker, *Moral repair: reconstructing moral relations after wrongdoing*, Cambridge University Press, Cambridge 2006.

[22] H.T. Tavani, *Levels of Trust in the Context of Machine Ethics*, in «Philosophy of Technology», 28 (2015), n. 1, pp. 75-90, p. 79.

[23] *Ivi*, p. 76.

[24] J. Buechner, H.T. Tavani, *op. cit.*, p. 42.

[25] H.T. Tavani, *op. cit.*, p. 81.

delegatee(s) requires to be equally reassessed. Even if the context in which delegation occurs is analogous, the two premises that we trust human beings, when we delegate tasks to them, and that AAs can substitute HAs as task executers, do not immediately imply that it is legitimate to place trust in artificial delegatees. The fact that, inside a zone of trust, «one simply engages in that behaviour, with little or no conscious reflection»[26] is not a justification of the behaviour itself – especially when one constitutive element of the situation in which the behaviour takes place is substituted by another that imitates it. In this case, on the contrary, it is crucial to maintain a critical focus on habitual trust attitudes to avoid deception and misplaced expectations. Even if it may feel natural to trust AAs, the question whether it makes sense to do so remains both relevant and unanswered. For this reason, it is still necessary to address the question: Can we trust AAs?

2. *Trust and Reliance*

In the context of HA→AA task delegation, the substitution of human trustees with technological products impacts significantly on the overall character of the relation. Therefore, the reasons to frame HA→AA relationships in terms of trust are not self-evident and require discussion. It thus becomes necessary to clarify whether it is accurate to transfer trust from forms of delegation that involve exclusively human beings to forms of delegation that encompass technological products as well. To this end, it must be determined first why trust is needed in HA→HA task delegation and, secondly, whether or not the same need arises in HA→AA task delegation. The point of the analysis, then, would consist in verifying whether the reasons why trust emerges in HA→HA forms of task delegation also occur in the case of HA→AA forms of task delegation. If the answer is positive, it is accurate to transfer trust from human to human-artificial contexts. Else, if the answer is negative, the concept of *robotrust* may need to be revised.

A similar enquiry is required since the theoretic decision of describing HA→AA task delegation in terms of trust necessarily implies that the usual way of understanding human relations to technological products has become unsatisfactory. As many scholars note[27], relationships between hu-

[26] J. Buechner, H.T. Tavani, *op. cit.*, p. 43.

[27] M. Dzindolet *et al.*, *The role of trust in automation reliance*, in «International Journal of Human-Computer Studies», 58 (2003), pp. 697-718; P.J. Nickel *et al.*, *Can We Make Sense of the*

man users and technological products are usually framed in terms of *reliance*. Reliance might be said to indicate a property of relations to tools that refers directly to the function that a tool is supposed to carry out. Reliability, in turn, indicates the capacity of a tool to achieve the ends it is built to serve or, which is the same, «the ability of the item to remain functional»[28], thus forming, by being available, «the basis of new relations between its users and their environment»[29]. Accordingly, «we expect the artefact to function, to do what is meant to do as an instrument to attain goals set by humans»[30]. Deciding to frame HA→AA task delegation by reference to trust implies that, in such relations, something *more* is at stake that cannot be accounted for only by reference to the functional notion of reliance. What is this additional element?

Trying to answer this question is critical not only because of the reasons previously exposed, but also on the account that trust and reliance are not easily distinguishable. Although the importance of differentiating between trust and reliance is often stated, the word "trust" is just as often used as a synonym of "reliance". More precisely, the word "trust" appears to include the meaning of the word "rely" among its possible usages[31]. This is, beyond any doubt, what Taddeo means when she writes that we trust elevators. Similarly, Coeckelbergh[32] speaks of «'trust as reliance'»; Kiran and Verbeek, while discussing reliability, write that «tools can only be used for doing something if they are trustworthy»[33]; and Pitt defines untrustworthy technologies as «products that do not performed as promised, that break easily»[34]. In sum, as Nickel *et al.* observe, sometimes «what it means to

Notion of Trustworthy Technologies?, in «Knowledge, Technology and Policy», 23 (2010), pp. 429-444; K. Hawley, *Trust. A Very Short Introduction*, Oxford University Press, Oxford 2012, pp. 3-6.

[28] A. Birolini, *Reliability engineering. Theory and practice*, Springer, New York 2007, p. 2. Reliability is first of all a technical notion. However, once a technological product is deployed in social contexts, the technical measurement of its reliability is psychologically reinterpreted – sometimes in wrong ways. Distorted perceived reliability may lead to disuse due to underutilisation or misuse due to complacency; well perceived reliability leads to correct use and appropriate reliance and thus must be enforced (M. Dzindolet *et al.*, *op. cit.*; R.R. Hoffman *et al.*, *Trust in Automation*, in «IEEE Intelligent Systems», 28, 2013, n.1, pp. 84-88).

[29] A.H. Kiran, P.-P. Verbeek, *Trusting our Selves to Technology*, in «Knowledge, Technology, and Policy», 23 (2010), pp. 409-427, p. 422.

[30] M. Coeckelbergh, *op. cit.*, p. 54.

[31] P. Pettit, *Trust, Reliance and the Internet*, in «Analyse & Kritik», 26 (2004), pp. 108-121.

[32] M. Coeckelbergh, *op. cit.*, p. 54.

[33] A.H. Kiran, P.-P. Verbeek, *op. cit.*, p. 410.

[34] J.C. Pitt, *It's not about technology*, in «Knowledge, Technology and Policy», 23 (2010), pp. 445-454, pp. 450-451.

trust (to a certain degree) a technical artefact [...] is more or less identical with what it means to rely (to a certain degree) on it»[35]. Sure enough, it would be of little significance to deny legitimacy to such usage. Nonetheless, the two words stand for different relationships that presuppose different scenarios and are based on different expectations, so that confusion on this point should be carefully avoided. It remains important, therefore, to reflect upon what is actually meant in these cases by the word “trust”, and if something else is meant when the same word is applied in exclusively human contexts.

In order to clarify the role of trust in task delegation, let us consider HA→HA relations, where trust is commonly acknowledged as a constitutive element. When a person delegates a task to someone else, it is sensible to expect that she carries out an evaluation of the delegatee’s overall adequacy to the task. Such adequacy is at least twofold: as Baier explains, «trust (...) is reliance on others’ competence and willingness to look after, rather than harm, things one cares about which are entrusted to their care»[36]. The delegatee then must be *able* and *willing* to execute the appointed task[37]. Delegation will be successful if, and only if, the delegatee is both capable of carrying out the task appointed to her and inclined to commit to the delegator’s requests. The first aspect concerns the delegatee’s resources and skills, whilst the second pertains to her will or intent[38].

In delegation, reliability and trust emerge within these two dimensions. Reliability refers to the skills, abilities, and expertise that the delegatee possesses and exercises once delegation has occurred. A reliable person, regardless her trustworthiness, is competent, i.e., has what it takes to succeed in carrying out the task. In a sense, when someone’s reliability is under scrutiny, she is already – even if partially – thought of as if she were a machine. She is indeed evaluated as an executer of predetermined tasks, i.e., of functions[39]. Reliability, as a measure of efficiency, is essentially relative to functional performances: when attributed to human beings, it

[35] P.J. Nickel *et al.*, *op. cit.*, p. 435.

[36] A. Baier, *Trust and Antitrust*, in «Ethics», 96 (1986), n. 2, pp. 231-260, p. 259.

[37] P. Pettit, *The cunning of trust*, in «Philosophy and Public Affairs», 24 (1995), n. 3, pp. 202-225; L.J. Camp *et al.*, *Trust: A Collision of Paradigms*, in P. Syverson (ed.), *Financial Cryptography*, FC 2001. Lecture Notes in Computer Science, vo. 2339, Springer, Berlin-Heidelberg 2002, pp. 91-105.

[38] A. Baier, *op. cit.*, p. 234.

[39] P.J. Nickel *et al.*, *op. cit.*, pp. 433-434; P. Pettit, *Trust, Reliance and the Internet*, cit., pp. 109-110.

indicates how well a person is able to serve as a means to a predetermined end in circumscribed contexts.

Trust, on the contrary, pertains to the second dimension of task delegation. Since human beings are free to choose what purposes to tend to, the delegator must secure the delegatee's commitment. Moreover, human beings can *fake* to tend to purposes other than those they actually tend to, so that assurance is even more required. Placing trust on the delegatee, who thus becomes a trustee, the delegator/trustor projects normative moral expectations on to the trustee's future behaviour[40]. In doing this, the trustor appeals to the trustee's sense of responsibility[41], impelling her to the task. As Luhmann writes, trust «serves to overcome an element of uncertainty in the behavior of other people which is experienced as the unpredictability of change in an object»[42]; or, in other words, its function is «the reduction of complexity in the face of the freedom of the other person»[43]. Trust puts social and moral pressure on the trustee, who is consequently motivated to align her own purposes to the trustor's ones. In the context of delegation, hence, trust is required when a delegatee, who is able to carry out the task, must also be persuaded to do so, since she may be uninterested in the delegator's demands or may fake interest, while having other purposes in mind. In this situation, trust adds an additional element to the relation between delegators and delegatees, which has the specific aim of motivating the latter to align their purposiveness to the former's one. Does the same need arise in HA→AA task delegation?

3. *Betrayal, Disappointment, and Robotrust*

Before addressing this question directly, it is necessary to spend few more words on why human beings place trust in order to delegate successfully. As already noted, trust enforces commitment, and commitment assures that human delegatees have assumed the delegators' end as their own. Since, when tasks are delegated, the delegatee's purposiveness may determine itself in counterproductive ways, measures must be taken so that it adjusts properly. In this context, trust is meant to influence the

[40] J. Buechner, H.T. Tavani, *op. cit.*, pp. 41-42.

[41] S.D.N. Cook, *Making the Technological Trustworthy: on Pitt on Technology and Trust*, in «Knowledge, Technology, and Policy», 23 (2010), pp. 455-459.

[42] N. Luhmann, *op. cit.*, p. 22.

[43] *Ivi*, p. 62.

delegatee's choice by placing a normative obligation on to her that appeals to her sense of responsibility. Feeling responsible not only for the task that has been appointed, but also for the trust that has been placed, the trustee is encouraged to carry out the delegated task and dissuaded to overwrite the trustor's purpose.

Consequently, in the context of HA→HA task delegation it is presumed that human beings enjoy a direct relationship to ends, i.e., that humans are self-determined purpose-setting agents. The delegator's attitude to trust stems from the presupposition that the delegatee has both the power to choose spontaneously among competing ends and preferences of her own. From such assumption follows that the delegatee may be uninterested in assuming the delegator's purpose or may fake to do so, while actually serving other purposes. The delegatee thus needs to be motivated so that she may truly take on ends set by somebody else. Trust, then, is a strategic response to nudge the unpredictable will of others by means of ethical and social pressure. Trust is required in HA→HA task delegation exclusively because it is assumed that HAs are *free* of determining their own will and choosing among different ends.

When a human delegatee, who acts as if she has taken on a given task, does not truly commit to the trustor's aim or has a personal agenda that collides with it, a breach of trust occurs. If a trustee fails to carry out the task entrusted to her, even though she was perfectly capable of executing it, the trustor feels *betrayed*[44]. In task delegation, betrayal is the accusation that trustors throw to noncomplying trustees, that is, to trustees who determine their own will regardless of their commitment to the delegators' purpose. Since normative moral expectations were placed on the trustee's behaviour, the trustor has the right to complain, to hold the delegatee morally responsible for breaching trust, to ask for justification or explanation, and perhaps even to take offence. In sum, the notion of trust in the context of HA→HA task delegation describes a property of such relation that has the function of minimising betrayal. This result, in turn, is accomplished through a dialectic of normative moral expectation and reactive responsible behaviour that nudges the delegatee's will in a way that fosters successful delegation.

In light of this, asking whether trust is needed in HA→AA task delegation equals asking whether AAs are capable of choosing autonomously among different ends and, thus, whether they must be motivated accord-

[44] A. Baier, *op. cit.*, p. 235; H. Tavani, *op. cit.*, pp. 86-88.

ingly so as to align their preferences to the delegator's ones. In other words, what must be clarified is whether AAs are self-determined purpose-setting entities, since only such entities can betray, thus making trust necessary. If yes, trust is required in HA→AA task delegation just as it is required in HA→HA task delegation, and it is correct to transfer trust from human to artificial delegatees. If not, AAs would be entities that serve predetermined ends. Thus, the only dimension of task delegation that would remain would be that of ability or efficiency. In this case, reliance should suffice to understand HA→AA task delegation and the introduction of trust would be spurious.

In my opinion, it is not possible to think AAs as entities which can spontaneously choose among competing ends and betray, since they do not exhibit a direct relation to purposes. No AA «*has* purpose and *acts on* purpose»[45] the way we do. While human beings are purpose-setting entities – or at least are supposed to be so in the context of task delegation – AAs are «purpose-built artifacts»[46] as any other technological product and there is no need to assume them to be anything more. Even though AAs display the distinguishing feature of being able of executing functions independently from human oversight or intervention, they are still fully understandable by reference to the category of tool[47]. As any other tool, AAs require a specified end to serve in order to be devised, designed, and manufactured. It is impossible to think AAs apart from the specific purposes they are built to serve – which are, therefore, always *given*.

Although AAs can carry out functions autonomously and, consequently, partially unpredictably (as no previous tool could), still they do not display the possibility of setting ends by themselves nor of intentionally serving unpredictable ends. Accordingly, tasks are delegated to AAs only in virtue of their efficiency in achieving ends that are valuable for their users. The range of AAs' autonomy and unpredictability extends exclusively to the execution of functions, that is, to the way in which given ends are accomplished. AAs serve a purpose or clusters of purposes that can always be traced back to their designers. At the same time, this purpose or these clusters of purposes identify with the reasons why they appear useful. AAs

[45] H. Jonas, *The Phenomenon of Life. Towards a Philosophical Biology*, Northwestern University Press, Evanston 2001, p. 119.

[46] J.J. Bryson, P. Kime, *Just an Artifact: Why Machines are Perceived as Moral Agents*, https://www.cs.bath.ac.uk/~jjb/ftp/BrysonKime-IJCAI11.pdf [accessed 1 October 2018], p. 1.

[47] F. Fossa, *Artificial Moral Agents: Moral Mentors or Sensible Tools?*, in «Ethics and Information Technology», 20 (2018), pp. 115-126.

are purpose-built artefacts that exist only within socio-technical contexts where ends are set and pursued[48].

As it seems sensible to frame AAs as purpose-built artefact, it seems also sensible to deny the possibility that AAs can betray[49]. Lacking the capacity of setting ends autonomously, AAs can neither be uninterested in the task they execute nor fake interest in the task, while intentionally serving other purposes in secret[50]. The alignment of any AA to the end of the function it executes is in fact merely a matter of design. Well-designed AAs will carry out their task as they are supposed to, whilst poor-designed AAs will not. In the case of HA→AA task delegation, then, there are no conditions for trust to emerge. Accordingly, true betrayal cannot occur in this context, but can be experienced only metaphorically.

When an AA fails to achieve the goal it is programmed to pursue, users ought not to interpret this failure in terms of betrayal, but rather in terms of *disappointment*[51]. Disappointment refers to functional expectations that are not met and, as such, is the appropriate reaction to reliability issues. AAs that repeatedly disappoint their users are unreliable – and "untrustworthy" only in this technical sense. Accordingly, it would be irrational to take offence at AAs or holding them morally responsible for failing to fulfil their commitment[52], just as it would be irrational to take offence at an elevator or hold it morally responsible in case of incident. Unreliable AAs can either be discarded or fixed so that they carry out their function as originally intended, in the most effective way possible. There are no untrustworthy AAs, just malfunctioning or poorly designed ones. In HA→AA task delegation, only the dimension of reliability emerges.

For this reason, in the case of task delegation there is no actual need to

[48] D. Johnson, *Computer Systems. Moral Entities, but Not Moral Agents*, in «Ethics and Information Technology», 8 (2006), n. 4, pp. 168-183.

[49] J. Simon, *op. cit.*, pp. 346-347; A. Van Wynsberghe, S. Robbins, *Critiquing the Reasons for Making Artificial Moral Agents*, in «Science and Engineering Ethics» (2018), https://doi.org/10.1007/s11948-018-0030-8 [accessed 1 October 2018].

[50] This does not mean, of course, that no technological product may run some functions explicitly while at the same time running other functions implicitly which do not align with the user's intentions. When this happens, however, it is not due to a lack of motivation in the AA, but it happens *by design*. Such AA would still be reliable, since it would do what is supposed to; however, the *social description* of the AAs would be untrustworthy, since it would not inform the users on what the AA does. This problem will be discussed in Section 5.

[51] L.J. Camp *et al.*, *op. cit.*; J.C. Pitt, *op. cit.*

[52] P. de Laat, *Trusting the (Ro)botic Other: By Assumption?*, in «SIGCAS Computer and Society», 45 (2015), n. 3, pp. 255-260.

move from reliance to trust once AAs are involved, even though they are, in a sense, autonomous and unpredictable entities. Framing HA→AA task delegation in terms of reliance rather than in terms of trust is not only more accurate, but also safer since it prevents anthropomorphism and the many social risks deriving from it[53]. Since trust is essentially linked to the need of steering someone else's will through motivation, placing trust in AAs would presuppose the existence of a practical dimension, that of purpose-setting freedom, which does not belong to artificial agency. On the contrary, such practical dimension characterises human agency, so that projecting purpose-setting freedom onto AAs would result in humanising them. Consequently, the relation with AAs might appear to require measures and precautions that, yet, would be misplaced. Moreover, possible malfunctions would risk being mistakenly charged with moral meaning while, at the same time, ill-suited moral expectations would develop. As Bryson notes, anthropomorphising AAs «invites inappropriate decision such as misassignations of responsibility and misappropriations of resources»[54]: framing HA→AA task delegation in terms of trust would probably risk a similar social effect. Finally, understanding AAs as direct objects of trust might divert attention from the social context in which HA→AA task delegation occurs – i.e., from the context related to AAs where trust does play a crucial role.

4. *Artificial Agents and Social Trust Mediation*

In light of what has been said, it proves more accurate to interpret direct HA→AA relations by reference to the notion of reliance, rather than to the notion of trust. Strictly speaking, trust does not pertain to the relationship between users and artefacts, though advanced they may be. To some extent, however, confusion on this matter may arise since trust permeates the social context in which HA→AA task delegation occurs. While executing delegated tasks, in fact, AAs mediate not only human agency, thus saving time and resources, but also trust relationships between various social actors. Being always embedded in social contexts, AAs become inter-

[53] P.J. Nickel *et al.*, *op. cit.*

[54] J.J. Bryson, *Robots Should Be Slaves*, in Y. Wilks (ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, John Benjamins Publish Company, Amsterdam 2010, pp. 63-74, p. 64.

section points of ethical normative expectations and responsibilities. In this *indirect* sense – i.e., as social trust mediators – AAs are the cornerstone on which trust relations between social actors are built. Such relations, moreover, play a critical role in shaping the social attitude towards automation, so that it is crucial neither to overlook such dimension nor to let it fade behind the misconception of direct HA→AA relationship as involving trust. Trust should not be mistakenly extended from the social milieu to the HA→AA relation itself: the two levels must not be confused.

In this light, a discussion concerning social trust as mediated by AAs is both possible and extremely relevant. However, such discussion can be properly carried out provided that, in the study of direct HA→AA relationships, trust is set aside. Only once it has been clarified why AAs cannot be trustees it becomes possible to ask who is truly trusted, when tasks are delegated to AAs.

In order to clarify in what sense AAs mediate social trust it is necessary to take a closer look to HA→AA task delegation. When a person considers whether to delegate tasks to an AA, it is rational to expect that she would try to establish if the AA in question "will do", that is, if it is capable of executing the desired task. Such evaluation concerns the AA's efficiency: it ponders over what purpose the AA is supposed to serve and how effective it is supposed to function. However, how can one know what a particular AA is supposed to do? Either the delegator has a deep understanding of the technology involved – which is arguably a rare case – or she will have to turn to a nontechnical description of the AA, to which she can meaningfully relate[55]. It follows that delegation will be most likely grounded on a description of the AA provided by those who happen to have the necessary expertise to express the AA's specifics in common language. Therefore, trusting the product means trusting the nontechnical description of its functions or utility; and this, in turn, means trusting the social actors who provide such description.

In HA→AA task delegation, then, direct relations between users and technologies are embedded in indirect relations between users and those who provide nontechnical descriptions of AAs. For the sake of the present argument, I will address those social actors as "stakeholders"[56]. Trust

[55] W. Pieters, *Explanation and trust: what to tell the users in security and AI?*, in «Ethics and Information Technology», 13 (2011), pp. 53-64.

[56] With the label "stakeholders" I mean all the subjects who are in different degrees involved in providing end-users with a comprehensible description of technological objects. In this specific sense, the label may apply to designers, programmers, engineers, advertisers, firms,

relations between end-users and stakeholders revolve around the adequateness of the AA's nontechnical description[57]. If this description is satisfying, the users will not need to worry about anything else than the AA's reliability. However, the description may be flawed or biased. For example, it may turn out that the AA executes other functions than those described, that it employs other means than those indicated, or that it also carries out undisclosed tasks. If any of these (or other) cases occur, users may reasonably feel betrayed; and since betrayal is a marker of trust, it suggests that the relation between end-users and stakeholders is one of trust. Inadequate descriptions lead to breaches of trust and, to this extent, causes AAs to appear untrustworthy (even if, perhaps, reliable).

The relation between users and stakeholders may be characterised as a case of HA→HA task delegation. HA→AA task delegation always occurs within a social context where AAs are presented to the public, their utility and features are advertised, and delegation itself is often encouraged. Knowingly or unknowingly, users delegate to stakeholders the task of providing an adequate description of the product they offer. This task, in fact, cannot be performed directly by the end-users, since they usually lack the necessary knowledge; and even if they could, to analyse meticulously every device one would want to use would still be extremely demanding in terms of time and resources. When AAs pass from the stakeholders' on to the end-users' hands, they bring along a description of the functions they carry out and the ends they serve, which translates the specifics of the artefacts in a user-friendly language. This description is the result of a human activity; and human beings can fake interest in delegated tasks, while intentionally serving other ends. Therefore, in this practical situation the conditions for trust apply. Users (as trustors) entrust to stakeholders (who become trustees) the task of providing a nontechnical description of the AA that would not be biased, incomplete, malevolent, or opaque. Placing

companies, institutions and so on. However, also the work of science communicators, journalists, and artists deeply influences the social understanding of technological products, which is then constantly negotiated. In its present form, the category is evidently ill-defined; nonetheless, it meets the need for which it is introduced in the argument. Further clarifications must be postponed. A similar use of the term "stakeholders" may be found in M. Hengstler *et al.*, *Applied Artificial Intelligence and Trust*, in «Technological Forecasting and Social Change», 105 (2016), pp. 105-120; P. de Laat, *op. cit.*; D. Pedreschi *et al.*, *Open the Black Box. Data-Driven Explanation of Black Box Decision Systems*, https://doi.org/0000001.0000001

[57] What elements make a nontechnical description "adequate" (transparency, honesty, completeness, clarity and so on) is an issue that requires much discussion and cannot be dealt with here.

trust, in this peculiar case, has the purpose of minimising the chance of betrayal by means of social and moral pressure. Hence, AAs indirectly mediate trust relationships between different social actors, i.e., *social trust*. "Untrustworthy" technologies are not such in themselves, but as devices made by untrustworthy producers or deployed by untrustworthy subjects.

The dimension of indirect trust mediation in HA→AA task delegation must not be overlooked, since much of the social acceptance of AAs depends from it[58]. Whether users will consider stakeholders trustworthy or not will affect their general disposition towards social robotics and AI systems in important ways. Low trust in stakeholders impinges significantly on the success of task delegation to AAs. Well-designed, reliable technologies will always appear in a suspicious light unless the companies and institutions behind them make an effort to earn the users' trust.

The trustworthiness of those who provide nontechnical descriptions of AAs represents undoubtedly a relevant issue to be addressed in the future from a social viewpoint. On this account, both ethical reflection and legal regulation must take on the task of indicating, recommending, and enforcing the right means to protect and maximise social trust. In conclusion, trusting AAs beyond reliance means trusting their nontechnical description and, thus, the social actors who provide it. Exclusively in this indirect, social sense, it seems possible to discuss trust, breach of trust, and distrust in situations involving AAs – which is entirely different from understanding AAs directly as trustees. Facilitating, protecting and enhancing trust between the human beings whose actions are practically mediated by AAs may be one of the most critical challenges posed by AI and robotics to future society.

Abstract

The aim of this paper is to clarify the extent to which relationships between Human Agents (HAs) and Artificial Agents (AAs) can be adequately defined in terms of trust. Since such relationships consist mostly in the allocation of tasks to technological products, particular attention is paid to the notion of delegation. In short, I argue that it would be more accurate to describe direct *relationships between HAs and AAs in terms of reliance, rather*

[58] M. Hengstler *et al.*, *op. cit.*; A.F. Winfield, M. Jirotka, *Ethical governance is essential to building trust in robotics and AI systems*, in «Philosophical Transactions A: Mathematical, Physical and Engineering Sciences», 2018 [in press].

than in terms of trust. However, as mediums of human actions to which tasks are delegated, AAs indirectly mediate trust between users and other social actors involved in their design, manufacture, commercialisation and deployment. In this sense, AAs mediate social trust. My conclusion is that relationships between HAs and AAs are thus to be understood directly in terms of reliance and indirectly in terms of social trust mediation.

Keywords: artificial agents; trust; robotrust; reliance; human-robot interaction.

Fabio Fossa
Università di Pisa
fabiofossa36@gmail.com