Foundations of Science (2023) 28:225–254
https://doi.org/10.1007/s10699-022-09836-2

Reviewing the Case of Online Interpersonal Trust

Mirko Tagliaferri[1]

Accepted: 18 December 2021 / Published online: 17 April 2022

Abstract
The aim of this paper is to better qualify the problem of online trust. The problem of online trust is that of evaluating whether online environments have the proper design to enable trust. This paper tries to better qualify this problem by showing that there is no unique answer, but only conditional considerations that depend on the conception of trust assumed and the features that are included in the environments themselves. In fact, the major issue concerning traditional debates surrounding online trust is that those debates focus on specific definitions of trust and specific online environments. Ordinarily, a definition of trust is assumed and then environmental conditions necessary for trust are evaluated with respect to such specific definition. However, this *modus operandi* fails to appreciate that trust is a rich concept, with a multitude of meanings and that there is still no strict consensus on which meaning shall be taken as the proper one. Moreover, the fact that online environments are constantly evolving and that new design features might be implemented in them is completely ignored. In this paper, the richness of the philosophical discussions about trust is brought into the analysis of online trust. I first provide a set of conditions that depend on the definition of trust that can be assumed and then discuss those conditions with respect to the design of online environments in order to determine whether they can enable (and under which circumstances) trust.

Keywords Doxastic theories of trust · Affective theories of trust · Online trust

1 Introduction

Trust[1] fosters cooperation (Gambetta, 1988) and it does so without requiring complex and expensive infrastructures (Williamson, 1993). Moreover, trust allows this cooperation to emerge in systems characterized by uncertainty Baier (1986) by decreasing the complexity of social environments (Luhmann, 1979) and, thus, allowing actions to take place even when data (or time) is not sufficient to perform a thorough analysis of the possible

[1] As will be emphasised in later sections, this paper is concerned with *interpersonal trust* only. Even though important, other forms of trust, e.g., institutional trust, won't be dealt with.

✉ Mirko Tagliaferri
mirko.tagliaferri@gmail.com

[1] Urbino University, Urbino, Italy

outcomes of said actions. Those facts seem to suggest that trust is an extremely important factor in environments where social interactions take place. Henceforth, the relation between trust and online environments has been object of study (Ess & Thorseth, 2011; Floridi & Taddeo, 2011; Grabner-Kräuter & Schratt-Bitter, 2013; Keymolen, 2016; Tagliaferri, 2019; Tagliaferri & Aldini, 2018a, 2018b). Online environments have become important parts of our daily life: we order food through apps, we book hotels through websites, we socialize with other people through social media and we learn new things through online courses. Basically everything that we used to do in the physical world until thirty years ago, can now be done online. Moreover, the increased possibility to interact online has become important also for technologies (e.g., smart homes sensors), which can exchange data and create huge networks of complementary services, given birth to what is now called the Internet of Things (IoT) Evans (2011).

Assuming that this shift from interactions in the physical world to interactions in online environments will gradually increase over time, it is useful to understand if trust (and which kinds of trust) can emerge in (which kind of) online environments. Specifically, what is needed is an answer to the following question concerning online trust[2]:

$$\text{Is online trust possible?} \qquad (Q_1)$$

In order to answer Q_1, it is necessary to understand whether online environments can satisfy the necessary conditions for trust and what is needed in order for them to do so. This will obviously depend on what trust is and how it is characterized. Ordinarily, the first step in finding an answer to Q_1 is to specify the definition of trust that is assumed. Then, a discussion usually follows about the features of specific online environments that might foster or inhibit the phenomenon of trust. Finally, such analyses produce a positive or negative answers based on the specific definition of trust assumed and the typology of environment which was analysed.

However, this *modus operandi* fails to appreciate that the concept of trust has always eluded a precise regimentation. Disciplines as diverse as sociology (Barber, 1983), economy Dasgupta (1988), political science Hardin (2002) and evolutionary biology Trivers (2002), computer science (Aldini et al., 2021; Aldini & Tagliaferri, 2020a, 2020b, 2020c; Tagliaferri & Aldini, 2022) dedicated some of their attention to trust, obviously prioritizing their specific needs and using their typical examination techniques. Not only, even inside the same disciplines, trust is often defined in different ways. A great example is the analysis of the nature of trust pursued in philosophy. Different authors provide profoundly different definitions of trust (Carter & Simion, 2021; McLeod et al., 2020), each one encompassing a multitude of examples that bring support to the specific definition provided.

Given the diversity of definitions of trust, finding a precise and definitive answer to Q_1 is almost impossible. What can be done, nonetheless, is to provide conditional answers based on specific analyses of trust and specific features required from online environments.[3]

[2] With the term *online trust* I refer to an occurrence of trust in an online environment. Later in the paper, definitions for the terms "trust" and "online environment" will be provided. For the moment, an intuitive grasp of what trust and online environments are is sufficient to understand the questions introduced.

[3] This is partially done when a specific definition is assumed and an answer is provided with respect to that specific definition. The difference between this common practice and the one followed in this paper is that in the latter case, multiple definitions of trust are taken into consideration at the same time, highlighting potential similarities and differences between those definitions and showing whether multiple forms of trust can emerge in an online environment with certain design characteristics.

Thus, instead of focusing on a specific definition of trust, a high-level analysis of the concept can be performed and different answers to Q_1 can be provided depending on how trust is conceptualized and which features of online environments are assumed.

The aim of this paper is exactly that of reviewing the philosophical literature surrounding trust in order to extrapolate different features of trust that can characterize the phenomenon in specific ways. Those features will then be scrutinized with respect to the characteristics of online environments and their design checking whether the specific forms of trust under analysis can emerge in such environments or whether specific design features are required to allow such emergence.

This means that the paper should be taken as a highly abstract guide in the exploration of various forms of online trust, providing insights on different conditions for trust and the feasibility of their fulfilment in online environments (and potential design choices that have to be made in order to enable such trust).

The paper is structured as follows. In section two, all the major definitions and assumptions that will form the bulk of the paper will be provided and explained. Most importantly, a definition of trust and of online environment will be provided. The definition of trust provided will include both mandatory and optional features that can qualify trust in different ways. In section three, those features of trust will be assessed and discussed trying to understand which design choices are necessary (if any) to enable trust in online environments. Finally, some conclusions will follow.

2 Definitions and Assumptions

In this section, the definitions that constitute the bulk of this paper will be provided. First of all, a definition of trust will be provided. This definition will include various conditional features that can help in characterizing trust as a specific concept. Moreover, a definition for the concept of *online environment* will be given. Finally, some concepts related to trust, i.e., *reputation* and *trustworthiness*, are analysed, showing how those concepts diverge from that of trust and why online trust might be relevant over and above them. Those sibling concepts will play a role also in subsequent analyses, since they are often employed as trust enablers in online environments.

2.1 Interpersonal Trust

The starting point for this paper will be *interpersonal trust* Potter (2020), Simon (2020). This means that no references will be made to theories that deal with trust which is not interpersonal, e.g., institutional trust (Hurley et al., 2013), group trust (Sapp et al., 2019), therapeutic trust (Hinchman, 2017) or trust conceptions that see trust as a property of relations (Primiero & Taddeo, 2012).[4]

Definition 1 (*Interpersonal Trust*) Interpersonal trust (from now on simply “trust”) is an **attitude** that an agent a_1 (the *trustor*) has towards another agent a_2 (the *trustee*) for a

[4] This does not mean that those characterizations of trust are not important for discussions on online trust. They simply fall outside the scope of this paper.

specific purpose ψ. In order for the attitude to qualify as trust, at least two elements are necessary:

1. The trustor must rely on the trustee in order to fulfil purpose ψ.
2. The circumstances in which trust is elicited must contain at least two elements of freedom. Specifically, the trustor must be free to choose whether to rely on the trustee or not; moreover, the trustee must be free to choose to betray the trustor by not contributing to the fulfilment of the purpose which is part of the trusting relationship.

Definition 1 is quite general. This is because such definition will only be employed to lay the foundations of all further conceptions of trust.[5] Starting from this definition, some important points can already be highlighted.

First and foremost, definition 1 states that trust is an attitude. This implies that, at the very least, the trustor must be an agent that can be an *attitude-bearer* and the trustee must be an agent towards which attitudes can be directed. Since the nature of the attitude depends on the different definitions of trust employed, no more can be said right now, since further specifications are necessary to better understand this requirement.

Second, the type of interpersonal trust that is analysed in this paper involves a three-part relationship. There will always be two agents (the trustor and the trustee) and a specific purpose that will determine whether the trusting attitude is present or not.[6]

Third, since the definition specifies the presence of two agents, some mandatory characteristics of those two agents must be made explicit. The major assumption that will be made on the nature of the agents is that the agents are able to *interact* in the environment in which they are placed. This means, among other things, that the agents must be able to *communicate* with each other and they must be able to *perform actions* that can *influence* the environment they are part of and the agents they interact with. The exact way the agents communicate, perform actions and influence their environment is not relevant, as long as they are able to do so. This means that both humans and artificial intelligent agents (AIs) might fill the role of trustor and trustee.[7] For the moment, a standard definition of what an AI is will be employed (Grodzinsky et al., 2010): an AI is taken to be a non-human entity that is autonomous, interacts with its environment and adapts itself as a function of its internal state and its interaction with the environment.

Finally, there are two necessary elements that characterize definition 1 which can be given more specific definitions.

Definition 2 (*Reliance*) An agent a_1 **relies** on another agent a_2 in order to fulfil a specific purpose if agent a_1 acts based on the supposition that agent a_2 will indeed contribute to the fulfilment of the purpose.

[5] This means that even though necessary, those conditions might not be viewed as sufficient for trust by many authors. However, those are the only elements that are assumed to be necessary independently from the theory of trust that is discussed.

[6] For a discussion of trust as a two-part relationship, see (Domenicucci & Holton, 2017)

[7] Given this possibility, when discussing whether trust is possible in an online environment, four different cases shall be taken into consideration: i) a human-human interaction; ii) an AI-AI interaction; iii) a human-AI interaction; iv) a AI-human interaction.

The concept of reliance is central to interpersonal trust. To understand its importance, take into consideration the effects of the betrayal of trust. If trust is distinct from reliance, then the following scenario would be possible. Imagine that you trust your friend to come and help you packing your belongings to move house, but, at the same time, you do not rely on him doing so. Given the absence of reliance on your friend, your actions would not be driven by the supposition that s/he will come and help you. Therefore, you would hire a private company to come and pack your belongings and then relocate them in your new house. Suppose that your friend does not show up. In such a case, it would be strange for you to feel betrayed by your friend; after all, even if s/he showed up, the company would have completed the job, without needing any help from your friend. Thus, the trust you placed on your friend was useless all along. On the contrary, if you relied on him, starting to pack your things and waiting for him to come to your old house and help you relocate all the packages and s/he didn't show up to help, then, it is likely that you would feel betrayed, since all your actions were performed on the supposition that s/he would come.

At the same time, it is possible to show that the two concepts do not collapse into the same one, since it is possible to rely on an agent without trusting them. Take the previous example as a reference: you might rely on the company to properly handle your belongings during the relocation, without necessarily trusting it. Your reliance, for instance, might be present because you know that the company is insured and if they do not complete the job properly, you would not pay them. At the same time, you might not have enough information about the company to fully trust them.

From this it follows that reliance is a necessary, although not sufficient, condition for trust.

Definition 3 (*Freedom*) An agent that is participating in a relationship is said to be **free** whenever s/he can autonomously choose to withdraw from the relationship (or not enter in it at all).

A trusting relationship envisages two distinct elements of freedom of the agents. The first element is that of the trustor, who must have the freedom to decide whether to rely[8] on the trustee or not. In case this first element of freedom is not present, then, instead of using the term "trust" to qualify the relationship, it would be better to talk about obligation. Usually, the trustor's freedom comes from the possibility of having alternative routes to fulfil his/her purpose, which would not require the intervention of the trustee. This alternative possibility does not need to be preferable, it only needs to be present. For example, you might trust your friend to help you to move your furniture, since you rely on him/her to get the job done. Obviously, you could have done it by yourself, without requiring his/her intervention. In this case, the job would have been more complicated and more difficult, but you indeed had an alternative possibility. The second element of freedom is that of the trustee. As with the trustor, the trustee must be free to decide whether to enter the trusting relationship or not and, moreover s/he must be able to withdraw from the trusting relationship whenever s/he wants. When the trustee has no choice but to enter the relationship, then it would be more appropriate to use the term "coercion" rather than "trust". Moreover, if the trustee cannot withdraw from the relationship once s/he entered it, then, again, it

[8] I am using the term reliance since in the previous paragraph it was assumed and argued that trust is indeed a form of reliance.

would be better to talk about obligation. For example, if the trustee signs a contract that forces him/her to collaborate with the trustor, then, this contractual obligation will make trust redundant in the relationship. Note that while the first element of freedom can directly affect only the trustor, the second element of freedom doesn't impact only the trustee, but might also affect the trustor. This is because the trustee is in a dominant position in a trusting relationship, since s/he might choose to exploit the relationship without contributing to it as much as s/he could (or even not contributing at all). This second element of freedom is what connects trust to risk and why the two are often discussed together. In fact, in order to have trust it must be the case that the trustor is exposed to a given level of risk that his/her expectations about the trustee's contribution to the relationship might be misguided (Corritore et al., 2003; Luhmann, 1979; Nickel & Vaesen, 2012).

The importance of those two elements of freedom for trust to be meaningful can also be seen by considering the fact that without them, there would be no need for trust in the first place. If the trustor had no choice but to rely on the trustee to fulfil a specific purpose, then such reliance would be a matter of necessity and even without trust being present, the trustor would simply have to rely on the trustee. On the other hand, if the trustee was not able to choose whether (and in which measure) to contribute to the fulfilment of the purpose of the trusting relationship, then, again, the trustor would simply need good forecasting abilities to correctly evaluate the (eventual) contribution of the trustee and act accordingly. Trust would be useless and/or redundant.

Now that the main elements that constitute trust have been discussed, some optional characteristics of trust that might capture specific definitions of the term will be introduced.

2.2 Optional Trust Characteristics

Over and above the necessary elements that characterize trust and that were introduced in the previous section, there are also more specific and discretionary elements that might characterize different conceptions of trust. Those elements exist because, as was mentioned earlier in the paper, trust is a multifaceted concept which is hard to nail down with a unique definition. Thus, in this section, some additional optional elements that might characterize trust will be discussed. Given the aim of the paper (i.e., that of enhancing the discussion surrounding online trust) different types of elements will be introduced and some references to theories assuming those elements as constituents of trust will be provided. Note, however, that it is almost impossible to give an exhaustive list of those elements; thus, focus will be placed on generic descriptions of the elements, rather than specific details about them. This should allow those principles to apply to most theories of trust that are described in the philosophical literature.

It is hoped that the introduction of those elements will enrich the discussion surrounding question Q_1, in order to do justice to the depth of the philosophical discussion around trust.

Two distinctive dimensions must be taken into consideration when trying to define trust in a more specific manner: i) the nature of the trusting attitude and ii) the contents of such attitude. The former specifies the type of attitude that trust is taken to be, while the latter indicates what such attitude is about. Concerning the first dimension, two different types of attitudes are usually linked to trust: doxastic attitudes and affective attitudes. Each type of attitude will describe a class of trust definitions that qualify trust as being an instance of that particular type of attitude. Concerning the second dimension, various contents of trusting attitudes have been presented and discussed in the literature about trust, each with peculiar features that might slightly change the nature of the phenomenon itself (i.e., they might be more apt at describing forms of trust which are different from interpersonal trust). In general, three distinct classes of contents can be identified: actions, motives and norms.

The two dimensions of trust interact with each other (i.e., one of the two possibilities for the nature of the trusting attitude is coupled with one or more of the classes of objects that specify the contents of the attitude), generating different trust theories that try to describe as precisely as possible the phenomenon of trust.

The two dimensions (the nature of trusting attitudes and the contents of such attitudes) will now be discussed in more details, trying to extract some elements that can be added to the ones described in the previous section. Note that the aim is not to justify one theory over the other or to decide which theory best explains examples of trust that can be encountered daily. The aim is to present different possibilities for the definition of trust and identify the main elements of such possibilities, in order to judge whether those components are compatible with the emergence of online trust. An important thing to keep in mind is that mixed approaches to trust are also possible. However, those won't be discussed. The reason is that understanding whether a mixed conception of trust is possible online would only require to take into consideration the relevant conditions that emerge from each separate element. For instance, if a theory claims that trust is both a doxastic and an affective attitude, understanding if trust is possible online would require a check of the conditions imposed by the doxastic components and the conditions imposed by the affective components. Thus, while interesting, discussing mixed approaches would not contribute in any way to the scope of this paper.

2.2.1 The Nature of Trusting Attitudes

When the nature of trusting attitudes is taken into consideration, two major possibilities are available. One possibility is that trust is a doxastic attitude. Starting from this intuition, it is possible to define doxastic theories of trust.

Doxastic theories of trust claim that trust implies (or is) a belief (those beliefs will be called *trusting beliefs*) Keren (2020). According to those theories the phrase "Agent a_1 *trusts* agent a_2 to do ϕ" (where ϕ indicates the possible actions that agent a_2 can perform to help fulfilling purpose ψ) implies (or is equivalent) to the phrase "Agent a_1 *believes* that agent a_2 will do ϕ".[9] Thus, doxastic theories of trust focus their attention on the beliefs (and their contents) that are tied to trust. In particular, the rationality and the epistemology of beliefs become central in the conceptualization of trust; the main point is that understanding when it is rational to hold a belief would define immediately the situations in which it is rational to trust another agent.

Doxastic theories of trust add to definition 1 the following requirement: the trustor must hold beliefs concerning the trustee (optional condition 3).

The second possibility is that trust is an affective attitude. Starting from this intuition, it is possible to define affective theories of trust.

Affective theories of trust differ from doxastic theories of trust due to the fact that they reject the idea that beliefs are necessary in order to have trust; they might be present, but are not essential. What is actually important for trust is the presence of an affective

[9] As the exemplary phrase might suggest, in this paper it is assumed that beliefs are propositional attitudes.

attitude Jones (1996). Those affective attitudes are often described in terms of emotions and/or mental states that are claimed to be different from beliefs. Therefore, affective theories of trust support the view that trust is present only when trusting relationships are grounded on emotional feelings directed towards a specific content (which might differ among the different affective theories of trust).

Affective theories of trust add to definition 1 the following requirement: the trustor must hold the appropriate affective attitude towards the trustee (optional condition 4).

2.2.2 The Contents of Trusting Attitudes

When the contents of trusting attitudes are taken into consideration, three major classes of objects have been analysed and presented as possibilities.

The first class of objects is that of actions. This first class is quite self-explanatory and it is typical of theories (*action-based theories of trust*) that claim that trust is directed towards the actions of others. In its simplest form, the content of the trusting attitude is just the evaluation of the probability that the trustee will indeed perform a specific action.

If actions are taken as constituents of trusting attitudes, then it is necessary to add the following requirement to definition 1: the environment in which trust has to emerge must allow the transmission of information that can be employed by the trustor to evaluate the potential actions of the trustee and their likelihood (optional condition 5). This requirement is needed because without it, it is not clear how the trustor is in a position to form the proper attitude towards the expected actions of the trustee. After all, if someone is unable to evaluate which actions might be performed by a given agent, then it is impossible to him/her to form sensible expectations that some course of action will take place.

The second class of objects is that of motives. This second class of objects is slightly more complicated compared to the first. In particular, theories of trust that assume that motives constitute the contents of the trusting attitudes are often built upon action-based theories of trust, i.e., trust must involve an attitude towards the actions of others, and adds a further element that requires that the attitude is also directed towards the motives that encourage someone to perform such actions. Thus, from the point-of-view of the trustor, not only the trustee must act in a certain way, s/he must also be motivated in doing so. This further requirement is assumed to be necessary to distinguish cases of exploitation from cases of genuine interest in the relationship.

If motives are taken as constituents of trusting attitudes, then it is necessary to add the following requirement to definition 1: the environment in which trust has to emerge must allow the transmission of information that can be employed by the trustor to evaluate the potential motives that move the actions of the trustee (optional condition 6). As with action, without this requirement, it would be impossible for a trustor to properly evaluate whether the trustee has the right motives to act in a fruitful manner.

The third and final class of objects is that of norms. This third class identifies theories of trust that claim that in order to trust someone, the trustor must have an attitude towards the existence of normative grounds that can press someone to act in a certain manner.[10]

[10] It is important to notice that if the trustor believes that the norms motivate the trustee to act in a certain way, then those further views can be seen as a subclass of motive-based doxastic theories of trust. However, the trustor might also believe that the trustee acts upon those norms unconsciously. In such case, the trustor would not attribute to the trustee an explicit motive to act and this would provide those normative-based doxastic theories an independent status.

As with motives-based theories of trust, also normative-based theories of trust are built starting from action-based theories of trust. In particular, the trustor must have a specific attitude towards the actions of the trustee and s/he must extend this attitude to the potential normative grounds that push the trustee to act in such a manner. Different normative-based theories of trust will differ on the specific nature of the norms, e.g., they might be social norms or moral norms, *etc*.

If norms are taken as constituents of trusting attitudes, then it is necessary to add the following requirement to definition 1: the environment in which trust has to emerge must have shared norms and values and those shared norms and values must be understood by the agents present in the environment who want to form a trusting relationship (optional condition 7). This requirement is needed because the benefits of having norms come from the fact that all members of a specific group accept those norms and act accordingly. For instance, take a game of basketball. If all players agree to abide to the rules of the sport, then it is possible to play. However, if every player follows his/her own rules, then it is highly unlikely that a basketball game is played, since there would be no way to judge whether someone broke a rule or not (everyone could make up arbitrary rules). This connection between the potential shared norm and value infrastructure and the evaluation of the behaviour of others is what is important in normative-based theories of trust. In order to be able to judge whether someone is fruitfully participating in the trusting relationship, it is necessary that there is a comparative framework against which judging the actions of that someone.

Mixing together the different natures of the trusting attitudes with the contents of such attitudes, it is possible to obtain different trust theories. As a reference: for doxastic action-based theories of trust see (Gambetta, 1988; Taddeo, 2010); for doxastic motives-based theories of trust see (Baier, 1986; Hardin, 2002); for doxastic normative-based theories of trust see (Cogley, 2012; Dasgupta, 1988; Kelp & Simion, 2020; Nickel, 2007); for affective action-based theories of trust see (Frost-Arnold, 2014; McGeer, 2008); for affective motives-based theories of trust see (Jones, 1996; McLeod, 2002); finally, for an affective normative-based theory of trust see (Lahno, 2017).

2.3 Online Environment

As with trust, the conception of an online environment is also quite vague. Many environments with quite different characteristics qualify as online environments. In order to avoid confusion, a standard model of for computing systems will be employed as a reference point. Starting from such model, a working definition of online environment will be provided.

Definition 4 (*OSI Model*) The Open Systems Interconnection model (OSI model) is a conceptual model that characterises and standardises the communication functions of a telecommunication or computing system without regard to its underlying internal structure and technology.

The OSI model describes Network Architectures (e.g., the Internet) and provides standards that can be used at different abstraction levels of such architectures. In particular, according to the OSI model, the communication between two systems can be split into seven different abstraction layers: the physical layer, the data link layer, the network layer, the transport layer, the session layer, the presentation layer and the application layer[11]. Each layer will determine which protocols should be used to exchange information at that specific level. For the purpose of this paper, only the application layer will be taken into consideration. At the application layer, the end users will interact through a software application, allowing them to communicate.

Definition 5 (*Online Environment*) An **online environment** is a virtual space in which two (or more) agents can interact through an interface that follows the protocols of the application layer of the OSI model.

Some specifications are needed in order to better understand the previous definition.

First, a virtual space is whichever space that offers an interacting interface that allow two agents to communicate. Usually, those spaces resemble physical spaces, even though the physical layer at which the communication occurs is different when compared to how the virtual space is perceived by the interacting agents. Examples of virtual spaces are chat rooms, blogs, forums and social media news feeds. As said, in order to count as an online environment, those virtual spaces must allow interactions between the agents that are present in such space. Those interactions are carried out through an interface, which is the specific device or program that enables the agents to communicate with each other. Examples of interfaces are graphical user interfaces (GUI) that allow communication between human beings and application programming interfaces (API) that allow communication between programs.

Such interfaces should be tied to applications that are constructed in accordance to the *application layer* of the Open Systems Interconnection model (OSI model), thus following specific protocols that are set forth by such layer. The interfaces of most network applications that allow communications between agents fall into this category. Examples of such network applications are web browsers and email systems. The protocols built for the application layer in the OSI model can be seen as normative constraints on how communications can take place.

This definition of *online environment* is quite broad and it is so in order to include different types of environments. Moreover, given definition 5, an online environment can allow interactions between agents of various kinds. Obviously, different interacting scenarios will call for different interfaces and different protocols being used.[12] Two human agents might communicate through the use of a news feed of a social media, which handles communication between the users using the HTTPS protocol. Also, a software (the client) could communicate with another software (the server) through a RESTful interface which employs, again, the HTTPS protocol.

[11] The choice of employing the OSI model as a base instead of the TCP/IP model is due to the fact that the OSI model is considered the *de iure* standard for communicating networks. Moreover, compared to the TCP/IP model, the OSI model has a finer grain in specifying the various layers of a network system, which allows to better select the abstraction level of the environment. This said, it is quite easy to move from the OSI model to the TCP/IP model; in particular, the application layer of the OSI model is related to the application layer of the TCP/IP model, which, however, also maps onto the presentation and session layer of the OSI model.

[12] It is important to appreciate that different protocols does not mean different models. The OSI model contains a multitude of protocols at each layer that can be adapted to specific scenarios.

As the examples show, the definition of online environment is broad enough to allow different kinds of communicating interactions between different kinds of agents, thus not limiting the scope of this paper to specific interacting scenarios.

In subsequent parts of this paper, further features might be added to this basic definition of online environment. In particular, it will be explored what kind of peculiar features are necessary in order to have specific conceptions of trust in those environments.

2.4 Trust Siblings

In this subsection, various concepts that are closely tied to trust will be discussed. This is done to clarify the importance of trust and to highlight the difference trust might bring when compared to those sibling concepts. The two concepts which are commonly associated, and often conflated, with trust in computer science are the concepts of *reputation* and of *trustworthiness*.

Definition 6 (*Reputation*) An agent's **reputation** is a public evaluation of such agent based on the opinions of a community. A reputation could be **informal** or **formal**. An informal reputation is a community-driven subjective evaluation based on, e.g., rumors, gossips, innuendo and indiscretions; a formal reputation is an community-driven objective evaluation based on opinions provided by the community and then manipulated through the use of appropriate algorithms.

Reputation basically indicates how a given agent is perceived by the community with which such agent interacts. Reputation could be good or bad and it is often the first element of evaluation new members of the community employ to produce initial assessments of the agents that are already part of it. An agent's reputation is normally determined both by the behaviour of the agent inside the community and the way the community itself perceives this behaviour.

Formal reputation models are often employed in computer science as substitutes of trust models (Jøsang, 2007; Pinyol & Sabater-Mir, 2013). The reason to employ reputation over trust is that reputation is easy to compute and manipulate according to specific goals. The only thing that is needed is a way to gather data (which will be provided by members of the community) and then merge this data together in order to obtain a unique value, which will provide a quantitative measure of an agent's reputation. However, while simple and practically very important, reputation models (and, concurrently, the concept of reputation itself) have important flaws that make them unable to properly substitute trust in online environments. The first flaw is that members of a community might have genuine different evaluation criteria, thus making the overall reputation of an agent unreliable. An Italian individual might be completely satisfied and happy if the bus he/she has to take is 5 minutes late and, thus he/she will judge the bus service positively. However, it would be problematic for a British individual to employ the reputation of such bus service as a guidance to his/her choices, since, if he/she had been in the same situation as the Italian individual, he/she would have probably judged the service as poor. The possibility of disparity between the reputation ratings based on cultural or personal traits, can become extremely problematic when those differences become substantial. Differently from reputation, trust is not subject to this first kind of issue. The reason is because trust is evaluated subjectively[13] and thus it does not need to accommodate different evaluations from different individuals. Nonetheless, it has to be pointed out that this comes at a price, i.e., the complexity of implementing trust models in online environments given the possibly high amount of information that each evaluating agent might need to gather. A second flaw in employing reputation models instead of trust models is related to the possibility of maliciously deceiving reputation models (Yu & Singh, 2003). This happens when groups of malicious agents collaborate to inflate or deflate reputation scores in order to obtain a personal advantage. This might happen for various reasons, from political (discredit someone who has different views) to economical (falsely inflate the reputation of a bad product in order to keep on selling it). Again, trust avoids this issue: if trust is based on personal evaluations, in order to maliciously alter the perception of the target of trust, it would be necessary to alter the personal evaluations of most (if not all) the agents that might interact with the target. Obviously, this might become extremely expensive or hard to achieve, compared to simply manipulating the reputation score of such target through the use of false ratings. Given those reasons, while reputation models maintain their usefulness (especially when it comes to simplicity of implementation), they are simply not enough to produce the benefits that can commonly be associated with trust.

The second concept which is tied, but different from, trust is trustworthiness. Providing a clear definition of what trustworthiness is is difficult, since there are as many definitions of trustworthiness as there are definitions of trust. The reason is that trustworthiness is often given a thin sense for which an agent is trustworthy if it can be trusted by another agent (thus moving the issue of providing a definition from trustworthiness to trust). Even though it is difficult to provide a thick sense of the term ’trustworthiness’, there is at least one feature which can be attributed to the concept, i.e., trustworthiness is a property of an agent (McLeod et al., 2020) rather than an attitude (as trust is). Another thing that could be said about trustworthiness is that what is relevant is its presence in the trustee, rather than in the trustor. It doesn’t matter for a trusting relationship whether the trustor is trustworthy, as long as the trustee is. This is because the recipient of the trusting relationship should be the trustworthy one, given that he/she is the one with the possibility of exploiting the risky situation in which trust emerged. This said, when it comes to online environments, concentrating on building trustworthiness might be as important as focusing on building trust. The reason is that the two concepts are normatively tied: we should aim at trusting trustworthy agents and distrust untrustworthy ones. However, the two are distinct concepts and they can easily come apart (Scheman, 2020), meaning that they should be treated separately, rather than together. A focus on trustworthiness would require a study on which features are needed in order to allow such property to emerge in agents present in online environments. However, possessing such property wouldn’t bring, by itself, the advantages that are commonly attributed to trust, e.g., facilitating cooperation. Without the possibility of correctly tracking trustworthiness and, thus, forming trusting relationships only with trustworthy agents, being able to build trustworthy agents in online environments would be sterile. Not only, if it is proved that online trust is not even possible, then, the task of building trustworthy agents in online environments would be completely useless. This explains why the goal of this paper is so important and primitive with respect to other analyses

[13] The exact way in which it is evaluated will depend on the specific theory of trust that is employed.

that focus on online trust; it sets the stage to justify other forms of reflections on the interaction between trust and online environments. This said, while it is recognized that having a proper analysis of online trustworthiness is important, this paper will just focus on online trust, eventually setting the stage for further analyses of concepts which are close to trust - e.g., trustworthiness.

3 Assessing the Definition of Trust

It is now time to assess whether online trust is possible. The goal of this analysis is twofold: first of all, the analysis should help to find an answer to question Q_1; moreover, it should pinpoint some requirements that must be imposed on the nature of online environments, i.e. which, if any, further features are required in order for trust to emerge in those environments.

All the requirements for trust will be discussed. In order to assess whether trust can emerge in online environments, it is first necessary to evaluate whether the two main conditions for trust can be met in an online environment[14], then discussions about the optional conditions will allow the reader to have an omni-comprehensive view of the issue of online trust.

Recall the two major conditions for trust:

1. The trustor must be in a position that allows him/her to rely on other agents.
2. The agents involved in the trusting relationship must be free to enter, withdraw or avoid the relationship at all times.

3.1 Reliance

As far as condition (1) is concerned, it must be evaluated whether the phenomenon of reliance can indeed occur in an online environment. In order to decide on the matter, an account of reliance and its normative grounds is needed. In this paper, I will assume a particular version of the *Mixed View* of Reliance as presented in Alonso (2016), Alonso (2014). According to such theory, reliance, at its core, is a cognitive attitude. Here "cognitive" is employed in contrast to "conative" Velleman (1992), where the former qualifies attitudes that describe the world as it is from the perspective of the attitude bearer (e.g., believing is a cognitive attitude since the phrase "agent a_1 believes that the house is yellow" means that, from agent a_1's perspective, the house is yellow), while the latter qualifies attitudes that describes the world as it should have been from the perspective of the attitude bearer (e.g., wishing is a conative attitude since the phrase "agent a_1 wishes that the house was yellow" means that, from agent a_1's perspective, the house should have been yellow). In this sense, reliance is a cognitive attitude since it describes the world as perceived by agent a_1, i.e., from agent a_1's perspective, agent a_2 will contribute to the fulfilment of the

[14] It must also be checked whether the assumptions made about the trustor and the trustee (that the trustor is an attitude bearer and that the two agents are able to interact in the environment) do indeed hold for both human agents and AI inside an online environment. However, the former assumption will be discussed later, since it is dependent on the specific attitude trust is, while the latter is trivially satisfied by the definition of online environment, in which it is explicitly stated that such environments must allow interactions between agents.

purpose for which s/he is relied on. Moreover, reliance guides thoughts and actions and is normatively grounded on evidence and pragmatic considerations.[15] Specifically, for agent a_1 to rely on agent a_2 in a given context it must hold that (i) a_1 has a relevant purpose to reach in the given context; (ii) a_1 has good reasons for holding the view that relying on a_2 is a means to fulfil the purpose (pragmatic considerations); (iii) a_1 does not have sufficient reasons to hold the view that s/he shouldn't rely on a_2 (evidence considerations); (iv) (ii) is partly motivated by (iii) (connection between pragmatic and evidence considerations).

Therefore, in order to fulfil condition (1), four subconditions must be met in an online environment:

1.a The trustor must be a cognitive attitudes bearer.
1.b The trustor must have a relevant purpose to fulfil.
1.c The trustor should be able to pragmatically evaluate that the trustee can help him/her to fulfil the purpose.
1.d There must not be information that indicates to the trustor that s/he should not rely on the trustee, i.e., there should be no negative evidence pointing at the fact that the trustee will not act appropriately in the trusting relationship.

Subcondition (1.a) places some restrictions on the type of agents that can take the trustor's role. Understanding those restrictions will clarify which kind of agents can partake in a trusting relationship online. Since both human-agents and AI systems are taken into consideration in this paper, it must be asked whether either (or both) of them can indeed be cognitive attitude bearers.

For human agents, the issue is trivial. Human beings are prime examples of creatures who have cognitive attitudes and there is no controversy involving such claim. For AI systems, the matter is not so simple. The main problem is that most AI systems are aimed at reproducing human cognitive phenomena, but they achieve their goal through functional rather than structural resemblance to those cognitive phenomena Lieto (2021). Take, as an example, the back-propagation method in a neural network. Simply put, the back-propagation method is a way of building algorithms that help neural networks in their learning process. During such process the back-propagation algorithms take the errors at the output level (with respect to the correct output that should have been given) and tries to understand which connections inside the neural network were problematic by propagating back the errors through all the different layers of the neural network. Then, once identified, those connections are modified, in order to improve the quality of the output, ideally matching the correct response that was expected. The back-propagation method is very fruitful as a learning technique for neural networks and it resembles closely the trial-and-error learning technique by human beings. It, thus, provides AI systems with a very useful method that is functionally equivalent (if not even strictly better) to human learning. However, it is also implausible that biological systems do use forms of back-propagation Crick (1989). Therefore, while functionally similar, AI systems based on back-propagation do not structurally represent well the cognitive phenomenon of trial-and-error learning of biological systems (among which we have human beings). This example could easily be extended to other cognitive phenomena. Take, as a further example, the cognitive attitude of believing. According to a standard view of beliefs Schwitzgebel (2019), the term "belief" refers to the

[15] I am not ascribing any anthropological meaning to the noun "thought".

attitude of taking something to be the case. Given this view of beliefs, it is pretty straightforward to identify features of AI systems that functionally represent such an attitude, e.g., simple memory retrieval mechanisms. However, those functionally equivalent systems, are not structurally equivalent to the way beliefs are formed and retrieved in human beings. *Mutando mutandis*, it is possible to apply similar reasonings to the cognitive attitude of relying. All those examples should highlight that in order to evaluate subcondition (a) for AI systems, it is important to decide whether functionalism is sufficient or structuralism is required for the representation on cognition in such systems. This is because if someone is a defendant of strong structuralism (i.e., the view that only structurally equivalent mechanisms allow the emergence of cognitive phenomena), then it follows that AI systems cannot be (at our current level of technology) the bearers of cognitive attitudes. This said, when studying trust in online environments what is relevant is whether online communities can benefit from the effects that trust can generate. Henceforth, even assuming that AI systems are only able to replicate cognitive attitudes through functionally equivalent mechanisms, this is likely enough to bring about the desirable effects of trust. For the issues related to this paper, it is therefore safe to assume a form of functionalism about cognitive phenomena. If it turns out that structuralism is indeed the correct view about cognitive phenomena, then it would mean that AI systems are only able to maintain surrogates of trusting relationships, without ever being able to generate real ones.

Subcondition (1.b) places some restrictions on the type of environment in which reliance can emerge. In particular, the agents inside such environment should have purposes to reach. This subcondition can be easily fulfilled in various online environments (as defined by definition 5) by both human agents and AI systems. In online environments agents seldom interact without specific purposes in mind. Obviously, those purposes could vary depending on the specific circumstances in which the relationship occurs (e.g., buying an item, finding love, gathering information, or sending a request for a service), but it is hard to imagine that such relationships are totally purpose-free. This means that even though subcondition (b) is a restriction on the kind of situations in which reliance (and consequently trust) can emerge, it is not a meaningful restriction.

Subcondition (1.c) is the first normative element that constitutes reliance. Such subcondition is pragmatic in nature, since the trustor is justified in relying on the trustee only if such reliance is instrumental to the trustor's purpose fulfilment. At this stage, the trustor only needs to take into consideration the ability of the trustee to be useful for the fulfilment of the purpose, without any other considerations coming into play.

Definition 7 (*Ability*) Ability is conceptualized as the capacity of an agent to perform actions that can impact the fulfilment of a specific purpose.

The trustor must be in a position to evaluate whether the trustee possesses the relevant abilities that are required in order to fulfil the purpose of the trusting relationship. Such judgement helps in creating realistic expectations about what the trustee can and can't do. Indeed, without having the possibility to evaluate whether those abilities are present, it would be troublesome for the trustor to form any sort of real expectation, since it would not be possible to determine whether the trustee is in a position to actively participate in the fulfilment of the purpose. This judgement plays a crucial role for the emergence of reliance - especially when it is related to interpersonal trust. For example, I would never rely on my mother to perform an open-hearth operation since the evidence I possess about her abilities is sufficient to determine that relying on her is not instrumental to the purpose of successfully carrying out the operation.

In order to satisfy this subcondition (1.c) some general qualifications about the online environments in which reliance might emerge are needed. Specifically, the satisfaction of this condition depends on the possibility for the trustor to judge if the trustee possesses some specific abilities tied to the purpose s/he wants to fulfil. It has been shown (Papadopoulou, 2007; Papadopoulou & Kanellis, 2019), through experiments, that it is possible, in an online environment, to create substitutes to the cues that are normally employed in the physical world to judge the presence or absence of abilities on the part of the trustee. In particular, having a virtual reality interface that can substitute physical world cues (Papadopoulou, 2007), having a reputation systems that can provide enough information to the trustor to form initial assessments Papadopoulou and Kanellis (2019) and allowing high levels of communication between the parties (trustor and trustee) Ryan (2012) all contribute to increase the salience of the trustee, thus putting the trustor in the right circumstances to form expectations about the trustee's abilities and/or competences. Then, those justified expectations are employed by the trustor to influence the emergence of reliance Vries (2006). This is in line with social experiments which show that increasing amount of communication can foster trust between agents through a better understanding of the characteristics and intentions of others (Wichman, 1970). For human agents, the virtual reality features are especially important, since it has been shown that persons employ visual cues as their prime source of information in an online environment (Wang, 2005). Moving to AI agents, the kind of cues that are looked at might be different from the ones employed by human beings, i.e., visual cues might not be so relevant when judging the abilities of the trustee. In scenarios where the trustor is an AI agent, it might be necessary to have different forms of evidence in order to allow the agent to assess the presence or absence of abilities in the trustee. One suggestion is the use of standardized certifications - e.g., ISO 25010 certificates for software quality assurance - which indicate whether or not a given agent has an ability.[16] Those certificates would be exchanged employing the communication channels available and, coupling those information exchanges with the potential of making queries and the presence of reputation scores about the trustees can allow the AI agent to form expectations about the abilities of such trustees. Interestingly, the requirements that reliance places on online environments that can foster it are similar to the ones that are also present in the physical world. Physical cues, reputation and certificates are often employed in the physical world to prove the presence of some kind of ability (think about university degrees or language certifications) and are therefore employed by agents to evaluate whether someone is capable or not to perform a given task.

Subconditions (1.d) is the second normative requirement about reliance and the last overall requirement. Such requirement can be perceived as an evidential test on reliance. Such test is that the trustor must not have sufficient evidence to form the expectation that trustee will not act appropriately towards the fulfilment of the purpose. In general, the same information channels that have been discussed in the previous paragraph on condition (1.c) can be employed to gather the relevant information. However, in order to avoid the possibility that the trustees provide false information and can manipulate their public

[16] Human beings could also employ such certificates as indicators of the presence of certain qualities and/or abilities. However, instead of providing such certificates directly to the trustor, it seems more reasonable that those certificates are represented by graphical elements that appear in the interface that the agent is using to interact, thus falling into the first category of required possibilities.

perception, some further remarks are required (Kamvar et al., 2003). Two design features are especially important: i) the environment should not assign any profit to newcomers. That is, reputation should be obtained by consistent good behaviour through several transactions, and it should not be advantageous for malicious peers with poor reputations to continuously change their online profile to obtain newcomers status; ii) the environment should be robust to malicious collectives of peers who know one another and attempt to collectively subvert the system. Taking those two design features into account when building the infrastructure that will be employed by users to interact is extremely important (it is especially important in the design of reputation models, that, as was said, are prime sources of information about the trustees' abilities). Without such features, it would be easy to produce false evidence about the potential abilities of the trustees and to hide true evidence about their inabilities. This, in turn, would make it extremely difficult for the trustor to make appropriate evaluations and reliance would hardly emerge in the environment.

Before moving on to the other conditions that must be fulfilled in order to have online trust, a further point that has been widely discussed in the philosophical literature (Nissenbaum, 2001; Turilli et al., 2010) about online trust deserves some attention. Such point is connected to the identity of the agents in an online environment. One of the arguments against the possibility of online trust is that it is not possible to ascribe identities to agents in online environments. If this was true, it would obviously make it extremely hard for trustors to gather information about the trustees and thus establish whether to rely on them or not. However, as claimed in Turilli et al. (2010), having the possibility of being physically anonymous in an online environment, does not imply that it is not possible to ascribe a diachronic identity to agents. There are multiple ways to establish someone's online identity: access control, passwords and IP identification are just some of the techniques that can be employed in an online environment to make sure that we are always dealing with the same agent. Even Nissenbaum (2001), one of the major proponents of the impossibility of having a diachronic identity online, discusses some of those security mechanisms and admits that they might help in making online identities transparent. However, she then argues that trust and security are incompatible, since security defeats the purpose of trust itself and thus renders it useless. Her argument is compelling: if the aim is to implement security mechanisms in every phase of an online relationship, then trust is indeed useless. However, this is seldom the case; often, security is implemented only in specific phases of online relationships and, thus, is only an accessory element and not a constitutive one. In the case of trust (and reliance alike), security mechanisms can be implemented only as a vehicle to establish someone's online identity, without directly affecting other elements of the relationship that is established between the agents. This means that security is not employed as a substitute for trusting relationships, but as a facilitator. It might still be argued that even though security is indeed necessary and useful, it might not be sufficient. The fact that digital identities are not tied to physical ones[17] allows agents to frequently change their online profiles, which might again generate a problem of anonymity. The solution to this problem is tied to the discussion carried out in the previous paragraph about condition (1.d). In that case, some design features were presented that could overcome the problem of reputation avoidance (the endeavour of creating new profiles to avoid a bad reputation score). Those features can easily be employed to solve the problem of the proliferation of profiles, thus limiting the possibility of agents of creating multiple digital identities they can use to

[17] Note that it might actually be possible to tie online identities to physical identities through, e.g., biometric security mechanisms.

interact. Implementing both the security mechanisms and the design features in the online environments can have beneficial effects for the creation of diachronic identities online, without limiting the usefulness and scope of trust.[18]

Summing up, there seems to be no obstacles to the possibility of having reliance attitudes in online environments. This said, there are some requirements that should be taken into consideration. First of all, if we wish to extend the possibility of online trust also to AI agents, then a form of functionalism about cognitive attitudes must be assumed. Moreover, online environments must contain features that can increase the amount of information that is exchanged between agents. Often, reputation models are employed to perform this task (but other alternatives are also available), by allowing agents to evaluate the behaviour of others and then using those evaluations to create a publicly perceived score for specific agents. Finally, it has been argued that some security mechanisms and design features are required in order to allow the emergence of diachronic identities for the agents that interact in online environments.

3.2 Freedom

As far as condition (2) is concerned, it must be evaluated whether online environments are compatible with the two levels of freedom required by trust. Not many authors discuss trust in terms of freedom of the agents and instead opt to relate trust to risk and uncertainty. However, this is misguided. After all, those authors agree that risk is relevant for trust in virtue of the fact that the trustee might betray the trusting relationship. This means that it is not risk *per se* that characterizes trust, but the freedom that the trustee has to maintain, interrupt or exploit the relationship. It's this freedom that makes the environment risky for the trustor and thus requires trust on his/her part to partake in the relationship with the trustee. Same goes for uncertainty. It is often argued that uncertainty is fundamental for trust, since trust would be useless in an environment without uncertainty. Again, as with risk, the issue is that uncertainty is taken as fundamental *per se*, ignoring the fact that it is what causes the uncertainty that is relevant. Take as an example a situation in which you buy a ticket from a flight company and you expect such company to get you to your target destination. In this case, there is no trust involved in the relationship since the company is obliged to fulfil the contract you stipulated when you purchased the plane ticket. Obviously, there is still some level of uncertainty involved in the relationship (e.g., the company might be forced to cancel its flights due to extreme weather conditions), however, this uncertainty plays no role in the emergence of trust. The example should show that it is not uncertainty *per se* that is related to trust, but the fact that such uncertainty is the product of the freedom that the trustee has to act differently from what was expected. Thus, instead to evaluating whether online environments are risky and uncertain environments (which they are), it is more accurate to discuss whether those environments allow the agents to be free to enter, participate and/or withdraw from a trusting relationship. One last remark that is needed before analysing whether online environments allow for those two levels of freedom: when the concept of freedom is discussed in this paper, the issue is not metaphysical.

[18] It is important to appreciate that even though most authors are worried with the diachronic identity of human agents, this issues can be applied, *mutando mutandis* to AI agents. However, also in the case of AI agents, the same solutions could be employed to overcome the problem.

The important fact is not whether there is free will or whether artificial agents following algorithms are truly free. What is important is the perception of this freedom being present, i.e., whether the agents involved consider it possible to change their course of action. For example, if a human agent is interacting with an artificial one that is simply following its algorithm, but the human agent does not know which algorithm the artificial agent is following, then, from the perspective of the current discussion, the artificial agent can indeed decide to withdraw from the trusting relationship and shall be considered free. This is so even if the artificial agent withdraws from the relationship in virtue of a previously established algorithm telling it to do so. In order to not count as free, the artificial agent would have to be forced to fulfil its commitment in the relationship in pain of getting punished in case it does not do so.

As was briefly mentioned in Sect. 2.1, in order to have the two levels of freedom required by trust, two subconditions must be fulfilled:

2.a The trustor must be in a position of having various choices concerning the agents to rely on in order to fulfil his/her purpose.
2.b The trustee does not have to have strict obligations to fulfil his/her part of the trusting relationship.

The two subconditions (2.a) and (2.b) place some restrictions on the type of environments in which trust can emerge. Note that in both cases, the nature of the trustor and trustee are not relevant.

As far as subcondition (2.a) is concerned, online environments might be actually better suited for trust compared to environments in the physical world. In particular, online communities are often larger than real-world communities, since there are no geographical and/or spatial barriers that might limit the dimension of those communities. It is therefore reasonable to assume that the trustors would have many different agents to choose from when trying to find partners that can help them to fulfil their purpose. Obviously, all previous considerations about the amount of information needed by the trustor and the conditions that must be imposed on newcomers to the communities would still need to apply to all those potential trustees, thus, the increased number of potential trustees might not produce an increase in the number of actual trustees. However, it is fairly safe to assume that, in online environments, trustors have a large degree of freedom in choosing the proper trustees.

As far as subcondition (2.b) is concerned, the first thing to understand is the difference between a strict obligation and an obligation *simpliciter*. In this paper, the term “strict obligation” will indicate a course of action to which an agent is forcefully bonded. This forceful bond can be due to legal enforcement or brute force. Simply put, someone subject to a strict obligation has no other choice but to follow the prescribed course of action in pain of severe repercussions in case s/he doesn’t. On the other side, an obligation *simpliciter* refers to a course of action to which an agent is committed due to morality or ethics. In this case, the agent ought to follow the course of action, but not in a forceful manner, which means that s/he might refrain at any moment from following such course of action. The most common form of strict obligation is contractual obligation, where the only possibility for the agent to not follow the course of action prescribed is to breach the contract and getting exposed to the risk of tangible penalties. Obviously, in order to count as a real case of strict obligation, the risk of being punished must be real. If, for example, an agent breaches the contract, but lives in a country in which it is not possible to prosecute him/her, then there is no strict obligation. The lack of strict obligations is an extremely important aspect of trust. Following the intuitions presented in Williamson (1993), trust can be seen as a facilitator for interactions and cooperation also because it eliminates the necessity of constructing a whole infrastructure that enforces certain courses of actions of the agent involved. Building forcing infrastructures (e.g., a legal system) is costly and might inhibit lower-level interactions between agents that do not have the economical capacity to cover the cost of being part of such infrastructure (e.g., they are not in a position to pay a lawyer). In those cases, trust substitutes the whole legal infrastructure, increasing the risk of betrayal or misconduct by the agents, but also eliminating all sorts of costs of partaking in the interaction.

The question then is: do online environments enforce some specific behaviours on the agents or are they free? The answer to this question can be easily given if malevolent episodes that commonly occur in online environments are taken into consideration. Take as an example, the recent scam about Magic the Gathering cards on the Amazon platform.[19] Recently, Wizard of the Coast (the company producing Magic the Gathering cards) started selling boxes of card on Amazon. Unfortunately, many users reported that the boxes of cards they received were manipulated so that the packs of cards were opened, rare cards were removed and then the packets were resealed. The likely explanation was that fraudulent users bought the boxes, tampered with them and then resealed them and sent them back to Amazon thanks to their return policies. This meant that subsequent users that were sent those very boxes would fall victim of the scam. Even though many of the scammed individuals could ultimately claim their money back and return the deficient product, the scammers were not punished for their behaviour. This example shows that it is fairly easy to exploit features of online environments to behave in ways that are commonly thought to be against the law. The major reason can be attributed to the fact that online environments connect agents from different parts of the worlds (either persons that live in different countries or AI agents that are based on servers that are located in different countries), which might have different legislatures and thus make it hard to apply specific laws. This lack of a world-wide international law infrastructure allows agents to behave in ways which would not be possible in face-to-face interactions. This fact, while often tied to negative episodes, is also important for subcondition (2.b). In most circumstances, the trustees are free to behave as they please, since the chances of concrete retaliation are often missing in online environments. Note that this freedom also applies to AI agents that strictly follow algorithms without being able to alter their code (i.e., agents that do not have a concrete possibility of behaving differently from how they do), whenever the trustor is not in a position to know the details of such algorithm: this is so because from the trustor's perspective, there is no prescribed behaviour of such AI agent.

Summing up, online environments seem to be suited to allow the two levels of freedom required by the basic definition of trust given in Sect. 2.1. Note that this freedom is most likely employed to exploit the trusting relationship rather than enhancing it. This is exactly the reason why trust is so important: it allows fruitful cooperation even when the agents might opportunistically opt out of the relationship at any moment.

Now that the two major elements of trust have been discussed, the optional features of trust will be analysed leading to the final answer about online trust.

[19] https://mtgrocks.com/mtg-player-reports-issue-with-buying-booster-boxes-from-amazon/ .

3.3 Doxastic and Affective Theories of Trust

In this section, the two alternatives for the nature of trusting attitudes will be discussed. In particular, it will be discussed whether human agents and AI agents can be bearers of the attitude that characterizes trust.

3.3.1 Beliefs

The first attitude that will be discussed is that of *believing*. As with reliance, believing is a cognitive attitude, i.e., it is an attitude that represents the world as it is from the perspective of the attitude bearer and not as it should be from his/her perspective. Again, as was the case with reliance, asking whether human being can be believers is an easy question to answer. In fact, it is uncontroversial to claim that the condition is satisfied when the trustor is a human being. Acting in an online environment or in the physical world does not change the fact that human being can have (and do have) beliefs. Therefore, the only possible issue with this condition is to understand whether AI agents can have such beliefs. In this case, what is relevant is the kind of AI agent that is built. What is needed in order to have an AI agent that can have beliefs is that such AI agent is provided with what is commonly called *a knowledge base*.[20] A knowledge base is simply a set of sentences that the AI agent employs to deduce new pieces of information about the world or to decide on a course of action. It is important to understand that those sentences are assumed by the IA to be true facts about the world.[21] However, they might indeed be false. Given this possibility of falsehood, it might be better to label those knowledge-bases as belief-bases, indicating that those kind of AI agents can have beliefs in the form of propositional attitudes Perlis (2000).

3.3.2 Affective Attitudes

The second type of attitudes that will be discussed are affective attitudes. As with the belief condition, the fulfilment of the affective attitudes condition seems to be unproblematic for human agents. Whenever a trustor is a human agent, the very nature of such agent allows the formation of affective attitudes and, thus, the emergence of behaviours based on such attitudes. What can greatly impact the emergence of trust defined through the use of affective attitudes are trustors and trustees that are artificial intelligent agents. This is because only very specific kinds of AI agents are able to simulate affective attitudes.[22] Not only there is a problem for AI agents to understand emotions (a problem of empathy), but there is also a huge problem in understanding how to code an AI agent to allow it to simulate emotions Calvo et al. (2014). Concerning the first problem, the issue is building AI agents

[20] The reader shall not be fooled by the terminology. In computer science, the distinction between belief and knowledge is often neglected. In fact, often a knowledge base is only a set of propositions that the artificial intelligence agent believes to be true based on its interactions with the world.

[21] The fact that those sentences are assumed to be true, does not mean that AI agents might not "change their minds". In fact, AI learning agents can constantly change the information that they possess in order to better represent all the new evidence they came across. Moreover, also static (i.e., non-learning) systems might be allowed to change their knowledge base through various techniques, e.g., (Alchourron et al., 1985).

[22] I talk here about simulation of those attitudes since it is still controversial if AI agents will ever be able to concretely *feel* those affective attitudes. I, however, will assume a slightly uncommon position for which simulating the emotion might be sufficient to foster trust.

that can detect emotions and act upon them. While it is possible to obtain algorithms that allow AI agents to endure this task[23], the number of AI agents that do include those components is limited and most AI agents that can be encountered in online environments do not have those features implemented. This means that if trust is taken to be an affective attitude, there might be serious limits to the emergence of trust between artificial intelligent agents in an online environment (and for that matter between an AI agent as a trustor and a human being as a trustee). Concerning the second problem, i.e., AI agents simulating emotions, great advances have been made.[24] This seem to suggest that it might be feasible to have AI agents that can employ affective concept of trust in their decision-making processes. However, as with the first issue of empathy, currently, the amount of AI agents that do indeed implement features that allow them to simulate affective attitudes is still way too limited to be able to claim that affective attitudes are present among artificial intelligent agents in online environments. Thus, the affective attitude condition seems to suggest that, whenever an affective definition of trust is employed (rather than a doxastic one), online trust might be limited to interpersonal trust among human agents.

Now that the nature of the trusting attitude has been discussed and analysed, the next step is to analyse the contents of those attitudes, in order to understand whether online environments make it possible to collect information about such content and therefore form the relevant kind of trusting attitude.

3.4 The Contents of the Trusting Attitudes

As was presented in Sect. 2.2.2, there are three potential contents of trusting attitudes: actions, motives or norms. Each will be discussed separately, trying to establish whether online environments are suited to have such contents and to transfer information about them among the agents that interact in the environment.

3.4.1 Actions

First note that the possibility of having actions being performed in the environment was already taken as an assumption in the discussion of definition 1. What must be established, therefore, is whether online environments allow the transmission of information that can be employed by the trustor to evaluate what are the potential actions of the trustee and their likelihood.

Central to the possibility of evaluating an agent's behaviour is the concept of integrity:

Definition 8 (*Integrity*) Integrity implies that the trustee will adhere to a set of rules and principles. The trustee will be reliable in following those rules and principles that dictate his/her actions. This means that s/he will tend to act in certain ways when specific triggers and suitable conditions are present.

Integrity is important for the evaluation of the potential actions of others because it provides a frame in which to judge the probability that agent a_2 will perform a specific action rather than another. If agent a_2 does not show signs of integrity, then it is likely that his/her

[23] See, as a reference, section 2 of Calvo et al. (2014).

[24] See section 5 of Calvo et al. (2014).

behaviours are completely unpredictable, since under the same conditions and given the same triggers, s/he might act differently at different points in time.

The discussion about whether it is possible to evaluate an agent's integrity in an online environment will be postponed to the next subsection, since it makes sense to merge such discussion with that on benevolence. The reason is that the same information providers that can be used online to evaluate benevolence, can also be used to evaluate integrity and thus it makes sense to discuss the two notions together.

3.4.2 Motives

As was mentioned in Sect. 2.2.2, if a motive-based theory of trust is assumed, what is needed in order to form a trusting relationship is a set of cues that can allow the trustor to get a clear comprehension of the motives that guide the behaviours of the trustee. This set of cues are tied to the concept of benevolence.

Definition 9 (*Benevolence*) Agent a_2 display **benevolence** towards another agent a_1 if agent a_2 acts with a genuine interest in agent a_1's interests or welfare, subordinating immediate short-terms personal gains for long-term reciprocal gains.

Benevolence is central to motives because it forms the base of all relevant positive motives that could push the trustee to act in a favourable manner towards the trustor. As was said in the subsection about actions, the concept of benevolence is tied to that of integrity.

Understanding whether benevolence and integrity are present in an online environment is a difficult task, mostly due the lack of physical cues that are commonly employed by human beings to intuitively assess them in the physical world Jarvenpaa and Leidner (1999) and the consequent difficulty in identifying parameters that could substitute them online (which creates the further difficulty of understanding which parameters must be provided to AI agents to assess those features).

The results that will now be presented shall be seen as a reference and starting point to understand what might be done to overcome the problems of identifying benevolence and integrity online. Those results are not expected to be conclusive nor they provide the only potential solution to the problem. However, they manage to show that it is indeed possible to have indicators for benevolence and integrity online and, thus, open up the possibility of having online trust.

The most impactful elements for the perception of benevolence and integrity (at least for human beings) are virtual reality surrogates of the bodily presence of agents. As shown in Papadopoulou (2007) virtual reality can become a beneficial factor in the development of trust in online environments (this would impact all relationships where the trustor is a human being, independently from the nature of the trustee). The main explanation for such a phenomenon is that virtual reality substitutes concrete physical cues with digitally mediated alternatives. Moreover, the usefulness of virtual reality for considerations concerning benevolence and integrity is also given by the fact that virtual reality enables functionalities that are commonly associated with the presence of the two conditions just mentioned. Specifically, in Papadopoulou (2007) it is shown that in a virtual reality setting, it is possible to make promises, enabling them and then keeping them all the way. One example of such possibility is given when the system is designed to assign to each agent entering the virtual reality a personal avatar, which can digitally interact in the environment, make recommendations, provide specific information and, finally, act on the environment itself. All those features substitute those experiences that an ordinary agent would have in the physical world and thus foster trust as if the trustor was in a concrete setting and not in a virtual one. This study shows how important abundance of information is for the emergence of trust. Finally, the study manages to show that the type of the medium (either the physical world or a virtual one) plays no role in the relevance of the information provided, since virtual reality is thought of as a relevant alternative to the physical world (especially in economy). Further deepening the study (and expanding the results in order to take into consideration AI agents as trustors), in Papadopoulou and Kanellis (2019) the authors explore the effects of various interaction stages on trusting beliefs[25], where trusting beliefs are defined as the beliefs (of trustors) about the ability, benevolence and integrity of the trustees. The interaction stages they examine in the study are A) before, B) during and C) after the interaction with an online vendor has happened. Since in this paper the main issue is about the emergence of trust, only results from (A) and (B) will be discussed.

Concerning (A), the authors showed that reputation has a significant impact on trust beliefs concerning the ability and integrity of the trustee, but it plays no significant role in the formation of trusting beliefs about the trustee's benevolence. This seems reasonable, since reputation scores can help in assessing the skills of a given party and they allow to evaluate the general set of principles that the trustee follows. However, reputation scores can hardly provide any information about the general attitudes of the trustee, since they provide evaluations about objective and measurable features. Thus, benevolence can't be influenced by reputation and some alternative feature must be looked for. Concerning (B), the authors showed that different acts related to promising can affect trusting beliefs. The three acts related to promising they analysed are: i) making a promise; ii) enabling a promise; iii) keeping a promise. They show that all three acts positively affect all trusting beliefs, with a specific emphasis on the relation between making a promise and benevolence beliefs, enabling a promise and ability beliefs, and keeping a promise and integrity beliefs. Those results, coupled with the results obtained in Papadopoulou (2007) - showing that (i), (ii) and (iii) can be obtained in a virtual reality setting as well as in the physical world -, show that some of the actions that are commonly related to the emergence of trust in face-to-face interactions can also be reproduced in online environments and that they have similar effects in those environments as they have in the ordinary world.

A final study that showed what could impact the benevolence and integrity features is Brengman and Karimov (2012).[26] In it, the authors explore the impact of the implementation of two distinct web communities in the online environment. In particular, they explore the effect of introducing social network sites (SNS) such as Facebook and corporate blogs. SNS were chosen mainly for the possibility they provide to share information between agents, thus allowing precise community-driven reputation scores to emerge; corporate blogs were chosen for the possibility they allow to communicate information that is considered relevant by the trustee. The interesting result is that the introduction of SNS in online environment has no distinctive effect on the emergence of trust (even though it might have effects in maintaining it), while the introduction of a corporate blog does have an impact on the formation of beliefs about the benevolence and integrity of the trustee.

[25] These authors talk about trusting beliefs specifically, but the term "beliefs" can be substituted with the term "attitudes".

[26] Note that, even though the study is concentrated on the e-commerce side of interactions, the results seem to apply equality well to other forms of interactions.

Finally, what is shown is that if both SNS and corporate blogs are implemented, the effects obtained by the corporate blog alone are increased.

Although interesting, those studies seem to provide insights only on the issue of benevolence and integrity with respect to human agents, while it says very little about the possibility of AI agents to obtain such information (with the notable exception of Papadopoulou and Kanellis (2019)). This is due to the fact that AI agents might not perceive the world as humans do and, moreover, there is a strict requirement of specifications for when to determine if benevolence and integrity are present. Given the peculiar nature of AI agents, it seems reasonable to hold the view that the information for such artificial agents must come from reputation models, which can provide enough insights into the past behaviour of the agents in the environment to determine whether such agents have a propensity to put shared gains in front of personal gains (benevolence) and tend to maintain deals and promised services (integrity).

Summing up, in order to gain information about the abilities and motives that characterize trustees, either community-based judgements must be present (in the form of reputation models) or surrogates of the physical world cues that are normally employed by human being to judge benevolence and integrity are needed. This means that simple interacting interfaces alone might not be sufficient to allow the emerge of online trust if trust is defined according to either the ability-based or the motive-based theories of trust.

3.4.3 Norms

If norms are taken to be the contents of trusting attitudes, then, as was argued in Sect. 2.2.2, the presence of a shared value and norm infrastructure in the online environment is mandatory. This puts immense pressure on normative-based theories of trust, since it is often assumed that there is an absence of this kind of infrastructure in online environments.[27] The main argument that is brought forward for the absence of such infrastructure is the heterogeneity of the community that populates online environments. Geographical distance between users implies a difference in legal and institutional backgrounds, and a distinction in their moral and ethical values. Thus, it seems like all theories of trust that require those normative conditions are condemned when it comes to online environments. Although this might be true if online environments are taken as a whole, the same might not be true for specific instances of such environments. Smaller scale communities in online environments often establish sets of rules of conduct that indicate what is acceptable and what is condemnable. Think about a MMORPG (Massive Multiplayer Online Role-Playing Game): in the initial phases of interaction in the community of players, individuals are told what is considered appropriate and what is not. In this way, the newcomer gradually learns what are the norms that govern such community and might decide to abide to them (become an active member of the community) or breach them, often resulting in a ban from the game (i.e., the newcomer is prohibited from playing the game). This said, when it comes to human being interacting online, sometimes the added normative infrastructure is not even needed. Even though it is commonly thought that Internet services and online environment can connect the whole world (and it is indeed true that they can), what is neglected is that such services are also employed to maintain contacts and interact with agents which are

[27] There are also concerns about the fact that the presence of such infrastructure might inhibit trust. See Turilli et al. (2010) for a position defending such claim.

already part of our physical world communities. Social media groups and forums are thematic environments where the participating agents are already part of a community with its rules and norms. The fact that those agents interact online rather than in the physical world, does not inhibit their adherence to those rules and norms. Thus, on a smaller scale, online environments could still show elements of shared cultural and institutional backgrounds, satisfying the condition required by normative-based theories of trust.

The issue is more complicated when AI agents are taken into consideration. As with AI agent being able to have affective attitudes, the task of building AI agents that abide to moral, ethical and social norms is a complex one. There are two issues in particular: the first issue is that it is hard to pinpoint proper objects that can be used as a reference to build those AI systems. It is almost impossible to identify a unique set of (moral, ethical or social) rules that can guide actions in every possible scenario. Shall we use the norms set forth by Romans? Are the moral rules presented in the Bible the best ones? The issue is that human beings are not even close to understand whether those questions can be answered, without even dwelling into potential answers themselves.[28] Moreover, even looking at rules that appear to be uncontroversial, e.g., it is morally despicable to kill someone, there are particular circumstances in which they are not easily applicable and thus cause problems, e.g., the trolley problem in moral philosophy.[29] This first issue shows that even admitting the existence of moral and ethical AI agents, it might still be the case that a shared normative infrastructure is missing, since those AI agents might be subject to different moral and ethical norms. However, this first issue is easily solvable through considerations similar to the ones made for human agents. Even though it is probable that it will not be the case that all AI agents will share common moral and ethical norms, this does not mean that they cannot share them in smaller communities or webs of services Greene et al. (2016). This would, however, require AI agents that are able to learn the specific norms that are present inside a community and autonomously decide whether to abide to those norms or not.

Assuming that the first issue is solved, a second important issue is that of actually implementing moral and ethical norms in AI agents. While the first issue is mostly philosophical in nature, this second issue is engineering-related and depends on the level of morality that is wanted for those AI agents. In particular, four distinct levels of morality can be presented for AI agents Moor (2006): ethical impact agents (e.g., robot jockeys), implicit ethical agents (e.g., safe autopilot), explicit ethical agents (e.g., using formal methods to estimate utility), and full ethical agents. While the first two levels of moral AI agents are pretty easy to construct, the last two are more difficult to come by. The main reason is that the best available engineering mechanism to implement moral and ethical norms into AI agents is to directly program it in their code (i.e., morality is produced operationally). However, some authors (Hakli & Mäkelä, 2019) argue that this way of implementing morality is misguided, since it would only allow AI agents to be competent about morality, but it would not allow them to understand morality. Another engineering possibility is to allow those AI agents to develop their own morality (i.e., morality is produced functionally). However, this way of proceeding is dangerous since the exact kind of morality that the AI agent might develop is unpredictable. Moreover, this would partially conflate with international public policies (e.g., the European guidelines on Trustworthy AI) that require constant supervision on the development of Artificial Intelligent systems.

[28] Although, for a first step in the right direction see Awad et al. (2018).

[29] See, e.g., [65]

Summing up, as was for affective attitudes, it seems that normative-based theories of trust are adequate in online environments only if the interacting agents are human beings. In fact, creating a shared normative infrastructure in online environments might be feasible for smaller scale communities of human agents. However, if it is assumed that operational morality is not genuine morality, then, given the current guidelines for the development of Artificial Moral Agents, it seems unlikely that AI agent will ever be able to fruitfully participate in those moral/ethical communities. Thus, when normative-based theories of trust are assumed, it appears as if AI agents are not suited to be trusting agents (given the premises, neither in online environments nor in the physical world).

4 Conclusion and Future Works

Going back to the main question of the paper:

$$\text{Is online trust possible?} \qquad (Q_1)$$

The analyses provided in this paper point at a clear response: *it depends*. The first element that can influence whether the answer is yes or no is the theory of trust that is assumed. For instance, doxastic theories of trust might be better suited to allow online trust to emerge, while affective theories of trust might cause some problems when the trustor is an AI agent. A second element that can influence the answer is the specific nature of online environments. It is unlikely that a positive answer can be obtained for online environments in general (it might not even be possible to obtain such a general answer for physical environments). The way specific online environments are built can have a huge influence on whether trust can emerge in those environments. As an example, take the first instances of the Internet. At the beginning of its life cycle, Internet was employed by researchers to spread their ideas and their results. Such environment was simply a direct extension of the physical world and the overall spirit of all agents present in the environment was a collaborative one. Social, ethical and moral norms were borrowed from those of academia and trust characterized most of the relationships that were formed in such environment. Now think of Internet today and think about the various mechanisms that are employed to manipulate data and influence people's behaviours.[30] Even though different, the technology did not change that much from the past. However, the elements that are contained in such environment changed it to the point that trust is actually a rarity rather than the norm. This should make it clear that it is impossible to discuss about the possibility of having trust in an online environment in general, but specific features of the environment must be discussed first. Nonetheless, in the paper, it was shown that it is indeed plausible to implement features in online environments that might enable the emergence of trust. Therefore, while the answer to Q_1 in general is *no*, when specific design techniques and feature implementation are taken into consideration, the answer might become *yes*. Another important aspect highlighted in the paper is that the nature of the agents involved could potentially limit the applicability of some definitions of trust. This is because artificial intelligent agents are still lacking the relevant features (i.e., affective capacities and genuine moral values) that allow the emergence of some kinds of trust. However, other forms of trust seem to be compatible

[30] A famous recent case is that of Cambridge Analytica.

with AI agents being able to trust each other (in particular, basic forms of trust seen as reliance).

Concluding, it must be noted that this is only a small step in the huge debate on online trust. However, given the increasing importance of online interactions, this step is an important one. Moving from here, similar analyses might be required also for other forms of trust, such as organizational trust or therapeutic trust. This is because without an omni-comprehensive view of all the possible trusting relationships, it would be really hard to understand how to obtain the full benefits of trust in an online environment.

References

Alchourron, C., Gardenfors, P., & Makinson, D. (1985). On the logic of theory change. *Journal of Symbolic Logic, 50,* 510–530.

Aldini, A., Curzi, G., Graziani, P., & Tagliaferri, M. (2021). Trust evidence logic. In *Symbolic and quantitative approaches to reasoning with uncertainty* (pp. 575–589).

Aldini, A., & Tagliaferri, M. (2020a). Logics to reason formally about trust computation and manipulation. In A. Saracino & P. Mori (Eds.), *Emerging technologies for authorization and authentication (ETAA2019)* (pp. 1–15). LNCS 11967.

Aldini, A., & Tagliaferri, M. (2020b). Logics to reason formally about trust computation and manipulation. In *Emerging technologies for authorization and authentication* (pp. 1–15).

Aldini, A., & Tagliaferri, M. (2020c). A trust logic for the varieties of trust. In *Software engineering and formal methods* (pp. 119–136).

Alonso, F. M. (2016). Reasons for reliance. *Ethics, 126,* 311–338.

Alonso, F. M. (2014). What is reliance. *Canadian Journal of Philosophy, 44*(2), 163–183.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563,* 59–64.

Baier, A. (1986). Trust and antitrust. *Ethics, 96*(2), 231–260.

Barber, B. (1983). *The logic and limits of trust*. New Brunswick: Rutgers University Press.

Brengman, M., & Karimov, F. P. (2012). The effect of web communities on consumers' initial trust in B2C e-commerce websites. *Management Research Review, 35*(9), 791–817.

Calvo, R. A., D'Mello, S. K., Gratch, J., & Kappas, A. (Eds.). (2014). *The Oxford handbook of affective computing*. Oxford: Oxford University Press.

Carter, J. A., & Simion, M. (2021). *The ethics and epistemology of trust*. Internet Encyclopedia of Philosophy. Retrieved January 2021, from https://iep.utm.edu/trust/.

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

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies, 58,* 737–758.

Crick, F. (1989). The recent excitement about neural networks. *Nature, 337,* 129–132.

Dasgupta, P. (1988). Trust as a commodity. In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 49–72). Oxford: Blackwell.

Domenicucci, J., & Holton, R. (2017). Trust as a two-place relation. In P. Faulkner & T. Simpson (Eds.), *The philosophy of trust* (pp. 149–160). Oxford: Oxford University Press.

Ess, C., & Thorseth, M. (2011). *Trust and virtual worlds*. New York: Peter Lang.

Evans, D. (2011). *The Internet of Things: How the next evolution of the Internet is changing everything*. CISCO white paper.

Floridi, L., & Taddeo, M. (2011). The case for e-trust. *Ethics and Information Technology, 13*(1), 1–3.

Frost-Arnold, K. (2014). The cognitive attitude of rational trust. *Synthese, 191*(9), 1957–1974.

Gambetta, D. (Ed.). (1988). *Trust: Making and breaking cooperative relations*. Oxford: Blackwell.

Grabner-Kräuter, S., & Schratt-Bitter, S. (2013). Trust in online social networks: A multifaceted perspective. *Forum for Social Economics, 44*(1), 48–68.

Greene, J., Rossi, F., Tasioulas, J., Brent-Venable, K., & Williams, B. (2016). Embedding ethical principles in collective decision support systems. In *Proceedings of the AAAI conference on artificial intelligence*, *30*(1).

Grodzinsky, F. S., Miller, K., & Wolf, M. J. (2010). Toward a model of trust and e-trust processes using object-oriented methodologies. *Ethicomp 2010 proceedings*.

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275.

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

Hinchman, E. S. (2017). On the risks of resting assured: An assurance theory of trust. In P. Faulkner, & T. Simpson (Eds.), *The philosophy of trust* (pp. 51–69).

Hurley, R. F., Gillespie, N., Ferrin, D. F., & Dietz, G. (2013). Designing trustworthy organizations. *Sloan Management Review, 54*(4), 75–82.

Jarvenpaa, S. L., & Leidner, D. E. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791–815.

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

Jøsang, A. (2007). Trust and reputation systems. In A. Aldini, & R. Gorrieri (Eds.), *Foundations of security analysis and design IV* (pp. 209–245).

Kamvar, S. D., Schlosser, M. T., & Molina, H. G. (2003). The eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the 12th international conference on world wide wide* (pp. 640–651).

Kelp, C., & Simion, M. (2020). What is trustworthiness? Manuscript.

Keren, A. (2020). Trust and belief. In J. Simon (Ed.), *The Routledge handbook of trust and philosophy*. London: Taylor and Francis Group.

Keymolen, E. (2016). *Trust on the line: A philosophical exploration of trust in the networked era*. Oisterwijk: Wolf Publishers.

Lahno, B. (2017). Trust and collective agency. In P. Faulkner, & T. Simpson (Eds.), *The philosophy of trust* (pp. 129–148).

Lieto, A. (2021). *Cognitive design for artificial minds*. London: Routledge.

Luhmann, N. (1979). *Trust and power*. New York: John Wiley and Sons Inc.

McGeer, V. (2008). Trust, hope and empowerment. *Australasian Journal of Philosophy, 86*(2), 237–254.

McLeod, C. Trust. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, Fall 2020 edition. Retrieved from https://plato.stanford.edu/archives/fall2020/entries/trust/

McLeod, C. (2002). *Self-trust and reproductive autonomy*. Cambridge: MIT Press.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21.

Nickel, P. J. (2007). Trust and obligation-ascription. *Ethical Theory and Moral Practice, 10*(3), 309–319.

Nickel, P. J., & Vaesen, K. (2012). Trust and risk. In S. Roeser (Ed.), *Handbook of risk theory* (pp. 857–876).

Nissenbaum, H. (2001). Securing trust online: Wisdom or oxymoron. *Boston University Law Review, 81*(3), 635–664.

Papadopoulou, P. (2007). Applying virtual reality for trust-building e-commerce environments. *Virtual Reality, 11*(2–3), 107–127.

Papadopoulou, P., & Kanellis, P. (2019). Online trust and the importance of interaction. *International Journal of Technology Marketing, 13*(1), 21–50.

Perlis, D. (2000). The role(s) of belief in AI. In J. Minker (Ed.), *Logic-based artificial intelligence* (pp. 361–374). Berlin: Springer.

Pinyol, I., & Sabater-Mir, J. (2013). Computational trust and reputation models for open multi-agent systems: A review. *Artificial Intelligence, Review, 40,* 1–25.

Primiero, G., & Taddeo, M. (2012). A modal type theory for formalizing trusted communications. *Journal of Applied Logic, 10*(1), 92–114.

Potter, N. . N. (2020). Interpersonal trust. In J. Simon (Ed.), *The Routledge handbook of trust and philosophy* (pp. 243–255). London: Taylor and Francis Group.

Ryan, R. M. (Ed.). (2012). *The Oxford handbook of human motivation*. Oxford: Oxford University Press.

Sapp, J. E., Torre, D. M., Larsen, K. L., Holmboe, E. S., & Durning, S. J. (2019.) Trust in group decisions: A scoping review. *BMC Medical Education 19*, Article Number 309.

Simon, J. (Ed.). (2020). *The Routledge handbook of trust and philosophy*. London: Taylor and Francis Group.

Scheman, N. (2020). Trust and trustworthiness. In J. Simon (Ed.), *The Routledge handbook of trust and philosophy* (pp. 28–40). London: Taylor and Francis Group.

Schwitzgebel, E. Belief. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy, Fall 2019 Edition*. Retrieved from https://plato.stanford.edu/archives/fall2019/entries/belief/

Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-Trust. *Mind and Machines, 20*(2), 243–257.

Tagliaferri, M. (2019). *A logical language for computational trust*, Ph.D. Thesis, University of Urbino.

Tagliaferri, M., & Aldini, A. From belief to trust: A quantitative framework based on modal logic. *Journal of Logic and Computation*, forthcoming.

Tagliaferri, M., & Aldini, A. (2018a). A trust logic for pre-trust computations. In *21th International conference on information fusion (Fusion'18)* (pp. 2010–2016), IEEE.

Tagliaferri, M., & Aldini, A. (2018b). From knowledge to trust: A logical framework for pre-trust computations. *12th IFIP international conference on trust management (IFIPTM'18), IFIP AICT 528* (pp. 107–123), Springer, Berlin.

Trivers, R. .L. (2002). *Natural selection and social theory: Selected papers of Robert Trivers*. Oxford: Oxford University Press.

"Trolley Problem". In Britannica. Retrieved from https://www.britannica.com/topic/trolley-problem

Turilli, M., Taddeo, M., & Vaccaro, A. (2010). The case of online trust. *Knowledge, Technology and Policy, 23*(3–4), 333–345.

Velleman, J. D. (1992). The guise of the good. *Nous, 26*(1), 3–26.

Vries, P. D. (2006). Social presence as a conduit to the social dimensions of online trust. In W. I. Jsselsteijn, Y. D. Kort, C. Midden, B. Eggen, & E.vD. Hoven (Eds.) *Persuasive technology* (pp. 55–59).

Wang, Y. D. (2005). An overview of online trust: Concepts, elements, and implications. *Computers in Human Behaviour, 21*(1), 105–125.

Wichman, H. (1970). Effects of isolation and communication on cooperation in a two-person game. *Journal of Personality and Social Psychology, 16,* 114–120.

Williamson, O. (1993). Calculativeness, trust, and economic organization. *Journal of Law and Economics, 36*(2), 453–486.

Yu, B., & Singh, M. P. (2003). Detecting deception in reputation management. *Proceedings of the 2nd international joint conference on autonomous agents and multiagent systems (AAMAS'03)* (pp. 73–80). ACM Press.

Mirko Tagliaferri received his bachelor in Philosophy at the University of Glasgow, where he followed a program mainly focused on logic and philosophy of science. Then, he completed a doctoral program at the University of Urbino, with a thesis on computational trust. In his thesis, he proposed a novel logical formulation of trust that could be employed to reproduce trust in digital environments. He is currently a PostDoc researcher at the University of Urbino and an Adjunct Professor at the University of Verona. He lectures on philosophical issues related to Artificial Intelligence and he continues his research on trust and ledger technologies, with a focus on digital representations of social concepts.