Formal Ontology in Information Systems
N. Aussenac-Gilles et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA231117

On the Relation of Instrumental Dependence

Luca BICCHERI [a,1], Stefano BORGO [a] and Roberta FERRARIO [a]
[a] *ISTC-CNR Laboratory for Applied Ontology, via alla Cascata 56c, 38123, Trento, Italy*

Abstract. The notion of trust has been traditionally investigated within many disciplines, ranging from sociology to economy, as well as politics, psychology, and philosophy. More recently, it is especially in the fields of AI, ICT, and Engineering (e.g., Critical systems), that the need for a discussion on the concept of trust, problematized in relation to the massive employment of technical artefacts in modern society, is becoming urgent. Yet, being a characteristic trait of human relationships, it is not clear whether the attitude of trust can also be directed towards artefacts. Moreover, with respect to the study of systems' failures, the engineering sciences provide cognate notions to that of trust, e.g. reliability or dependability, which highlight our dependence on complex systems to fulfil certain tasks in a context of risk, uncertainty and vulnerability. In order to understand how far we can rely on technology, we should be able to understand, first of all, which kinds of dependencies are at stake. To this aim, in this paper, we will briefly review and discuss the main theoretical points related to trust and the technical notions mentioned, looking at both humanities and engineering literature. Then, we shall propose a preliminary ontological analysis aiming at comparing the specificities of the concepts concerned, all sharing a form of instrumental dependence.

Keywords. Instrumental dependence, trust, dependability, reliability, confidence, technical artefact, goal-oriented agents

1. Introduction

Modern society is characterized by technologically dense environments, in which the actions of persons, collectives and technical artefacts are inextricably intertwined (e.g. industry 4.0). Such environments have been analyzed, in the last decades, through the lenses of the socio-technical systems paradigm. Among socio-technical systems, critical systems (like, e.g. hospitals or nuclear plants) are – not surprisingly – gaining increasing attention since, as technology advances, we rely more and more on it and, though on the one hand it makes the task we have to accomplish easier, on the other, its complexity decreases the control we can have on it and, thus, makes us more vulnerable (to its failures and its misuse). What has definitely increased with the complexity of the systems is the interdependence of their components (including agents) for the correct execution of the systems' tasks and the accomplishment of their goals. For this reason, there is an urge to develop a high-level framework to understand the ways in which agents depend

[1] Corresponding Author: Luca Biccheri, via alla Cascata 56C, 38123, Trento, Italy; biccheriluca@gmail.com

on other agents or artefacts in view of bringing about certain goals, while tolerating certain levels of risk, uncertainty and vulnerability. To this aim, we will analyze, from an interdisciplinary perspective, the long-lasting philosophical debates on the notion of trust by linking it to the engineering field of dependable systems. Our starting point will be the intrinsic instrumental value and the goal-oriented structure of the relation of trust and our enquiry will revolve around the question whether there is something specific in the relation of trust that makes it applicable only between human beings or if, taken in its generality, it can be applied also to artefacts and systems. The paper will proceed as follows: in section 2 we will discuss the main controversies emerging from the philosophical literature on trust, while in section 3 we will analyze some cognate notions developed in the literature in engineering, such as reliability and dependability. Section 4 is the core of the paper and is dedicated to the comparison of what has emerged from the different debates, with the purpose of introducing a high-level ontological relation that we shall call 'instrumental dependence (ID)'. By keeping focused on *ID*, we shall be able to argue that there is a sense in which the attitude of trust can be directed both to human agents and artefacts. Furthermore, our analysis will show that if on the one hand, the notions of 'trustworthiness' and 'dependability' are the same, on the other hand, these must be distinguished from that of 'reliability'. In addition, *ID* will also help us to clarify the difference between the attitude of trust and what we are used to call 'confidence'. Finally, section 5 concludes the paper and suggests the directions of its future developments.

2. Theories of trust

In this section, we will provide a brief overview of the most recurrent themes on the notion of trust, while underlying the ways in which scholars have debated in the literature the sense of this notion with respect to artefacts. It is customary to distinguish between 'trust' on the one hand, and 'trustworthiness' on the other. While the former corresponds to an attitude of the trustor (i.e. the one who trusts) toward the trustee (i.e. the one who is trusted), the latter is a property instantiated by the trustee[2]. While hardly anyone is willing to question that the trustor must be a cognitive agent, many disagree on whether the role of the trustee can be played only by agents or also by artefacts. Those who claim that only cognitive agents, and more specifically only human beings, can play the role of trustee appeal either to necessary features related to the trustor stance (e.g. the possibility to feel betrayed by the trustee) or to characteristics that are deemed to be necessary to explain why, i.e. the motivations according to which the trustor grants trust to the trustee (e.g. their capabilities, self-confidence, willingness, persistence, morality, etc). As has been stressed in [2], all these arguments leverage the fact that features that characterize trustworthiness and related implicatures on the attitude of trust cannot be possessed by artefacts or, more generally, inanimate objects. On the other hand, scholars who claim that also artefacts can play the role of trustee may appeal to Dennett's theory of intentional stance [3] (arguing that it is an 'abstraction' useful to ascribe trustworthiness to artefacts in order to predict their behaviour), or to a weaker sense of trust than the full-fledged notion of trust between humans [2,4,5] or, ultimately, to the claim that in using

[2]For a general introduction to the concept of trust, see [1].

technologies we are in the face of a scenario in which trust is typically involved, i.e. a condition of vulnerability, uncertainty and risk [6,7] (think about the issues raised by AI technologies). Another option available is the pluralistic view, i.e. the attitude of trust comes in many forms that are reflected in both common linguistic usage and specialized academic jargon, although such forms share a common root [8]. Be that as it may, a relevant assumption that emerges from the literature is that the trustor grants their trust to the trustee on the basis of certain goals to be achieved. This points out a reason that could partially explain why, generally speaking, it is difficult to establish specific reasons or motivations to trust. From time to time, such motivations depend on the trustor's goals. So, someone or something is not deemed to be trustworthy in absolute terms, but always with respect to some tasks or goals (which in turn depends on their desires, beliefs, values emotions, etc). So, at first glance, if x stands for a trustor, y for a trustee, and z for a goal, then we can represent trust as it is classically understood, i.e. a ternary relationship, that is $T(x,y,z)$. Moreover, an additional issue to be raised is that of reliance. The word 'reliance' refers to an attitude similar to, although distinct from, trust. A common idea shared within the literature on trust is that the notion of trust is stronger than that of reliance, in the sense that the former implies the second, but not vice versa [1]. The notion of reliance refers to a kind of predictive belief or expectation related to the behaviour of the trustee in the future. Yet, the definition of trust is sometimes directly reduced to that of reliance *qua* predictive expectation. In this respect, trust can be explained in tandem with the notion of risk and subjective probability. So for example, the sociologist Diego Gambetta provided a definition of trust in which such attitude is conceived of as "the subjective probability with which an agent expects that another agent or group of agents will perform a particular action on which its welfare depends" [9]. The definition of trust in terms of probability has been dubbed 'the rational-choice account' of trust [10]. In criticizing the rational-choice account of trust, Castelfranchi and Falcone [11,12] have stressed that trust implies more than the notion of risk related to the possible outcome of a choice based on subjective probability. In fact, the attitude of trust is a complex cognitive attitude that consists of a set of beliefs, needs, desires, intentions, evaluations and expectations about the trustee that, ultimately, lead the trustor to take the decision to rely on the trustee in view of a certain goal. Castelfranchi and Falcone dubb the decision to rely on the trustee with the term 'reliance'. Note that Castelfranchi and Falcone make therefore use of the term 'reliance' in a way that is not equivalent, in the sense that 'is not reducible', to that of 'predictive expectation' abovementioned. Yet, also for them trust is more than reliance, given that reliance (i.e. decision) is part of the attitude of trust. That is, in a nutshell, the core mental part of trust according to Castelfranchi and Falcone. Then, this decision is followed by an action which they call 'delegation'. The action of delegation is something that 'happens in the world', i.e. is observable by other agents. Now, the use of the term 'delegation' to refer to an action may be misleading, since it can suggest that between the trustor x and the trustee y there is always a more or less informal commitment (e.g. a handshake) or an explicit commitment (e.g. the signature of a contract). But this is not always the case. In fact, Castelfranchi and Falcone [12,13] distinguish between two kinds of delegations: a) strong delegation, that is at stake whenever there is a commitment between x and y; b) weak delegation, that does entail that y is aware of the fact that x trusts y. From the definition of weak delegation, it follows that on the role of the trustee can be applied both to agents (when they are not aware of the trust of the trustee side) and artefacts (that are not aware *tout court*). Another

critique of the rational choice account of trust has been offered by Paul Dumouchel who, however, also rejects *in toto* the idea of trust as a cognitive attitude, given that, for example, cognitive theories of trust do not provide clear criteria for distinguishing trust from other kinds of actions that agents perform based on the expectations about the behaviour of other agents [14][3]. Accordingly, Dumouchel reduces the notion of trust to an action that has particular implications for the trustor. In Dumouchel's words "To trust is to act in such a way as to give another agent power over us. In other words, when I trust I increase my vulnerability to another agent through an action of my own, and that action is precisely what trust is. If I had not acted I would not be vulnerable, or at least not as vulnerable to the other agent" [14]. Moreover, note that when x trusts y for a goal z, the achievement of the goal z is not wholly dependent on y. There are unforeseen variables within the environment that can influence the goal achievement. This represents the dimension of uncertainty. Furthermore, when x trusts y for a goal z, the domain in which the action takes place matters. e.g. parents can trust their oldest child to take care of their younger child in the home garden, but not outside of it. In the literature on trust all of this is summarized in the notion of context. From what has been said so far, it follows that the classic triadic relation of trust is sometimes enriched with further elements. So, in the literature trust is also represented as a quaternary relation involving a trustor, a trustee, an action, and a goal [6], or a quinary relation that, in addition, also has a context [12]. As a final remark, note that the notion of trust is dynamic with respect to time. For example, x can trust y at time t, but not a t' with $t' > t$. Yet, things are often more complicated, given that the property of trustworthiness is not a matter of all or nothing, i.e. it comes in degrees that, ideally speaking, can range from 0 to 1. The fact that the overall notion of trust can be described as a dynamic complex mental attitude relative to a goal, which is context-dependent and implies risks has been highlighted in a recent paper that offers an ontological account of trust applied to modelling the case of citizens' trust in central bank digital currency [15].

3. Engineering: from reliability to dependability

Looking at the development of the notion of dependability during the second half of last century, what stands out immediately is its historical connection to the engineering concept of reliability. Even if in everyday English the term 'reliability' is synonymous with 'dependability', within the engineering jargon the meaning of these terms does not necessarily overlap. Literature highlights that the first in-depth scientific discussions on the notion of reliability started around the 1950's [16,17,18]. In this respect, some remarkable epistemological and socio-economic changes that occurred during the World War II, i.e. the consolidation of statistical sciences and mass production, as well as the advent of the electronic revolution, were the main triggers (cf. [16,17]). Moreover, a decisive event was the large-scale use of the vacuum tube, which enabled a number of both military and commercial applications. However, the tube "was also the chief source

[3]Dumouchel seems to entirely reduce the cognitive perspective on trust to the rational-choice account of trust, which is not the case, as shown by the cognitive theory of Castelfanchi and Falcone, who agree that trust is more than an expectation. Yet, Dumouchel has the merit of clearly highlighting the fact that the dimension of vulnerability associated with the notion of trust is entailed by the actions performed as a consequence of the decision to trust.

of equipment failure. Tube replacements were required five times as often as all other equipments" [18]. In short, scholars agree that the need to accurately estimate failures using statistics, under the boost of mass production, led to the emergence of a new discipline, that is Reliability Engineering. Saleh and Marais [17] have stressed that during the 60's the science of reliability followed two main directions, namely an increasing specialization of the discipline and a shift of perspective from the component-level analysis to the analysis of complex systems, including safety issues, related to domains such as oil and gas industry, chemical industry, and nuclear power industry. These systems are typically called 'critical' to the extent that the occurrence of catastrophic system failures may cause fatalities or environmental disasters, as well as significant economic losses [19]. Furthermore, "as software became more pervasive, and was increasingly dependant upon in a growing number of applications, the need for software reliability assessment and improvement developed" [17]. And it was both the software and hardware system-level analysis to represent the turning point for introducing a new concept, i.e. that of dependability, meant to expand the notion of reliability towards a more general set of concepts and definitions useful to "discuss problems that might occur either within or between system components at any level of a system" [20]. Moreover, the computer scientist Jean-Claude Laprie was the first to suggest the label 'dependability' to refer to a broader concept than the former notion of reliability (cf. [20,21]). The studies made by Laprie and his colleagues were systematised in the book entitled 'Dependability: Basic Concepts and Terminology' [22]. According to Laprie, dependability can be defined "as the trustworthiness of a computer system such that reliance can justifiably be placed on the service it delivers" [22, p. 4][4]. We can observe that, within the definition, the emphasis is placed on the ability of the system to deliver a service. What prevents the system's ability to deliver services is the occurrence of failures. The notion of failure strictly depends on the users' perspective, in the sense that a "failure occurs when an error "passes through" the system-user interface and affects the service delivered by the system" [22, p. 19]. A failure is therefore a transition from the state of correct service to the state of incorrect service. The delivery of incorrect service is called 'system outage'. A system is characterized as "an entity having interacted or interfered, interacting or interfering, or likely to interact or interfere with other entities, i.e., with other systems" [22, p. 8]. Such other systems collectively represent the environment in which the system interacts. Services' users are conceived of as a system in their turn and are included into the environment. Furthermore, systems' specifications concern the description of both the expected functions and the expected services. The specifications are derived from the agreement (be it implicit or explicit) between persons or corporations. It is relevant to stress that the social facet of agreements plays a fundamental role in determining the dependability of a system, given that these agreements are "necessary in order that the specification can serve as a basis for adjudicating whether the delivered service is acceptable or not, or, equivalently, whether a failure has occurred or not" [22, p. 10]. At any rate, the word 'dependability' ultimately refers to a system property that can be decomposed into a set of basic attributes, which are the following: a) reliability: continuity of correct service; b) availability: readiness for correct service; c) safety: absence of catastrophic consequences on the user(s) and the environment; d) confidentiality: absence of unauthorized disclosure of information; e) integrity: absence of improper system state alterations; f)

[4]Note that the term 'trustworthiness' appears in the definition. As we shall see in section 4, there is no reason to distinguish the notion of 'trusworthiness' from that of 'dependability'.

maintainability: ability to undergo repairs and modifications. The theory of dependability is thus presented as a profitable way to subsume many relevant system's aspects under a single conceptual framework, also taking into account the threats (i.e. faults, errors, failures) that impact on the system's services, as well as the means (i.e. fault tolerance, fault removal, fault forecasting, fault prevention) through which such threats are faced. However, as noted in [23], a number of dependability concepts have been renamed or reinvented defining new terminologies in many fields related to dependability, leading to confusion. This seems to be confirmed also by [24], in which a comparative analysis of widely-used concepts within the area of information systems is made. The authors report that the integration of heterogeneous elements (including human agents) into complex systems has resulted in numerous different vocabularies to describe systems' performance across different fields, by underlying also the need to develop a common understanding of such concepts without referring to a specific discipline. In addition, it is common to think that the various aspects of dependability can be understood either in qualitative terms or in quantitative terms [22]. For instance, the concept of 'reliability' can be seen as an attribute or property of the system, i.e. continuity of correct service or as the measure employed to evaluate the continuity of correct service itself. Typically, a common measure of reliability is Mean Time To Failure (MTTF) or, alternatively, it can be expressed as a function of time [25]. More precisely, it can be defined as the probability $R(t)$ that no failure will occur from the time the system is put into operation to the time that the system's service will drop below a certain threshold, usually relatively high (e.g. $R = 0.9$). However, not all the concepts that describe the dependability of a system are easily quantifiable in mathematical terms, and this is especially true, for instance, for the attribute of security. Rather than having to do with random system's failures, the dimension of security mainly concerns malicious/intentional human attacks that are not analysable through probabilistic risk assessment techniques. Moreover, even when probabilistic methodologies are applicable, we have to deal with different kinds of uncertainties including simplifications related to mathematical modelling, insufficient consideration or knowledge of faults and errors, and inaccurate data (e.g. limited information available on new applications/technologies) [24]. Finally, dependability is not a binary concept (all or nothing), in the sense that a system is said to be more or less dependable as long as its services match acceptable thresholds, which in turns depend on both the application (e.g. communication, process control, data processing, etc.) and the domain (e.g.construction, transportation, industrial control) [26]. Thus, the notion of dependability is context-dependent.

4. Instrumental Dependence

4.1. Some theoretical remarks on the topic

After having illustrated various senses in which the notion of trust, as well as the notions of reliability and dependability, have been formulated within the literature, we are going to introduce our proposal on the topic, with the main attempt to outline an ontological framework broad enough to account for different relations in which agents and artefacts are engaged. To this aim, our analysis will start by looking at a set of core principles that can be generally applied either to humans or artefacts, taking its cue from the cognitive

model of means-end reasoning, which will lead us to introduce a high-level form of ontological relation that we shall call 'instrumental dependence (ID)'. The main reason for introducing *ID* is that we need a general concept that encompasses a wide range of dependencies between agents and artefacts, for understanding how agents and objects interact in order to achieve certain goals. For one thing, the introduction of *ID* will allow us to individuate the core assumptions shared by different forms of dependencies, for another thing, once identified such assumptions, we shall be able to better characterize the specific nature of the dependencies at stake. Before entering into the details of our proposal, we want to make a few remarks on the main hypothesis and motivations we decided to embrace in this work.

First of all, the kind of instrumental dependence we shall take into consideration is directly involved in the technical notions of reliability and dependability, as well as in those concepts that are usually characterized by a more cognitive and social aspect, such as the attitude of 'reliance' and 'confidence', which are deeply entangled with the broader notion of 'trust'. With respect to the latter notion, we are inspired by the pluralistic view (see section 2). So we do not claim to be able to provide a definitive answer to the question 'what does it mean to trust in a genuine sense?'. Nonetheless, we believe that it is possible to enucleate a handful of fundamental conceptual tools to grasp a substantive aspect of the notion of trust and, at the same time, engage in a fruitful discussion relative to the problem of contextualising such notion with respect to technology. In this respect, as we have seen in sections 2 and 3, the need for an interdisciplinary and thorough understanding of how much and in which sense we trust (or more generally depend on) artefacts such as AI and complex systems clearly emerges from the literature. Accordingly, in this section, we will try to highlight the meeting point between the theories of trust and those of dependability engineering. As we shall see, our analysis suggests that in either case, the agents' attitude and the resulting actions directed towards the goal achievement share the same instrumental value. Moreover, such value, although with different nuances, is also at stake within that kind of intimate relationship that we are used to call 'confidence'.

This being said, let's start our investigation, by gradually bringing to the attention some of the theoretical elements mentioned in sections 2 and 3, which we deem to be useful to define the notion of instrumental dependence (ID) involved in both trust and engineering literature. Speaking about the notion of trust, we said that a relevant assumption is that the trustor grants their trust to the trustee with respect to certain goals to be achieved. So, someone or something is not deemed to be trustworthy in absolute terms, but always with respect to some tasks or goals. Therefore, the notion of trust highlights an intrinsic instrumental value and a goal-oriented dependency. And the same goes for dependable systems. According to Laprie (cf. [22, p. 6-7]), the use of the term 'dependability' is meant to highlight the increasing dependence of contemporary society on complex systems' abilities to provide functionalities to satisfy human interests. As we have seen, the concept of 'dependable system' must be put into perspective, i.e. a system is dependable only with respect to the specific tasks assigned to it.

An additional issue to be raised is that of the attitude of reliance and the related engineering notion of reliability. With respect to the review of the theories of trust offered in section 2 , we said that a shared idea is that the notion of trust is stronger than that of reliance, which is a predictive belief or expectation related to the future behaviour of the trustee with respect to the goal achievement assigned by the trustor. Moreover, we

have also seen how some scholars [12,14] have criticised the so-called 'rational-choice account' of trust that instead reduces the notion of trust to that of reliance, quantified in terms of subjective probability.

Now, what has been said so far shall allow us to clearly show the link between the notion of trust and the other relevant engineering notion we are interested in, i.e. dependability. As we have observed in section 3, at some stage of development of the history of engineering, the technical notion of reliability has been questioned because there was the need to account for a broader notion, i.e. that of dependability (which includes, among others, the notion of reliability itself), so as to model the overall aspects of critical systems. Let us remind that these systems are typically called 'critical' to the extent that the occurrence of catastrophic system failures may cause fatalties or environmental disasters, as well as significant economic losses. Within the engineering theory of dependability, the notions of safety (i.e. absence of catastrophic consequences on the user(s) and the environment) and security – aka confidentiality – (i.e. absence of unauthorized disclosure of information) are those that most contribute to qualify some kinds of systems as 'critical' in the sense above mentioned, and which therefore require an account for our dependence on complex systems that extends far beyond the ability of a system to remain functional over time[5].

The parallel between the notion of trust and that of dependability, whereof the dimensions of safety and confidentiality imply the dimension of vulnerability usually ascribed to agents that entertain the attitude of trust, is therefore evident. So it is no coincidence that, besides being used within the classic definitions of dependability (see section 3), in the recent engineering literature the term 'trustworthiness' begins to be used directly in place of the word 'dependability' [26]. Notice that, within the engineering literature, the word 'trust' is not used to point at an attitude, rather the term 'trustworthiness' is employed, which is, as said, a property, and not an attitude. So, one may claim that the use of the terms 'dependability' or 'trustworthiness' applied to complex systems underlines the fact that engineers are directly interested in evaluating the property of a system, i.e. what makes a system trustworthy or dependable in an objective sense, leaving aside the subjective agents' attitude towards these systems. Yet, this is not completely true, as shown by the influential work of Laprie [22] which links the meaning of what counts as a failure to the agents' perspective and make them parts of the system. So, the agents' attitude towards the dependability of a system, although being somehow implicit, is still there and plays a fundamental role in establishing why and when a system is dependable. To sum up, it is interesting to note that the literature on trust highlights that the attitude of trust is more than reliance, which implies that the property of 'trustworthiness' is more than that of 'reliability', and, similarly, the engineering literature suggests that the property of 'dependability' is more than that of 'reliability'. This means that both trustworthiness and dependability must be thicker notions than reliability and the related attitude of reliance.

As a final remark, let us sum up other theoretical elements that emerged from section 2 and section 3 and that are relevant to our purposes. Generally speaking, we underlined that when an entity x is deemed to be trustworthy or dependable, it is so always w.r.t. a goal to be achieved, and not in absolute terms. There are unforeseen variables within the environment that can influence goal achievement. This represents the dimension of un-

[5] Other attributes of dependability, like availability and maintainability, although related to security and confidentiality, are more directly linked to the functionalities of the system.

certainty that is often called into question both in the literature on trust and dependability engineering. This notion shall be distinguished from risk estimation, which instead concerns those aspects related to the attitude of reliance and the associated property of reliability that are quantifiable through probabilistic or stochastic methodologies. Finally, note that the notion of trust, as the engineering notion of dependability, is dynamic with respect to time, comes in degrees, and is context-dependent.

In conclusion, from what has been said, we claim that using the word 'trustworthiness' or 'dependability' with respect to complex artefacts such as critical systems is more a linguistic matter, than an ontological issue. Furthermore, we believe the property of 'trustworthiness' can be instantiated by both human agents and artefacts. And this brings us back to the debate raised in section 2 concerning whether the attitude of trust can be directed only towards human agents or also towards artefacts. Let us explain better. Generally speaking, we suggest that such debate overlooks the most noticeable feature of trust, i.e. its instrumental structure. By keeping focused on this evidence, the debate can be seen under a simpler perspective. As we shall see, when we say that an agent or an artefact is 'trustworthy', even taking into account some non-negligible differences, we are putting in place the same form of instrumental dependence. Such instrumental dependence is triggered by a complex mental attitude that is involved in all kinds of means-end reasoning, including the attitude of trust. In other words, we are claiming that the attitude of trust in its general sense is nothing but a type of instrumental reasoning, which however has specific consequences on the action side. Moreover, the same agents' attitude and the resulting kinds of actions are also at stake in other contexts in which agents and artefacts are engaged, including the scenario of dependability engineering. So, all in all the term 'trustworthiness' can be used as a synonym for 'dependability' in all respects. To see why, first we have to introduce the theoretical picture of means-end reasoning, after which we will present the relation of instrumental dependence (ID).

Cognitive agents are limited in terms of computational power. Agents live in a physical and social dynamic context where it is difficult to foresee what will happen. The environment pushes us to take decisions quickly. Thus, agents are resource-bounded in a double sense. On the one hand, they have limited access to information that could be relevant for the choice. On the other hand, in order to make an inference, agents have at their disposal limited time and memory space. Due to the challenges of a fast-changing environment and the limitations of decision-making processes, very detailed plans are not very useful in the long term. Yet, to achieve complex goals we have to think rationally, structuring all relevant actions step by step according to our intentions. A plan can be considered as mental attitude on par with intentions. Yet, it is a complex mental attitude that involves different kinds of mental states, such as beliefs, desires, needs, expectations, etc. However, intentions as mental attitudes represent the fundamental elements to coordinate plans, intentions are "so to speak, the building blocks of such plans; and plans are intentions writ large" [27, p. 8]. So we propose to better qualify the kind of complex attitude we mentioned above as a plan for the decision making process that leads to actions.

4.2. The ontological structure of instrumental dependence

Now it is time to introduce the notion of instrumental dependence (*ID*), which is meant to represent the fact that an agent is instrumentally dependent on another agent or artefact

to bring about a goal. To this aim, we introduce such a relation as a quaternary relation, which is formally defined as follows[6]:

$$ID(x,y,z,w)APO(x) \wedge POB(y) \wedge PD(z) \wedge Goal(w) \wedge \exists t.(sat(z,w,t) \wedge PC(y,z,t)) \quad (1)$$

With reference to DOLCE categories, the variable x ranges over agentive physical objects (*APO*), which can be more accurately characterized as intentional agents, i.e. agents whose mental states (*MS*) exhibit the feature of *intentionality*, also known as *aboutness*[7]. Differently from x, the variable y ranges over physical objects (*POB*), which can be other intentional agents[8] or technical artefacts (see [29]), that is to say, every physical object intentionally designed by agents so as to accomplish some goals. In line with [30], agents' goals are conceived of as intentional mental states that are about those entities called 'satisfiers'. According to our framework, satisfiers (*sat*) are perdurants (*PD*) that, occurring at a certain time, satisfy agents' goals. Finally, thanks to the relation called participation (*PC*), agents or artefacts come into play with respect to the occurrence of perdurants. Thus, overall, *ID* expresses the idea that an agent's goal can be satisfied only through the occurrence of a perdurant in which the entity towards which such an agent is instrumentally dependent participates. Furthermore, note that so far we have talked about actions that must be performed in order to achieve a goal, given that we wanted to keep our discussion aligned with the literature examined in the previous sections for the sake of clarity. However, as it should be clear from the above, goals can be achieved by means of every kind of perdurant (e.g. events and processes, as well as actions). But for the sake of simplicity, we shall keep talking about actions in the examples which follow. This being said, let's make an example contextualizing *ID* to trust. The plan is triggered by a goal w, arguably an intention, that the trustor would like to bring about. Yet, for whatever reason, the agent is not able/does not want to reach w by themselves. So, inevitably, the trustor has to think about how to achieve w by means of some trustee. In our view, the role of trustee is played by a physical object. Let's take the more familiar case in which the trustee is an agent. Now imagine the case of a job interview in which there are alternative candidates y, p, q. Let's say that the trustor finally chooses y on the basis of a set of beliefs, needs, desires, evaluations and expectations about the trustee y, and y accepts the job. As the last step, the trustor and the trustee establish a commitment, i.e. they sign a contract. This commitment is what Castelfanchi and Falcone [12] call 'strong delegation', which is an action that is at stake whenever there is a commitment between x and y, whereas their notion of weak delegation does not entail that y is aware of the fact that x trusts them. Yet, both strong and weak delegation refer to a kind of action that is different from the kind of action that constitutes an argument of the relation *ID*. Such delegations *qua* actions seem somehow to 'objectify' (make public, observable) the final result of the trustor's decision-making process as a complex mental attitude. Even

[6] At this stage of research, we prefer not to provide a definition of instrumental dependence in terms of necessary and sufficient conditions. There are still thorny aspects that need to be clarified. To begin with, we need to better understand which temporal constrains have to be applied to the relation between agents' goals and the occurrence of the perdurants that satisfy them. Therefore, in the statement below, we simply aim to express a necessary condition and to give an interpretation to the variables of the main predicates at stake.

[7] For a wider study of the theory of intentionality and its application to DOLCE, see [28].

[8] Note that, as we shall see in section 4.3, agents can be instrumentally dependent on their proper parts, such as cognitive parts or bodily parts.

if these kinds of actions are relevant for trust, we are looking elsewhere to capture the idea of instrumental dependence.

Coming back to our example, suppose that the job interview has been carried out to find someone suitable to play the role of 'Professor of Phonetics' and agent y has been chosen. Now y should be able, competent, willing, whatever you want, to carry out the goal w, let's say, 'teaching phonetics', which actually is a complex goal decomposable into many subgoals. Now, generally speaking, what's relevant for the trustor is the action (or set of actions) that the trustee performs to fulfill the main goal. The action of delegation (the signature of the contract in our example) that Castefranchi and Falcone introduce comes into play just to assign a goal w that must be fulfilled by the trustee through the performance of an action z that is distinct from the action of delegation. So in our view, within the quaternary relation $ID(x, y, z, w)$, the variable z stands for the action (or set of actions) that the trustee performs to fulfill the main goal. That is to say, the action of delegation is not captured by the quaternary relation, so we implicitly suppose that that act of delegation has been performed earlier.

We can observe how *ID* binds the trustor to the trustee, as a consequence of the act of delegation in view of the achievement of a goal. *ID* presupposes the decision-making process that *qua* plan is in line with the analysis of trust as a complex mental attitude offered by the cognitive theories of trust [9]. Now, *ID* is at stake even when an agent is instrumentally dependent on an artefact for the goal achievement. Also in this case, *ID* presupposes that the agent has formulated a plan. If the notion of plan is suitable for capturing the idea of trust as a complex mental attitude, we see no reasons why the *ID* relation between agents and artefacts should not be triggered by the attitude of trust. Obviously, we are not claiming that there are no differences between trusting an agent and trusting an object. It is quite clear that, on the one hand, these differences lie in the content of the beliefs, needs, expectations, intentions, etc. involved in the plan. In trusting agents we are interested in evaluating properties that are different from the properties that would be at stake in evaluating artefacts. In this respect, the complex property of 'trustworthiness' will refer to a different set of properties. Moreover, we build very different emotional bonds with agents and artefacts. All these are common platitudes, which we won't discuss. Rather, we are saying that, both in the case of humans and artefacts, the trustor's means-end reasoning structure and the relation of *ID* are the same. In other words, at a high ontological level, the attitude of trust does not differ from a classic kind of plan which includes another agent or artefact and may lead to the *ID* relation.

4.3. Trust, vulnerability and confidence

Now, as we shall see, the consequences of *ID* on the action side, which follow the act of delegation, are the essential aspects to define trust. Furthermore, if the property of 'trusworthiness' can also refer to artefacts, there is *a fortiori* no reason to distinguish the

[9] However, theories of trust classically represent an agent as a rational agent, i.e. an agent whose actions are motivated by their personal interests and who take a decision intentionally and consciously. Yet, the literature offers counterexamples in which agents act irrationally as long as they may trust people even if knowing that they are untrustworthy [31]. Biological factors, like the production of oxytocin, unconscious biases and emotions may alter the deliberative process that leads a trustor to trust the trustee [32]. Therefore, it is evident that the cognitive perspective is a simplified model that does not take into account non-intentional elements that influence the attitude of trust.

notion of 'trusworthiness' from that of 'dependability'. The *ID* relation is clearly instrumental, given that the trustor achieves the goal only by means of the actions performed by the trustee. Note, however, that the performance of these actions is just a necessary condition to the goal-achievement. For example, there are unforeseen variables within the environment that compromise goal achievement (recall the dimension of uncertainty). In addition, trust does not exclude cooperation, thus some actions on the trustor side could also be required to the goal-achievement [12]. The relation of instrumental dependence can shed light on a pivotal feature of trust, i.e. vulnerability. In general, vulnerability is a recurrent relevant theme within the literature on trust and is entailed by the notion of safety and security mentioned in the engineering literature on dependability (see sections 2 and 3). However, note that from the perspective of the theory of trust, the dimension of vulnerability should not be just considered as a relevant dimension among others, rather it is constitutive or essential to the meaning of trust. To see why, note that, leaving aside vulnerability, trust would not be distinguishable from other forms of *ID* involving the interactions between agents, such as cooperation which, as trust, involves a scenario of decision-making, delegation, risk and uncertainty. Yet, given that cooperation can take place even in absence of trust, trust must entail something more than cooperation, i.e. vulnerability [14]. However, the notion of vulnerability is very broad, what does it mean to be vulnerable?

First of all, the meaning of being vulnerable depends on the trustor, so it is quite subjective. Yet, it can be parameterized to the value attributed to the goal achievement: the higher the value of the goal that the trustor intends to bring about by means of the trustee, the more vulnerable the trustor, if such a goal would not be achieved by the trustee. From this follows that vulnerability is strictly related to the autonomy of the trustor as an agent. In literature, we can find several definitions, which often differ significantly in meaning [33]. Nonetheless, it is possible to individuate a minimum set of requirements to define autonomy in its broadest sense [34], that is: I) agents should be able to set their own goals based on their beliefs, desires, intentions etc; II) agents should be able to reach goals on their own. Now, the second sense of autonomy is directly at stake in *ID* between the trustor and the trustee, making the former more vulnerable: by renouncing their autonomy, the trustor gives the trustee power over them with respect to the achievement of a certain goal.

Differently from trust, instrumental dependence (*ID*) is just about goal achievement from a means-end point of view, regardless of the specificity of trust that instead implies vulnerability. So *ID* is more general than trust. From what has been said, it follows that *ID* can also cover the dimension of cooperation between agents. Yet, *ID* is more general than cooperation as long as it can be applicable to cases in which, for example, the agent is instrumentally dependent on the actions performed by a particular artefact. When this happens, *ID* can cover the case in which we are entertaining an attitude of trust with weak delegation, as defined by Castelfranchi and Falcone [12]. Furthermore, *ID* can also be generally applied to complex systems – as critical systems (see section 3) – understood as relational networks in which agents and many different kinds of artefacts interact, with the latter performing actions related to various goals of the former.

Finally, there is a form of interaction between agents and artefacts that differs from all others, e.g. artefacts toward which the agents' attitude is directed and that become part and parcel of the agents' functions (examples range from walking sticks to advanced bionic limbs). In these cases, it is as if the mental attitude (involving risk, evaluation,

decision etc) disappeared from the context, leaving space only for the action side. Within the literature on trust, this kind of interaction has been recently emphasized by Nguyen [2] who claims that, for example, when climbers trust their ropes to climb a mountain, a specific form of trust comes into play, which is dubbed 'unquestioning attitude'. In this respect, climbers stop questioning whether their ropes will perform their functions, integrating such artefacts into their own agency[10]. Yet, as Nguyen himself suggests, this does not mean that when an agent trusts an artefact in this sense, they never question it at all, rather it implies that agents have a general disposition not to do it [11]. Moreover, Nguyen extends this type of trust beyond the interactions with artefacts, covering the case in which such an attitude is related to agents' proper parts, such as cognitive parts (e.g. memory) or bodily parts (e.g. hands). In such cases, the unquestioning attitude is at stake as long as these parts start to not properly perform their functions (e.g. shaky hands, faltering memory). In our view, the kind of agent's relation implied by the unquestioning attitude is clearly another form of *ID*, as long as an agent is dependent on artefacts, parts of artefacts, as well as on their own proper parts in order to bring about certain goals. However, it is not clear whether the unquestioning attitude is a form of trust to the extent to which the former can be applied also to cognitive parts and bodily parts of agents, which do not presuppose that kind of explicit means-end reasoning that, *qua* mental attitude, is at stake in the cognitive theories of trust. Thus we prefer to call this kind of attitude in which agents are involved 'confidence'. Finally, one may say that as long as identity counts as a limit case of parthood, we may call 'self-confidence' that type of agent's unquestioning attitude directed towards themselves.

5. Conclusions

In this paper, a careful comparative analysis of the literature on trust and dependability engineering, followed by the introduction of the instrumental dependence relation (ID) developed through a means-end reasoning perspective, has led us to conclude the equivalence between the notions of 'trustworthiness' and 'dependability', underling the differences from the weaker notion of 'reliability'. Moreover, the analysis has allowed us to individuate a core set of conceptual elements implied by the cognitive account of trust, by arguing in favour of its application w.r.t. both humans and artefacts. Finally, the notion of 'confidence' has been problematized regarding that of trust. Our main concern in the current work has been showing that, from the structural point of view and under many other respects, the relation of *ID* that we have described has only agents in the domain, but both agents and artefacts in the codomain. However, concerning the account of trust *qua* plan that triggers the relation of *ID* with its implications of vulnerability, we are aware that trusting an agent or an artefact for accomplishing a goal are very distinct experiences. For this reason, a working hypothesis for the prosecution of our study is to leverage Brian Epstein's theory of *grounding* and *anchoring* [38]. Though Epstein has developed his theory to account for social facts, the metaphysical mechanisms seem to be generally applicable. In a nutshell, the idea is that the grounding conditions are common

[10] As is well known, there is empirical evidence that artefacts can be integrated into an extended sense of body agency, see e.g. [35,36].

[11] Another notion of trust, which is called 'simple trust', that somehow implies the exclusion of the cognitive aspect of trust, as well as a lack of control on the trustee, has been formulated in [37].

to all instrumental dependence relations, while the anchoring conditions vary depending on whether an agent is instrumentally dependent on other agents or artefacts, as the context, reasons and perceptions involved are very different. To say it more explicitly, for all x, y, z and w, if x is an agent, y is an agent or an artefact, z is an action, w is a goal, and x counts on y to perform z and achieve w (which is x's goal), that grounds the fact that x instrumentally depends on y to achieve w. Thus, all relations of instrumental dependence share the same grounding conditions. On the other hand, for Epstein the anchoring conditions are the facts that put in place the grounding conditions and, in the case of the relations of instrumental dependence, these are different in case such dependence is from another agent or from an artefact, as the reasons for choosing exactly that agent or that artefact are very different, possibly involving the moral status that x ascribes to y, or some power that x has over y etc., which cannot be taken as anchor for relations of dependence on an artefact. This is an interesting future direction of research we deem worth pursuing, which shall be added to the long-term goal of formalizing the theoretical picture presented in this paper, leveraging the foundational ontology DOLCE.

Acknowledgement

The authors acknowledge support by the European project OntoCommons (GA 958371, www.ontocommons.eu). Roberta Ferrario was supported by the BRiO Project, funded by the Italian Ministry of University and Research under the PRIN Scheme (Project no. 2020SSKZ7R).

References

[1] C. McLeod. Trust. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[2] C. T. Nguyen. Trust as an Unquestioning Attitude. In John Hawthorne Tamar Szabó Gendler and Julianne Chung, editors, *Oxford Studies in Epistemology*, volume 7. Oxford studies in epistemology (online edn, oxford academic, 15 dec. 2022) edition, 2022.

[3] P. T. Lewis and S. Marsh. What is it like to trust a rock? a functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, 72:33–49, 2022.

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

[5] A. Ferrario, M. Loi, and E. Viganò. Trust does not need to be human: it is possible to trust medical AI. *Journal of Medical Ethics*, 47(6):437–438, 2021.

[6] M. Chen. Trust and trust-engineering in artificial intelligence research: Theory and praxis. *Philosophy & Technology*, 34(4):1429–1447, 2021.

[7] P. J. Nickel. Design for the value of trust. *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*, pages 551–567, 2015.

[8] T. W. Simpson. What is trust? *Pacific Philosophical Quarterly*, 93(4):550–569, 2012.

[9] D. Gambetta et al. Can we trust trust. *Trust: Making and breaking cooperative relations*, 13(2000):213–237, 2000.

[10] P. J. Nickel, M. Franssen, and P. Kroes. Can we make sense of the notion of trustworthy technology? *Knowledge, Technology & Policy*, 23:429–444, 2010.

[11] C. Castelfranchi and R. Falcone. Trust is much more than subjective probability: Mental components and sources of trust. In *Proceedings of the 33rd annual Hawaii international conference on system sciences*, pages 10–pp. IEEE, 2000.

[12] C. Castelfranchi and R. Falcone. *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons, 2010.

[13] C. Castelfranchi and R. Falcone. Principles of trust for mas: Cognitive anatomy, social importance, and quantification. In *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*, pages 72–79. IEEE, 1998.

[14] P. Dumouchel. Trust as an action. *European Journal of Sociology/Archives européennes de sociologie*, 46(3):417–428, 2005.

[15] G. Amaral, T. Prince Sales, and G. Guizzardi. Ontological foundations for trust dynamics: The case of central bank digital currency ecosystems. In *Research Challenges in Information Science: 16th International Conference, RCIS 2022, Barcelona, Spain, May 17–20, 2022, Proceedings*, pages 354–371. Springer, 2022.

[16] S.S. Bhamare, O.P Yadav, A. Rathore, et al. Evolution of reliability engineering discipline over the last six decades: a comprehensive review. *International Journal of Reliability and Safety*, 1(4):377–410, 2007.

[17] J.H. Saleh and K. Marais. Highlights from the early (and pre-) history of reliability engineering. *Reliability engineering & system safety*, 91(2):249–256, 2006.

[18] A. Coppola. Reliability engineering of electronic equipment a historical perspective. *IEEE Transactions on Reliability*, 33(1):29–35, 1984.

[19] J. Rushby. Critical system properties: Survey and taxonomy. *Reliability Engineering & System Safety*, 43(2):189–219, 1994.

[20] B. Randell. Software dependability: A personal view. In *Proceedings of the 25th International Symposium on Fault-Tolerant Computing (FTCS-25), Pasadena, California, USA, 27-30 June 1995*, pages 35–41. IEEE Computer Society Press, 1995.

[21] P.M. Melliar-Smith and B. Randell. Software reliability: The role of programmed exception handling. In *Proceedings of an ACM conference on Language design for reliable software*, pages 95–100, 1977.

[22] J.L. Laprie, editor. *Dependability: Basic Concepts and Terminology*. Springer Vienna, 1992.

[23] B. Randell. Dependability-a unifying concept. In *Proceedings Computer Security, Dependability, and Assurance: From Needs to Solutions (Cat. No. 98EX358)*, pages 16–25. IEEE, 1998.

[24] M. Al-Kuwaiti, N. Kyriakopoulos, and S. Hussein. A comparative analysis of network dependability, fault-tolerance, reliability, security, and survivability. *IEEE Communications surveys & tutorials*, 11(2):106–124, 2009.

[25] N. Edwards. Building dependable distributed systems. *ANSA, Feb*, 1994.

[26] M. Farrukh Khan and A.P. Raymond. Pragmatic directions in engineering secure dependable systems. In *Advances in Computers*, volume 84, pages 141–167. Elsevier, 2012.

[27] M. Bratman. *Intention, plans, and practical reason*. Harvard University Press, 1987.

[28] L. Biccheri. *Needs as Mental Attitudes. An Ontological Study for PA's Service Design*. PhD thesis, University of Urbino Carlo BO, 2021.

[29] S. Borgo, M. Franssen, P. Garbacz, Y. Kitamura, R. Mizoguchi, and P.E Vermaas. Technical artifacts: An integrated perspective. *Applied Ontology*, 9(3-4):217–235, 2014.

[30] L. Biccheri, R. Ferrario, and D. Porello. Needs and intentionality. In *Formal Ontology in Information Systems*, pages 125–139. IOS Press, 2020.

[31] O. O'Neill. *Autonomy and trust in bioethics*. Cambridge University Press, 2002.

[32] A. Damasio. Brain trust. *Nature*, 435(7042):571–572, 2005.

[33] C. Carabelea, O. Boissier, and A. Florea. Autonomy in multi-agent systems: A classification attempt. In *Agents and Computational Autonomy: Potential, Risks, and Solutions 1*, pages 103–113. Springer, 2004.

[34] C. Castelfranchi. Founding agents'" autonomy" on dependence theory. In *ECAI*, volume 1, pages 353–357, 2000.

[35] T. A Carlson, G. Alvarez, D. Wu, and F.A.J Verstraten. Rapid assimilation of external objects into the body schema. *Psychological science*, 21(7):1000–1005, 2010.

[36] A. Clark and D. Chalmers. The extended mind. *analysis*, 58(1):7–19, 1998.

[37] A. Ferrario, M. Loi, and E. Viganò. In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*, 33:523–539, 2020.

[38] B. Epstein. *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. Oxford University Press, 2015.