Philosophy & Technology (2020) 33:523–539
https://doi.org/10.1007/s13347-019-00378-3

RESEARCH ARTICLE

In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions

Andrea Ferrario[1] · Michele Loi[2] · Eleonora Viganò[2]

Received: 11 June 2019 / Accepted: 10 September 2019 / Published online 23 October 2019

Abstract

Real engines of the artificial intelligence (AI) revolution, machine learning (ML) models, and algorithms are embedded nowadays in many services and products around us. As a society, we argue it is now necessary to transition into a *phronetic* paradigm focused on the ethical dilemmas stemming from the conception and application of AIs to define actionable recommendations as well as normative solutions. However, both academic research and society-driven initiatives are still quite far from clearly defining a solid program of study and intervention. In this contribution, we will focus on selected ethical investigations around AI by proposing an incremental model of trust that can be applied to both human-human and human-AI interactions. Starting with a quick overview of the existing accounts of trust, with special attention to Taddeo's concept of "e-trust," we will discuss all the components of the proposed model and the reasons to trust in human-AI interactions in an example of relevance for business organizations. We end this contribution with an analysis of the epistemic and pragmatic reasons of trust in human-AI interactions and with a discussion of kinds of normativity in trustworthiness of AIs.

Keywords Artificial intelligence (AI) · Trust · Trustworthiness · E-trust

✉ Andrea Ferrario
aferrario@ethz.ch

Michele Loi
michele.loi@uzh.ch

Eleonora Viganò
eleonora.vigano@ibme.uzh.ch

[1] Mobiliar Lab for Analytics at ETH, and Department of Management, Technology, and Economics, ETH Zurich, Zurich, Switzerland

[2] Digital Society Initiative and Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland

1 Introduction

Computer systems (also known as AIs) able to learn from data and generate predictions are increasingly leveraged in a myriad of products and services deployed in domains such as healthcare diagnostics, news broadcasting, criminal justice systems, social media, e-commerce, and job as well as financial markets.[1] AIs "increasingly mediate our social, cultural, economic, and political interactions" (Rahwan et al. 2019), affecting the lives of individuals and shaping organizations and societies as a whole. In the context of this paper, we define AIs as machines that can learn, take decision autonomously, and interact with the environment (Russell and Norvig 2009). They can support human decision-making and perform an ever-increasing number of data and information processing–related tasks. The AIs of interest in this paper consist of (1) complex ensembles of machine learning (ML) models and algorithms to generate predictions from input data and (2) scalable as well as (semi-)automated data processing pipelines; they are embedded as "cognitive engines" in products and services, which, on the other hand, can be referred to as AI-powered solutions.

The constantly decreasing costs of prediction-generating processes have further boosted the diffusion of AI-powered solutions in our societies (Agrawal et al. 2018). Yet, the ethical reflection on the use, effects, and challenges of AIs and an in-depth analysis on the properties, reasons, and dynamics of trust in AI have not kept pace with their technological advancement and diffusion. Trust in human-AI interactions has recently received attention from interdisciplinary communities or researchers (Ribeiro et al. 2016), even though a solid understanding of trust-based relationships involving AIs has yet to be reached. Statements like "Trust is a prerequisite for people and societies to develop, deploy and use AI" (European Commission 2018) are not providing a clear setting for the above discussion. In fact, we argue that the overall level of awareness in society on topics like AI (and considering its opportunities and risks) is still quite low; most users do leverage AI-powered solutions (in the form of a plethora of products and services) without being aware of the presence of AIs, or in the absence of truly accessible alternatives.

In this paper, we discuss the topic of trust and trustworthiness in human-AI interactions by proposing an analysis of trust that overcomes the traditional split of cognitive vs. non-cognitive accounts of trust, as well as the discussion around e-trust, in an incremental model in which we describe *simple*, *reflective*, and *paradigmatic* forms of trust.[2] Compared to alternative and more classical approaches, our proposal accounts for all the phenomena related to trust and thus, it is able to explain consistently both trust in human-human interactions and trust in human-AI interactions. As cornerstone of our model, we choose the concept of "control," i.e., the collection and processing of further information on a potential trustee (i.e., an agent we may trust) to perform an

[1] Some examples of AI-powered solutions: news ranking and social media bots, credit (or risk) scoring, digital financial coach/advisors, digital job assistants, automated insurance claim processing bots, online pricing algorithms, automated financial trading solutions, local policing and recidivism algorithms, autonomous driving and ride-sharing solutions, smart home solutions, online dating platforms, elderly care solutions, and intelligent weapons.

[2] The idea of specifying a paradigmatic form of trust and to distinguish it from degenerate forms is suggested by Bieber Friedemann and Juri Viehoff "A Paradigm-Based Explanation of Trust," unpublished manuscript. Our account of both paradigmatic and non-paradigmatic forms differs from theirs.

assessment of the grounds that make this agent reliable to perform an action to pursue a given goal of relevance for the trustor (i.e., the trusting agent). Roughly speaking, the willingness to rely in the absence of control is what we call simple trust, which is the cognitively less demanding account of trust in our model. On the other hand, reflective trust is having beliefs on the trustworthiness of the trustee, while paradigmatic forms of trust combine both simple and reflective ones.

We start by introducing definitions of AI, ML models, and algorithms, before moving to a quick analysis of salient features of AIs as technological artifacts. We then discuss the concept of trust and e-trust, and the aforementioned incremental model of trust in some detail. We end this contribution with an example of human-AI interactions around an "AI-powered solution case" that is of relevance for any company seeking for performance improvement, cost savings, or both through AI.

2 AI, ML, and Algorithms

In this section, we briefly introduce a minimum amount of notation on AI and ML in order to guide the reader through the paper and then move on to a quick analysis of society-relevant AI, ML, and algorithms peculiarities.

2.1 Definitions

AI is the multidisciplinary endeavor to build machines that can learn, take decisions, and act intelligently in the environment (Russell and Norvig 2009). Decisions can be outputs of a learning process, as well as inputs to generate new ones. Machines are technological artifacts comprised of a combination of software and hardware components. The rising interest on AI in domains like healthcare, retail, marketing, and financial services is due to the ability of AIs to endow products and services with "cognitive functions" through their capability to learn and suggest decisions from digital data. The effectiveness of AIs in delivering performance (e.g., supporting financial growth and cost savings or beating human experts in computer vision tasks) has further boosted their penetration in modern societies. An important component of AI's success is represented by ML, which is the discipline that combines statistical modeling and science of algorithms to create computer systems able to automatically generate predictions and support decision-making by learning inductively from input data (Mitchell 1997; Vapnik 2000). Following Mitchell: "[e]ach ML problem can be precisely defined as the problem of improving some measure of performance P when executing some task T, through some type of training experience E" (Mitchell 1997).

Therefore, given a ML problem at hand (where the task T is, for example, the classification of an email into "spam" or "not spam," the computation of personalized premiums for a given insurance product or the assessment of the risk level of a bank customer), the *statistical model* defines the theoretical structure of its solution. It comprises mathematical constructs describing the task T, the performance measure P to assess results, and the structure of the set of input data encoding experience E. On the other hand, the *algorithm*[3] is the procedure implemented into computer-understandable

[3] More formally, an algorithm is "any well-defined computational procedure that take some value, or set of values, as input and produces some value, or set of values, as output" (Cormen et al. 2001).

language to generate the solution itself, for example, by computing the parameters of the chosen ML model using available input data.[4] This is what technically is referred to as "training the ML model": it is the core process to deploy those AIs which learn inductively on data. In most applications, the result of training a ML model is an object in a given programming language that can be used to generate predictions to support decision-making, once it is fed with new data and embedded in an ad hoc IT architecture. This dynamic infrastructure is a key component of the design of AI-powered products and services; depending on their technical complexity (e.g., the number of customers they are supposed to reach, the number of transactions per second) and the structure of the organization promoting them (start-up vs. well-established organization), such infrastructure can be fully cloud-based, hybrid, or developed and managed in-house. AIs typically use multiple ML models, algorithms, and automated data processing pipelines to generate predictions[5]: on the other hand, communication with end users is driven by interfaces that can take the form of apps, web-based applications, and, more recently, even augmented reality devices.

2.2 Peculiarities of AI

In the current literature on ethics of AI, interpretability, transparency, fairness, and reliability are among the most discussed *desiderata*, as they provide a basis to better understand AIs and their outcomes as well as a support for the contestability of AI-generated decisions by end users and society (Lipton 2016; Miller 2017; Doshi-Velez and Kim 2017). The need to discuss such *desiderata* stems from two features that distinguish AIs from other technological artifacts: *inscrutability* and *nonintuitiveness* (Selbst and Barocas 2018). Inscrutability suggests that "models available for direct inspection may defy understanding," while nonintuitiveness "suggests that even where models are understandable, they may rest on apparent statistical relationships that defy intuition" (Selbst and Barocas 2018). Although one could be skeptical of whether end users of AI-powered solutions are really interested in the possible mitigation of inscrutability[6] (which is relevant for AI architects, instead), nonintuitiveness seems to be of high relevance for society due to the consequences stemming from the generation of personalized outcomes based on nonintuitive—and potential spurious—correlations in data. In addition to inscrutability and nonintuitiveness, one has also to consider that in most AI-powered solutions, the underlying ML models are updated (i.e., retrained) with frequencies that depend on the domain of applicability of the solution itself.[7] This implies that an AI potentially generates different outcomes for the same customer, patient, convict, job-seeking individual, depending on the moment at which the outcome is generated: any

[4] This is the pragmatic definition of learning proposed by Vapnik (2000).

[5] In this contribution, we will consider only those AIs whose goal is to generate predictions to support decision-making. Other type of intelligent machines, e.g., those devoted to data-driven support solutions like information retrieval and enterprise knowledge management, data extraction, transformation, load, and automated processing are out of scope.

[6] It is a matter of debate between legal experts to which extent the GDPR—considering both the legal text and the non-binding recitals—involves a right to an explanation of algorithmic decisions, either in the form of an *ex ante* logic of the decision-making or of an *ex post* explanation of individual ones (Wachter , Mittelstadt, and Floridi 2017).

[7] For example, in case of online learning, ML models are continuously retrained, i.e., training occurs at every arrival of a new data point, or batches of them.

explanation or justification of this outcome (e.g., for the purpose of contesting or auditing it) depends on time, as well. This intrinsic dynamism is what characterizes AI-powered solutions in which ML models and algorithms learn on ever changing input data; it poses a challenge to the identifiability of AIs. The inherent difficulties in trusting an artifact that changes itself is a crucial point deserving more exploration. Arguably, no program that learns can be as trustworthy as a program that does not learn, unless the learning is sufficiently controlled. A full exploration of this aspect falls, however, outside the scope of this paper.

3 Trust, e-trust, and an Incremental Model of Trust

3.1 What Is Trust?

People use the expression trust in a manifold of different ways to describe a variety of human affairs. People talk easily about trusting their friends, peers, or even strangers; people trust their own intuitions or themselves; they trust science and possibly the scientific community. Sometimes they (even) trust politicians or institutions; in certain cases, they can also express a trusting attitude towards non-human agents. Trust is thus a construct enriching a wide variety of relationships people establish, nurture, and interrupt in everyday life.

When talking about trust, people refer to specific goals and contexts: for example, I trust my friend's competence as a scholar versed in microbiology, but I would never trust him to post my letter. Similarly, I trust my accountant to accurately file my tax return respecting the fixed deadline, but not as a political advisor. Therefore, people do not trust each other in every possible way, but assess the competence of others in a specific context, with respect to a predefined goal they care about. Sometimes, they express trust in someone or something, sometimes they simply trust someone or something, and sometimes they even trust that something is the case.[8] In most of our experiences, trust is interpersonal and "face-to-face" or it is mediated by technology, as in the case of digital environments and the phenomenon of e-trust (Taddeo 2009). Therefore, trust is a relevant construct to be considered in everyday life, at different levels; however, "there is not yet a shared or prevailing, and clear and convincing notion of trust" (Castelfranchi and Falcone 2010).

The literature on trust comprises different research domains like philosophy (Gambetta 1988; Baier 1986; Holton 1994; Pettit 1995), sociology and psychology (Luhmann 1979; Deutsch 1962), organizational science and management (McEvily and Tortoriello 2011; McAllister 1995; Mayer and Davis 1999), and economics (Dasgupta 1988) as well as marketing (Komiak and Benbasat 2006; McKnight et al. 2002). These

[8] Philosophers have long debated whether the primitive concept of trust involves three arguments ("A trusts B to X") or two ("A trust B") (see Faulkner and Simpson 2017 for a defense of two-argument trust). We assume the more common three-argument analysis of trust. Notice that this is logically compatible with A trusting B to make choices that are in A's best interest, in general, i.e., for a large range of Xs. In the case of simple trust, that implies that A relies on B for a large range of different Xs without controlling B. In the case of reflective trust that implies that A considers B to have properties (e.g., a high level of competence when dealing with practical problems and a general motivation to further A's interest) that are objective reasons for A to rely on B, for a large range of different Xs, without controlling B.

accounts of trust intercept selected properties of the agents involved in the trust engagement, their intrinsic motivations, and the dynamics of the trusting process itself. This fragmentation leads to a context-specific plethora of definitions around trust and trustworthiness, where the latter is the concept encapsulating the characterization of the properties of the agent endowed with trust.

A significant division in the literature of trust is between those definitions that emphasize the cognitive aspects of trust and those that emphasize the non-cognitive aspects. Non-cognitive theories define trust in terms that do not essentially involve beliefs about, or assessments of, trustworthiness or reliability. We summarize all considerations in the taxonomy of Table 1, which should allow the reader to grasp the inherent complexity that is faced when dealing with trust and its accounts.

3.2 The Case of e-trust

E-trust is an interesting approach to trust in digital environments and in the presence of artificial agents (of which AIs are part of). E-trust is "trust applied to digital contexts and/or involving artificial agents" (Taddeo and Floridi 2011) and occurs "in environments where direct and physical contacts do not take place, where moral and social pressures can be differently perceived, and where interactions are mediated by digital devices" (Taddeo 2009). Therefore, e-trust becomes relevant in the presence of interactions with electronic commerce platforms, group chats and online communities, technology-mediated self-services, multi-agent systems contexts, and, in general, whenever humans or artificial agents, or both interact in a digitally mediated environment. As starting point, an account of e-trust is necessarily required to define trust, since the latter is the reference point, and to tackle the question whether e-trust is an occurrence of trust or an independent phenomenon. As the existence of a shared and institutional background and the certainty about the trustee's identity are usually identified as necessary conditions to develop any trust theory (Taddeo and Floridi 2011), detractors of e-trust discuss the impossibility of having trust in digital environments due to the absence of such conditions (Pettit 1995; Nissenbaum 2001).

Defenders of e-trust argue that e-trust is suitable for any analysis involving artificial agents (AAs), which are entities endowed with the properties of interactivity, autonomy, and adaptability (Floridi and Sanders 2004), once an appropriate level of abstraction (Floridi and Sanders 2003) has been introduced to describe a system of interest, its dynamics, and context. AIs are examples of AAs, where interactivity, autonomy, and adaptability properties rely on the ML algorithms embedded in the AI itself. Not only AAs communicate with

Table 1 Taxonomy of trust accounts

Reliability/trustworthiness belief (assessment, etc.)	Yes	Cognitive theory	Rational assessment	Yes	Rationalist theory
				No	Non-rationalist theory
	No	Non-cognitive theory	Non-cognitive mental states (e.g., conscious intentions to rely on the trustee or to cooperate)		
			Behaviors or behavioral dispositions (e.g., dispositions to cooperate in specific circumstances)		

human agents in digital environments, but also with other artificial agents in distributed systems (multi-agent systems). In the case of AAs[9] in distributed systems, Taddeo (2010) introduces e-trust as a second-order property of first-order relations where a trustor "rationally selects the trustee on the basis of its trustworthiness" (Taddeo 2010), and this latter is a value computed from the outcomes of past performances. Trustworthiness is a "guarantee required by the trustor that the trustee will act as it is expected to do without any supervision" (Taddeo 2010); this implies that for AAs in distributed system, the emergence of e-trust is advantageous and efficient: in its presence, trustors avoid to perform a given action and to supervise the trustees' performances. On the other hand, in a context involving both human and artificial agents, Taddeo (2010) identifies the necessity to introduce trust and e-trust at an appropriate level of abstraction, due to the complexity of the conditions leading to their emergence. e-trust is again a second-order property of first-order relations, where a trustor "chooses to achieve its goal by the action performed by the trustee" and "considers the trustee a trustworthy agent" (Taddeo 2010).[10]

Other authors engaged themselves in defining accounts of trust in digital environments; for example, Castelfranchi and Falcone (1998) argued that e-trust is based on a "threshold value that is the result of a function of the subjective certainty representing the subjective certainty of the beliefs held by an artificial agent." Clearly, the attribution of mental states, feeling, or emotions to artificial agents is controversial and rejected by most of authors (Taddeo 2010). On the other hand, Tuomela and Hofmann (2003) ground e-trust in some ethical principles stemming from rational social normative trust theories. Finally, we mention that Grodzinsky et al. (2011) introduce a notion of trust—denoted by TRUST—comprising both trust and e-trust and defined as a "decision by a to delegate to b some aspect of importance to a in achieving a goal" (Grodzinsky et al. 2011), whenever the artificial agents are endowed with the ability to take decisions. In the model of trust deriving from TRUST, two aspects are isolated: the mode of communication (via physical proximity or telecommunications) and the type of entities that constitute the two parties in a trust interaction (human or artificial agents). By combining the two aspects, eight kinds or subclasses of trust are derived (Grodzinsky et al. 2010).[11] We argue that the dependency of e-trust on the notion of trust and the necessity of establishing its functional relationship with it, as well as the presence of different accounts of e-trust indicate the need of alternative conceptualizations of trust in human-AI interactions.

[9] Taddeo (2010) considers rational artificial agents by design, as her analysis "rests on a Kantian regulative ideal of a rational agent, able to choose the best option for itself, given a specific scenario and a goal to achieve" (Taddeo 2010).

[10] Taddeo's definition of e-trust in a mixed context (e.g., where human and artificial agents interact) is analogous to our definition of reflective trust (see Section 3.5). For Taddeo, too, trustworthiness is the object of a belief entertained by a (human) trustor; in Taddeo (2010), Section 5, it is, too, a matter of reliability and delegation in the absence of control during the performance of the trustee, which Taddeo calls "supervision." However, Taddeo's definition, unlike ours, requires a guarantee of reliability. That is implausible, in our view. We think that in most cases, the expectation that the trustee will act as expected is not supported by a guarantee. Rather, the trustor merely believes, with a confidence (which does not have to be certainty), that the trustee will act reliably enough without supervision, for reliance to have a positive payoff to the trustor (in Nickel's (2009) classification, this makes our account a "staking" account, while Taddeo's is a "predictive" account). Moreover, in the absence of such a guarantee, a trustor may instinctively trust without forming any trustworthiness belief. This is why we introduce simple trust in Section 3.4.

[11] Although the binary aspects of a trust relationship are two, the resulting kinds of trust are eight and not four because the entity aspect occurs twice, as the parties of a trust relationship are two.

Buechner and Tavani (2011) try to avoid this problem by analyzing trust between artificial agents in terms of "diffuse trust," which is trust in a system or environment. For example, when using an airline, we expect "reliable, courteous, and orderly service" (Walker 2006), by "whatever individuals are filling organizational roles to perform effectively to the end we rely on" (Walker 2006). This is a moralized account of trust, which we consider problematic for e-trust.[12] Ascribing moral obligations to unspecified individuals or to a system of interacting agents as a whole is however notoriously problematic (Smiley 2017), so a non-moralized account has distinctive advantages in this domain.

3.3 An Incremental Model of Trust: Definition

We now propose a model of trust, which takes care of both cognitive and non-cognitive accounts *incrementally* and in a finite sequence of steps.[13] The model explains the many ways in which people—both ordinary people and scholars—talk about trust, as well as the relation between trust and trustworthiness. It can be applied to relationships between humans and artificial agents, with focus on those artificial agents endowed with cognitive capabilities stemming from ML algorithms. Our model of trust T consists of the triple

$$\mathrm{T} = (\textit{simple trust}, \textit{reflective trust}, \textit{paradigmatic trust}),$$

whose elements are constructed on the 5-tuple (X, Y, A, G, C), where X and Y denote interacting agents and A the action to be performed by the agent Y to achieve a goal G of relevance for X in a given context C. For simplicity of exposition, this latter is kept fixed in the forthcoming discussions. The remainder of this section is devoted to the discussion of the triple T.

3.4 Simple Trust

The first element of our model T—called simple trust—is a non-cognitive account of trust based on the concept of reliance. We say that X simply trusts Y $=_{def}$ X is willing to rely on Y to perform an action A pursuing a goal G, and X plans to rely on Y without intentionally generating and/or processing further information about Y's capabilities to achieve G. In case of simple trust, by definition, the agent X shows a mental attitude or predisposition comprising the intention (i.e., willingness) to depend on an agent Y

[12] A moralized account of trustworthiness, derived from Nickel's "moral expectation account" of trust qualifies Y as trustworthy to do A for X if and only if Y has properties that give X good reasons to morally expect Y to do A (Nickel 2009). This is unreasonable if Y is an AI since, arguably, current AI cannot be *morally obliged* to do anything, due to its lack of capacity as a moral agent (Talbot et al. 2017). Since X's moral expectations towards existing AIs are always unreasonable, Nickel's moralized account implies that existing AIs cannot be reflectively trusted. However, we notice that this is compatible with arguing that developers of AI have moral obligations related to the expectations of agents placing trust in AI. This opens up the possibility for moralized trust definitions in the AI domain, following for instance Nickel's diffusion of trust in engineering framework (Nickel 2019).

[13] This model is first presented and defended in more detail in [Loi et al.] "the logic of trust," unpublished manuscript.

without performing a further assessment of the grounds that make Y reliable to perform the action A to pursue a goal G. For example, in the case of trusting a human, the grounds that make Y reliable could be Y's capabilities and motivations, so a trustor X who trusts a trustee Y intends to depend on Y without the intention to further control Y's capabilities and motivations, from the point in time at which X decides to trust. Our definition of simple trust is in line with Castelfranchi and Falcone's (2010) layered notion of trust, where trust in its basic sense is considered as an attitude or disposition towards and agent Y. However, the authors describe the attitude as a cognitive one, i.e., the beliefs of expectation and evaluation. On the contrary, simple trust is non-cognitive and does not presuppose any specific representation of an attitude, property or inclination of Y. Notice however, that simple trust describes trust as property of X, an intention of the trustor with respect to Y. In the definition of "simple trust," we stipulate that "further" means additional information in the case in which some information about Y has already been collected and processed, but it can also refer to cases in which X has not yet generated any information and plans to generate some. So "further" excludes information that X already has.

Lack of control on Y (=$_{def}$ the generation or processing of further information about Y's reliability) can be due to either contingent or structural causes. Contingent causes can be ascribed to lack of space and time capacity: a typical example is time pressure conditions under which resources in corporations take decisions by trusting new co-workers or employees in performing a certain task through a form of non-cognitive delegation. Structural causes can be exemplified by the level of maturity of a child's cognitive capability when he interacts with people around him, instead. The way most people speak about small children suggests that they can trust. For example, it is often said that children instinctively trust their mothers, or they tend to trust all adults. It seems problematic to ascribe to children a sophisticated belief about trustworthiness, though. Other structural causes for the lack of control are represented by social as well as cultural barriers between X and Y, strongly limiting X's capabilities to plan a meaningful activity of control of the agent X relies on. For example, X—a tourist in a foreign country—trusts Y in showing him the right path to the hotel, although he does not speak Y's language and relies on his hand gestures only to (simply) trust him. Notice that simple trust does not require a belief in Y's trustworthiness, but it is compatible with both not having this belief (what we shall call, deviant simple trust), and with having previously acquired it (in what we shall call paradigmatic trust). For example, due to the existence of a certain number of past positive interactions between X and Y, X has collected information that she considers sufficient for judging Y trustworthy and, as a result, X could form the intention to rely on Y without control.

3.5 Trustworthiness and Reflective Trust

Reflective trust gives satisfaction to theories stressing the importance of cognitive aspects of trust. Reflective trust essentially involves the cognitive belief in X as trustworthy. Given these premises, and based on the definition of simple trust, we can move to the second element of the model T: trustworthiness. In line with the philosophical literature, we introduce trustworthiness as a property of the trustee, Y. In particular, we say that Y is trustworthy for X to do A for the sake of G = $_{def}$ Y has properties that are objective reasons for X to simply trust Y to perform A to reach the

goal G. This is equivalent to state that Y has properties that are objective reasons for X to rely on Y in the absence of control[14] (here and in what follows, we omit action and goal for brevity's sake). Trustworthiness is therefore the condition of being worthy of simple trust; a trustworthy entity is one that there are good reasons to rely on, even when it is not controlled. Our definition does not depend on specific commitments or motivations of the trustee: the "good reasons" of X in considering Y as trustworthy can in fact encapsulate a variety of considerations.

3.6 Reflective and Paradigmatic Trust

Having defined trustworthiness, we can finally move to the definition of reflective trust, which simply describes the condition of having a belief about the trustworthiness of an entity. More precisely, X reflectively trusts Y to perform A to achieve the goal G = $_{def}$ X believes Y to be trustworthy to perform A for the goal G. Notice that, in reflective trust, X believes Y to be trustworthy, that is, X believes that Y has certain objective features. Reflective trust, as we define it, is not by definition well-supported trust: although the properties X believes Y to have are conceived by X as good reasons to trust Y and X's belief that Y has such properties may be false. The measure of reflective trust is the confidence (i.e., subjective probability) with which the trustworthiness belief is held.[15] The following is a deviant form of reflective trust: given a trustworthy Y and X nonetheless plans and executes control, e.g., X has a conscious belief that Y is trustworthy but due to unconscious causes, he does not trust it. An example could be represented by two co-workers from countries that have been in war with each other and have developed ethnically motivated hatred; each worker concludes he has most reasons to regard the other colleague as trustworthy, based on past work interaction, but none is able to overcome the instinctual suspicion due to the those psychological-historical factors. Other causes of deviant reflective trust can be unconscious sexism and racism. If X has both simple and reflective trust in Y, i.e., X believes that Y is trustworthy and he is willing to rely on Y without control, then we say that X exercises paradigmatic trust, which is —according to our incremental model—the most complete form of trust we propose. Arguably, most people have in mind paradigmatic trust when talking about trust. The two deviant forms of trust highlighted in Table 2 are clearly conceptually possible and we suggest that they are also psychologically possible—this is the psychological hypothesis we propose.

We avoid discussing accounts of trust between artificial agents in detail here; as this is a less familiar problem for our readers, it is best to leave it to a distinct article. Our approach is compatible with Taddeo's idea that "the analysis of e-trust among AAs endorses a non-psychological approach and rests on a Kantian regulative ideal of a rational agent, able to choose the best option for itself, given a specific scenario and a goal to achieve" (Taddeo 2010). We therefore agree with Taddeo that an AA may operate based on a representation of the trustworthiness of another AA calculated in a rational way, based on past interactions. In other words, most cases of mutual trust

[14] Thus, every Y that is trustworthy according to Taddeo's (2010) definition is trustworthy according to our definition, but the reverse does not hold, in general. Y can be trustworthy in the absence of a guarantee of its reliability without control.

[15] This measurement is clearly linked to Gambetta's definition of trust as a "particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action" (Gambetta 1998).

Table 2 Summary of the different types of trust

		X believes that Y is trustworthy	
		Yes	No
X is willing to rely on Y without control	Yes	Paradigmatic trust	Deviant simple trust
	No	Deviant reflective trust	No trust

between AAs would be paradigmatic in terms of our incremental model: they would result from a combination of reflective trust (a cognitive representation by the trustor-AA of the trustee-AA's trustworthiness) and simple trust (the programmed disposition of the trustor-AA to rely without control on the trustee-AI deemed trustworthy).[16] Like Taddeo's own approach, our approach has the advantage of using a non-moralized account of trust, which is more easily applicable to AA-AA relations.

3.7 Incremental Model of Trust in Human-AI Interactions: a Simple Example

Let us now move to a simple example of human-AI interactions. To do so, let us imagine a proactive company pushing for AI-powered solutions to generate business performance through innovation and new technologies. The top management of the company recently decided to approve the design and test launch of a new product for a selected (and high priority) portfolios of customers. At the core of the product lies a two-staged cognitive engine, i.e., an AI comprising of multiple ML algorithms and automated data processes. The business decisions are taken based on both machine-generated predictions and human expertise (for example, considering a nested IF/ELSE logic structure). Health insurance solutions collecting wearables data, telematics products, or personalized marketing campaigns based on credit card usage are all examples of the above. The conception, design, development, testing, and deployment of the solution is a highly complex endeavor, involving a mix of domain expertise and technical capabilities in product development, data science, IT engineering, and project management. For this means, the company decides to hire a team of seasoned consultants to support in-house resources (*in primis* the project sponsor and his manager) to finalize and launch the AI-powered solution.

Which type of trust can we identify in the human-human and human-AI interactions occurring in the conception, design, development, testing, and launch phases of the AI-based solution, in this hypothetical case? Let us start with human-human interactions. We argue that it is reasonable to imagine that some, most or even all professionals in the team of externals (e.g., data scientists, data engineers, or experts in data-based product management) paradigmatically trust colleagues in performing certain routines to achieve a goal (e.g., designing the first blueprint of the AI or collect business requirements). This is because they believe such colleagues to be trustworthy, based on a

[16] The main difference between our approach and Taddeo's one concerns the way in which the trustor-AA should assess the trustworthiness of the trustee-AA. We argue that an analysis of trustworthiness should consider not only the probability of successful actions, but also the assessment of (1) the benefit for the trustor deriving from the successful action of the trustee, (2) the harm following from non-successful actions, and (3) the cost for the trustor of controlling the performance of the trustee.

sequence of past, successful interactions, as they have been working together on similar engagements for quite some time.

Typically, the company project manager, Alyssa, paradigmatically trusts the manager of the team of consultants, Bob (e.g., in delivering a high-level description of a new set of solution features for a catch-up meeting) due to his reputation, built on a track record of positive performances in that type of activities. In typical cases, Bob's reputation is the basis of Alyssa's *weak* trustworthiness belief. However, it is reasonable to expect that, due to the nature of the business relationship between the company and the consulting society, a certain amount of time and energy to accurately evaluate the capabilities and motivations of all professionals involved in the project has been performed by Alyssa's company in the past. Past control is not incompatible with paradigmatic trust, but in order for paradigmatic trust to be achieved for a given goal, Alyssa should cease control of Bob's team in that given instance. After repeated interactions, in case of a positive outcome, Alyssa considers some (or all) of Bob's consultants as trustworthy, and thus paradigmatically trust them. By contrast, Bob may not have a confident belief in the capacity of a young data scientist, Charles, to design a first example of predictive model for the AI solution and convince the client. In fact, Charles has been recommended by Emily, who is Bob's executive director, and in the past she did not always recommend the right resources to the consultancy company's engagements. Therefore, Bob may rely on Charles only in the presence of a significant amount of control (distrust). If Charles proves unable to execute the task assigned to it, Bob may eventually form a belief that there are good reasons not to simply trust Charles (reflective distrust). Let us now turn our attention to some examples of human-AI-powered solution trust interactions in our scenario of AI-based solutions development. We argue that a senior data scientist, called Diana, in the team of consultants responsible for the training, testing, and deployment of the ML algorithms in the cognitive engine of the solution, can manifest paradigmatic trust in the solution itself to achieve the goal G to generate personalized predictions reliably. This is due to her understanding of the structure and behavior of the cognitive engine she contributed to design; in fact, these are reasons to consider the solution as trustworthy and to will to rely on it without further collecting information (e.g., through additional testing). In this specific case, she can entertain a rational belief on the AI trustworthiness.[17] In the case of some machine learning generated models, in particular *inscrutable* ones, the senior data scientist Diana may have very little trust in the model: she may not be confident that the model should be employed without investing some resources to monitor its performance and, consequently, she would invest resources for monitoring. On the other hand, we argue that Diana could manifest only simple trust (without reflective trust) towards the ability of the AI solution to achieve the goal G, with G being the company overall business goals and not the specific machine learning problem at hand. She may not have the possibility to entertain a belief, sound or not, on the capability of the solution to do something good for the company, due to a lack of domain knowledge and in-depth understanding of all business specificities of the company sponsoring the project. Due to social motivations and constraints, including the need to do business, Diana may have only deviant simple trust in the AI's ability to contribute to the

[17] We are not implying that paradigmatic trust needs rationality of the trustworthiness belief; in fact, paradigmatic trust can also be irrational.

company's goals. This example involving the senior data scientist is paradigmatic as it shows that a manifold of different trust-relations (according to our model) are developed in a given human-AI interaction. They depend, among others, on the relationship between multiple dimensions of expertise (data science, domain knowledge) of the human agent, the specific goal the AI is supposed to achieve through its functioning, and on the context in which AI is implemented (e.g., a context in which it is very costly or there is no time to gather further information on the machine). In the case of paradigmatic trust, the senior data scientist is able to entertain a belief on the AI's trustworthiness in reaching a goal, that is completely understandable using the methods of data science; this is not true in the deviant case of simple trust, as she does not have the background and the capabilities to grasp the nuances of the business problem at hand, and it is not even part of her role to try to gain such understanding. Finally, we need to consider the perspective of an end user of AI-powered solutions. Typically, end users have a rather low level of data science expertise and variable levels of domain knowledge. We consider as costly for end users to improve both expertise dimensions. We illustrate the psychological hypothesis that some end users manifest deviant simple trust towards the AI-based solution to achieve both goals or reliable prediction generation and encapsulation of company's values and objectives. Simple trust can be coupled with a weak trustworthiness belief (reflective trust) and be an instance of weak paradigmatic trust. Such belief may rely on the reputation of the company's brand and services, for example.

3.8 Reasons of Trust in Human-AI Interactions and Kinds of Normativity in Trustworthiness

As seen in our incremental model, the *good* reasons for X to trust Y—and which render Y trustworthy—are various and their validity depends on the context of the trust relationship. These reasons can be epistemic (related to Y's trustworthiness) or pragmatic (related to X's need to trust some Y). Both epistemic and pragmatic reasons can be *objective*, when they are *facts*, or *subjective*, if they are relative to X's beliefs about Y. In the case of AI, the properties of an AI system that make it trustworthy for achieving a goal are objective reasons to trust that system; for instance, the accuracy with which the AI-powered solution computes personalized outcomes of customers in the selected portfolio.[18] Subjective reasons to trust such an AI-powered solution are, for instance, the properties that the managers of the company think that make it trustworthy, e.g., achieving the business goal faster or with lower costs.

The kind of pragmatic reasons that are relevant to AI are *prudential reasons*, namely reasons that are grounded in the fact that trusting Y increases the trustor's well-being. In our example, the company launching a new AI-powered solution may trust the latter as this increases the profits, which can be approximated to the company's interest. Similarly, a police department may have reason to trust an AI system such as software of face recognition because the latter frees police employees for more valuable tasks, thus increasing the available resources of the department. The most common subtype of prudential reason that is involved in the human-AI interaction is the cost of controlling the performance of an AI-powered solution, its functioning, features, and the normative

[18] In Nickel's (Nickel 2009) vocabulary, these are the objective reasons of X's *predictive* expectations.

principles that supported its design. The end user of an AI-powered solution may trust the latter without looking for further information on it because the search for information is a costly activity that decreases the end user's well-being (e.g., the individual should invest time and energy in studying computer science and applied mathematics). This is an incentive to simply trust the AI-powered solution and the data scientists who designed it, as previously remarked.

Epistemic reasons regard the trustor's belief in the trustee's trustworthiness. They are based on the strength of evidence for and against the trustee's trustworthiness. In AI, the strength of evidence of trustworthiness regarding properties of AI systems decreases with the decreasing expertise of the agent interacting with the systems. In fact, the collection and analysis of such evidence is easier for the data scientist that designed the AI system than for the end user with no data science training, as we discussed above. Notice that the same properties may be both pragmatic reasons for simple trust and epistemic reasons for reflective trust. For example, the fact that an AI is the cheapest solution in the market to achieve a goal G and the only solution X can afford and the fact that X is not able to control the AI in a meaningful way can be both a pragmatic reason for X to simply trust the AI, and a reason that X reflectively endorses as being a good reason to simply trust the AI (reflective trust) in her specific circumstances.[19] There is no contradiction in X reaching the conclusion that she has to trust the AI, given the limited options at her disposal, while wishing that she would not be in a condition in which she had most reasons to trust the AI. A summary of reasons in human-AI interactions is provided in Table 3, for the sake of readability. In our account, the reason that makes somebody or something trustworthy determines the scope of trustworthiness (absolute or relative). If X's reason to consider Y trustworthy is the fact X has no better alternative, then Y is *relatively* trustworthy. Relative trustworthiness depends on the context of the trust relationship and thus Y has this property from X's perspective, but not from another person's perspective. By contrast, if the reason that makes Y trustworthy makes it trustworthy for all, e.g., its competence in a field relevant to achieve G, then Y is *absolutely* trustworthy. This means that, in different contexts, with the same G, Y remains nonetheless trustworthy and is such from everyone's perspective.

For example, an unknown baby sitter can be (relatively) trustworthy for a debt-ridden family who desperately needs someone to take care of the child; a good baby sitter is trustworthy (there are reasons to simply trust her) for all, even the (super) affluent households.[20] Similarly, a cheap, inaccurate AI is relatively trustworthy for a company that needs to cut costs and has no better alternative. It is not trustworthy *absolutely*, e.g., a company that can afford the same job to be done by a team of more accurate humans does not have reasons to use it. An absolutely trustworthy AI would be one that even resourceful companies, that can access to a wide range of options, have reasons to trust.

For relatively trustworthy AI, there are cases in which agents with a limited option set end up paradigmatically trusting AIs—including correctly believing that the AIs are trustworthy *for them*—due to lack of feasible alternatives. Hence, no

[19] These are the objective reasons in support of X's "*staking expectations,*" consisting in X's "*judgment that it is worth staking something of value on the [...] action of*" Y (Nickel 2009).

[20] Our account of absolutely trustworthy AI avoids Nickel's babysitter's counterexample to non-moralized accounts of trust (in which we do not blame a baby-sitter chosen out of necessity), without introducing moral elements in the definition of trust and trustworthiness (Nickel 2009).

Table 3 Summary of reasons to trust in human-AI interactions

Reasons of trust in AI	Objective	Subjective
Pragmatic reasons for simple trust: X simply trusting Y increases X's well-being	It is a fact that X simply trusting Y increases X's well-being	X's beliefs about Y imply that X simply trusting Y increases X's well-being
Epistemic reasons for reflective trust: it is rational for X to believe that Y has features that make it rational for X to simply trust Y	It is a fact that Y has features that make it rational for X to simply trust Y	X believes that Y has features that make it rational for X to simply trust Y

inference can be made from the fact that AIs are trustworthy for specific agents to their being good AIs in general. An agent may even reasonably feel guilty for ending up in a condition in which it is in fact rational for her to paradigmatically trust an AI that, had her option set been better, she would never had trusted. Thus, *relatively* trustworthy AI is not a meaningful moral goal. A more meaningful moral goal is often to improve the alternatives available to users, even if that means that they no longer have reasons to trust those AIs that they *now* have reasons to trust for lack of better alternatives.

4 Conclusions

AI is a phenomenon affecting individuals and their lives, organizations, and societies as a whole. The ability to perform complex tasks and support decision-making thanks to ensembles of ML models and algorithms prima facie supports the adoption of AI in multiple domains. Therefore, it is necessary to discuss the nature and dynamics of trust in the presence of human-AI interactions, with focus on the properties of trustworthy AI. In this paper, we have introduced an incremental model of trust that can be applied to human-human and human-AI interactions combining both cognitive and non-cognitive accounts of trust. We have then applied the incremental model to an example of human-AI interactions of relevance for business organizations, where the different layers of the incremental models, in case of both human-human and human-AI interactions, are discussed in detail, highlighting the differences between AI designers (e.g., data scientists) and end users in the trusting relationships. We finished our analysis by describing the nature (epistemic or pragmatic) and level of objectivity (subjective or objective) of the reasons to trust AIs. Finally, we observed that AI's trustworthiness, can be a relative or absolute construct. Relative trustworthiness is not a meaningful normative goal of AIs development. Those using the concept of trustworthy AI to indicate a moral goal or objective should carefully define what they mean by trustworthiness.

Acknowledgments Michele Loi and Eleonora Viganò acknowledge the support of the European Union's Horizon 2020 Research and Innovation Programme under agreement no. 700540 and of the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0052-1 (CANVAS: Constructing an Alliance for Value-driven Cyber-Security). Eleonora Viganò acknowledges the support of the Cogito Foundation's project 17-117-S "Intrapersonal conflicts of values: scientific and philosophical

understanding of the present self's concern forher future self." The authors thank theanonymous referees whose comments helped improve the manuscript.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Boston: HBR Books.

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

Buechner, J., & Tavani, H. T. (2011). Trust and multi-agent systems: applying the "diffuse, default model" of trust to experiments involving artificial agents. *Journal Ethics and Information Technology, 13*(1), 39–51.

Castelfranchi, C., Falcone, R. 1998. Principles of trust for MAS: cognitive anatomy, social importance, and quantification paper presented at the proceedings of the third international conference on multi-agent systems.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: a socio-cognitive and computational model*. Hoboken: John Wiley and Sons, Ltd.

Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (2001). *Introduction to algorithms* (2nd ed.). MIT Press.

Dasgupta, P. (1988). Trust as a commodity, in Gambetta (ed.) 1988.

Deutsch, M. (1962). Cooperation and trust: some theoretical notes. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Lincoln: Nebraska University Press.

Doshi-Velez, F., Kim, B. (2017). Towards a rigorous science of interpretable machine learning. Available: https://arxiv.org/abs/1702.08608 (accessed 20.05.19).

European Commission, High-Level Expert Group on Artificial Intelligence. (2018). *Ethics guidelines for trustworthy AI - working document for stakeholders' consultation*. Available at https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai. (accessed 20.05.19).

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

Floridi, L., Sanders, J. W. 2003. The method of abstraction, in M. Negrotti, ed., The yearbook of the artificial. Issue II, Peter Lang, Bern.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *J. Minds and Machines, 14*(3), 349–379.

Gambetta, D. (1998). *Trust: making and breaking cooperative relations*. New York: Basil Blackwell.

Grodzinsky, F. S., Miller, K., & Wolf, M. J. (2010). Towards a model of trust and e-trust processes using object oriented methodologies. In *Ethicomp 2010*. Ed. Arias-Oliva: Bynaum, Rogerson, Torres-Coronas.

Grodzinsky, F. S., Miller, K., & Wolf, M. J. (2011). Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?". *Journal of Ethics and Information Technology, 13*(1), 17–27.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72*(1), 63–76.

Komiak, S., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly., 30*, 941–960. https://doi.org/10.2307/25148760.

Lipton, Z. C. (2016). The mythos of model interpretability. Available at https://arxiv.org/pdf/1606.03490.pdf (accessed 18.01.19).

Luhmann, N. (1979). *Trust and power. Two works*. Chichester: Wiley.

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: a field quasi-experiment. *Journal of Applied Psychology, 84*, 123–136.

McAllister, D. J. (1995). Affect and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal, 48*, 874–888.

McEvily, B., & Tortoriello, M. (2011). Measuring trust in organisational research: review and recommendations. *Journal of Trust Research, 1*(1), 23–63.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: an integrative typology. *Info. Sys. Research, 13*, 334–359.

Miller, T. (2017). Explainable AI: Insights from the social sciences. Available at https://arxiv.org/abs/1706.07269 (accessed 20.05.19).

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

Nickel, P. J. (2009). Trust, staking, and expectations. *Journal for the Theory of Social Behaviour, 39*(3), 345–362.

Nickel, P. J. (2019). Trust in engineering. In D. P. Michelfelder & N. Doorn (Eds.), *Routledge Companion to Philosophy of Engineering (Routledge, forthcoming)*.

Nissenbaum, H. (2001). Securing trust online: wisdom or oxymoron. *Boston University Law Review, 81*(3), 635–664.
Pettit, P. (1995). The cunning of trust. *Philosophy and Public Affairs, 24*(3), 202–225.
Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. *Nature, 568*(7753), 477.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': explaining the predictions of any classifier. In *Knowledge discovery and data mining (KDD)*.
Russell, S., & Norvig, P. (2009). *Artificial intelligence: a modern approach* (3rd ed.). Upper Saddle River, NJ, USA: Prentice Hall Press.
Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review, 87*, 1085.
Smiley, M. (2017). Collective responsibility. The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.). https://plato.stanford.edu/archives/sum2017/entries/collective-responsibility. Accessed 4 August 2019.
Taddeo, M. (2009). Defining trust and e-trust: old theories and new problems. *International Journal of Technology and Human Interaction (IJTHI), 5*(2), 23–35.
Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines, 20*(2), 243–257.
Taddeo, M., & Floridi, L. (2011). The case of e-trust. *Ethics and Information Technology, 13*(1), 1–3.
Talbot, B., Jenkins, R., & Purves, D. (2017). When robots should do the wrong thing. *Robot Ethics, 2*, 258–273.
Tuomela, M., & Hofmann, S. (2003). Simulating rational social normative trust, predictive trust, and predictive reliance between agents. *Ethics and Information Technology, 5*(3), 163–176.
Vapnik, V. N. (2000). *The nature of statistical learning theory*. New York: Springer.
Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law, 7*(2), 76–99.
Walker, M. U. (2006). *Moral repair: reconstructing moral relations after wrongdoing*. Cambridge: Cambridge University Press.