Minds & Machines (2017) 27:565–568
https://doi.org/10.1007/s11023-017-9450-5

Trusting Digital Technologies Correctly

Mariarosaria Taddeo[1,2]

Published online: 15 November 2017

Trust is a facilitator of interactions among the members of a system, whether these be human agents, artificial agents or a combination of both (a hybrid system). Elsewhere, I have argued that the occurrences of trust are related to, and affect, pre-existing relations, like purchasing, negotiation, communication, and delegation (Taddeo 2010a, b). Trust is not to be considered a relation itself but a *property of relations*, something that changes the way relations occur. Consider, for example, a case of communication. Alice talks to Bob and she informs him that the grocery store down the road is closed for the day As Bob trusts Alice, he believes her and decides not to walk to the shop to double check, instead he starts searching for an alternative place to shop for his groceries. Between Alice and Bob there is a *first-order relation*, the communication, which ranges over the two agents, and there is the *second-order property* of trust that ranges over the first-order-relation and affects the way it occurs.

As a property of relations, trust changes the way relations occur by minimising the effort and commitment for the achievement of a given goal of the agents who decides to trust (the trustor). It does so in two ways. First, the trustor can avoid performing the action necessary to achieve her/his goal her/himself, because s/he can count on the trustee to do it—Bob does not walk to the shop to check whether it is actually closed. Second, the trustor can decide not to supervise the trustee's performance. Bob does not ask Alice how she knows about the opening times of the shop. Delegation without supervision characterises the presence of trust (Taddeo 2010a, b). I define trust as follows:

✉ Mariarosaria Taddeo
mariarosaria.taddeo@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

[2] The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

> Assume a set of first order-relations functional to the achievement of a goal and that two agents are involved in the relations, such that one of them (the trustor) has to achieve the given goal while the other (the trustee) is able to perform some task in order to achieve that goal. If the trustor chooses to achieve his goal through the task performed by the trustee, and if the trustor considers the trustee a trustworthy agent, then the relation has the property of being advantageous for the trustor. Such a property is a second-order property that affects the first-order relations occurring between agents and is called trust (Taddeo 2010a).

It is its facilitating role that makes trust crucial for systems to work. Without trust, delegation would be much more problematic as it would require supervision. And this, in turn, would encroach on the distribution of tasks necessary for most systems to function. Imagine a society in which there is no trust in doctors, teachers, or drivers. This would require that all the members of the society to spend a significant portion of time and resources controlling the way others perform their tasks, at the expenses of their own tasks.

At the same time, not all systems require the same amount of trust to work and flourish. In the medium- and long-term, too little trust may affect the internal dynamics of the system and limit its development; but too much trust may lead to the collapse of a system if it results in a lack of control and coordination. Striking the right level of trust it is a delicate matter and requires considering the system (mature information societies, more on these presently), the expectations of the trustor (the tasks that human agents delegate), and the nature of the trustee (e.g. the digital technologies).

Mature information societies are hybrid systems, involving human and artificial agents. And trust is a crucial component of these systems. Trust among the members of information societies is transversal. It occurs among human agents. It also characterises some of the relations among artificial agents (Primiero and Taddeo 2012). Consider, for example, how your computer trusts your smart phone and exchanges information with it at all times. In mature information societies, human agents also trust some artificial agents. Floridi distinguishes mature information societies from immature information societies on the basis of *'their members' unreflective and implicit expectations* to be able to rely on digital technologies (Luciano Floridi 2016a). I agree with this view. It identifies a crucial, minimalist, criterion to identify mature information societies.

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. We trust machine learning algorithms to indicate the best decision to make when hiring a future colleague or when granting parole during a criminal trial; to diagnose diseases and identify a possible cure. We trust robots to take care of our elderly and toddlers, to patrol borders, and to drive or fly us around the globe. We even trust digital technologies to simulate experiments and provide results that advance our scientific knowledge and understanding of the world. This trust is widespread and is resilient. It is only reassessed (rarely broken) in the event of serious negative consequences.

Digital technologies are so pervasive that trusting them is essential for our societies to work properly. Supervising each run of a machine learning algorithm used to make a decision would require significant time and resources, to the point that it would become unfeasible to resort to these technologies. At the same time, however, the tasks with which we trust digital technologies are of such relevance that a complete lack of supervision may lead to serious risks for our safety and security, as well for the rights and values underpinning our societies. This is a lesson that we learned when discovering the implicit bias of COMPAS, and the breach that its deployment poses for the human right to fair trial.[1] COMPAS is a machine-learning based software that has been deployed in the US to assess the probability of a criminal defendant re-offending, and which has been shown to provide prediction that is strongly biased against black Americans, as ProPublica reports:

> black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).[1]

It is crucial to identify the correct way to trust digital technologies so that we can harness their value, while protecting fundamental rights and fostering the development of open, tolerant, just information societies (Floridi 2016b; Floridi and Taddeo 2016).

Digital technologies are not just tools to preform actions. Rather they are an interface through which we interact, change, perceive, and understand others and the environment surrounding us. These technologies share with the environment and with human agents the same informational nature (Floridi 2011). For this reason they can blend in the *infosphere* (Floridi 2002, 2014) to the point of becoming an *invisible* interface—one that we trust and about which we forget, and of which we remember only when something goes (badly) wrong, as in the case of COMPAS. The ‘trust and forget’ dynamic is problematic, as it erodes human control over digital technologies and over their impact on our societies. It is this dynamic that we avoid by putting the correct level of trust into digital technologie.

Design can play a crucial role in addressing this problem. Pop-up messages alerting users to algorithmic search engine results that have taken into account the user’s online profile or messages flagging that the outcome of an algorithm may not be objective are two good example of how design solutions can avoid ‘trust and forget dynamics’.

However, while resorting to better technological design may help, it is not the appropriate strategy in the medium- and long term. Two main reasons support this point. First, design solutions are *ah hoc*—they address specific problems following the specific implementation of a given technology in a given context. They do not provide an overall strategy to fine-tune trust dynamics in mature information societies. Second, the correct trust in digital technologies is defined accordingly to the way we design our societies (for example as open, pluralistic, tolerant, and just) and not according to the way in which we design digital technologies. A different approach is needed.

[1] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Mature information societies require a normative infrastructure to limit the spreading of ‘trust and forget dynamics’. This infrastructure should encompass norms enforcing transparency on the way digital technologies are deployed, for example when making decisions concerning human beings (Wachter et al. 2017); prescribing meaningful human oversight, for example when deploying autonomous technologies for security (Taddeo 2013) and national defence purposes; defining policies to ascribe liabilities of designers, providers, and users of digital technologies (Floridi 2016c). The alternative is to risk losing stewardship of the deployment of digital technologies and hence of the development of the societies that rely on them.

References

Floridi, L. (2002). On the intrinsic value of information objects and the infosphere. *Ethics and Information Technology, 4*(4), 287–304.

Floridi, L. (2011). *The philosophy of information*. Oxford: Oxford University Press.

Floridi, L. (2014). *The fourth revolution, how the infosphere is reshaping human reality*. Oxford: Oxford University Press.

Floridi, L. (2016a). Mature information societies—A matter of expectations. *Philosophy & Technology, 29*(1), 1–4. https://doi.org/10.1007/s13347-016-0214-6.

Floridi, L. (2016b). Tolerant paternalism: Pro-ethical design as a resolution of the dilemma of toleration. *Science and Engineering Ethics, 22*(6), 1669–1688. https://doi.org/10.1007/s11948-015-9733-2.

Floridi, L. (2016c). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112.

Floridi, Luciano, & Taddeo, Mariarosaria. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360.

Primiero, Giuseppe, & Taddeo, Mariarosaria. (2012). A modal type theory for formalizing trusted communications. *Journal of Applied Logic, 10*(1), 92–114. https://doi.org/10.1016/j.jal.2011.12.002.

Taddeo, M. (2010a). Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines, 20*(2), 243–257. https://doi.org/10.1007/s11023-010-9201-3.

Taddeo, M. (2010b). An Information-based Solution for the Puzzle of Testimony and Trust. *Social Epistemology, 24*(4), 285–299. https://doi.org/10.1080/02691728.2010.521863.

Taddeo, M. (2013). Cyber security and individual rights, striking the right balance. *Philosophy & Technology, 26*(4), 353–356. https://doi.org/10.1007/s13347-013-0140-9.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics, 2*(6), eaan6080. https://doi.org/10.1126/scirobotics.aan6080.