Chapter 11

Trustworthy humans and machines

Vulnerable trustors and the need for trustee competence, integrity, and benevolence in digital systems

Sara Degli-Esposti and David Arroyo

Introduction: trust and digital mediation

In the future happening today coders dream of erasing discrimination and corruption by replacing traditional institutions with new digital systems such as Distributed Ledger Technologies, or DLTs, in an attempt to restructure old institutions by means of computer code rather than through collective action. Satoshi Nakamoto's (2008) blockchain proposal to generate electronic transactions and cryptocurrency "without relying on trust" exemplifies this attitude, namely the use of *lex cryptographia* to restore institutions (De Filippi and Loveluck 2016). The problem with these kinds of proposals is that dependence on Information and Communication Technologies (ICT) may lead to an overabundance of trust in untrustworthy, yet credible and sometimes dependable, systems.

Our objective in this chapter is to discuss issues of dependence in the trust relationship that limit the ability of transparency to guarantee the trustworthiness of the trustee. We embrace Onora O'Neill's (2017) invitation to focus on what really matters about trust, which is people's ability to trust the trustworthy and distrust the untrustworthy in the context of digitally mediated interactions, where cryptography is reshaping the relationship between computer code and legal compliance in unforeseeable ways. We deal with the need to establish mechanisms to ensure that trustees—those humans who design and operate the machines on behalf of others whose life depends on those systems and machines—are trustworthy.

We argue that a fundamental distinction needs to be drawn between dependability and trustworthiness. We agree with Helen Nissenbaum's (2004) view that visions of trust as security lead to surety—that is, safety and certainty—in a best-case scenario, but not to trust conceived as "the accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one" (Baier 1986, 235). We contend that we need to move from dependability to trustworthiness to be able to deal with uncertainty. Under "unknown unknowns," which are risks that come from situations that are unexpected—a topic widely discussed in security studies—mechanisms to guarantee the

DOI: 10.4324/9781003120827-15

trustees' competence, integrity, and benevolence are necessary to build trust in institutions and organizations (Mayer, Davis, and Schoorman 1995). Similarly, when trustors are highly vulnerable and dependent—for example, in the case of citizens versus law enforcement agents—transparency plays a limited role in giving them control over trustees' actions. Under these types of circumstances, those interested in designing resilient organizational or technical systems would look for mechanisms to ensure trustees' competence, integrity, and benevolence. Benevolence, for example, has been demonstrated to be particularly important for trust relationships in the context of digital surveillance technologies used by law enforcement agencies (Degli Esposti, Ball, and Dibb 2021).

This chapter hopes to contribute to the dialogue between social science and computer science by replacing the traditional trust-as-control paradigm with a vision of trust-as-care. We focus on the implications of this view of trust for the field of security engineering, which is devoted to ensuring the dependability of systems and devices. We argue that this new vision would be better suited to articulating the relationship between humans and machines, so important in the path toward trustworthy artificial intelligence, or AI (AI-HLEG 2019a, 2019b).

Trust as control: the rationalistic instrumental paradigm

Trust represents a sort of leap of faith in another person's willingness to cooperate with us. A trust relationship involves two specific parties: a trusting party—that is, the individual rendering trust judgments (trustor)—and a party to be trusted (trustee) (Jones and Shah 2016). The trustee seems to be motivated either by self-interest or by benevolence toward the trustor. Hardin's (2002, 4) influential definition of trust as "encapsulated self-interest"—"I trust you because I think it is in your interest to attend to my interest in the relevant matter"—represents the mainstream approach foregrounding self-interest. According to Hardin (2002), there are three mechanisms by which the trustee can encapsulate the interest of the trustor. First, the two of them have established an ongoing relationship, which is valuable for the trustee. Second, the trustee loves or is a friend of the trustor; thus, the trustor can count on the trustee's benevolence. Third, the trustee wants to maintain his or her good reputation, which provides motivation to behave in a trustworthy manner.

The rationalistic instrumental paradigm of trust has been criticized for its individualistic, utilitarian assumptions, which emphasize individual self-interest over collective benefits. Experimental methods, games and abstract dilemmas, and disembodied human interactions have repeatedly questioned the validity of this approach. According to Michael Hechter (1992, 34), there is "ample reason to be skeptical of the sufficiency of game theory for

the solution of real-world collective action problems." As many empirical studies show, there is no society in which behavior is consistent with the selfishness axiom (Henrich et al. 2004). The problem is that self-interest does not explain sacrifice; sacrifice generated by affection or by a duty of care is central to the experience of those who care about other people's survival. In the view of psychologist Roderick Kramer (2009), "human beings are naturally predisposed to trust" because "it's a survival mechanism that has served our species well." The "care-giving we provide to others is as fundamental to human nature as our selfishness or aggression" (Taylor 2014, 4). Trustors' and trustees' shared experiences and destinies irreversibly forge their identities. This vision of intertwining paths, which should lead toward beneficial collective outcomes, is ignored by players trapped within a utilitarian logic.

Another limitation of the rationalistic instrumental paradigm is its tendency to deny the role of history and social norms. Collective history offers guidance to individuals on whether norms of trust and reciprocity exist and will be respected in each context (Berg, Dickhaut, and McCabe 1995). Some scholars argue that people appear to follow an "injunctive norm," which impels them to trust the character of the other person (see e.g., Fetchenhauer, Dunning, and Shlösser 2017). Those who believe that cooperation is beneficial and are willing to cooperate are also more inclined to believe that other people will share the same view and will behave accordingly. In ongoing relationships, expectations of reciprocity facilitate cooperation (Axelrod 1997) and may also influence perceptions of trustworthiness, which relates to the trustor's confidence in the trustee based on experiences or beliefs (Berg, Dickhaut, and McCabe 1995). Of course, when the relationship is sporadic—so that the trustee does not face any negative consequence caused by the trustor's lack of future cooperation—the incentives to deceive the other person may increase.

To conclude, we may assume that a good proportion of humans are wise and willing to care for human survival and thus acknowledge the value of cooperation and reciprocity. These humans may decide to set trust as a default systemic parameter. The assumption that trust—rather than distrust—is taken as the default position in many cultures finds additional support in the next section, where we consider some psychological studies and introduce a new characterization of the trustee–trustor relationship.

Trust as care: on the trustor's vulnerability and the trustee's benevolence

Mayer, Davis, and Schoorman (1995, 712) interpret trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." In their theoretical model, the level of trust is determined by the trustee's ability, benevolence, and integrity and by the trustor's propensity to trust.

The way the trustor interprets the context of the relationship affects the need for trust, risk assessment, and the evaluation of the trustee's trustworthiness. The tendency to trust another party is a function of the type of motivation attributed to the other: the more a person perceives another person to be benevolently motivated, the more likely they are to like and trust that person (Colquitt, Scott, and LePine 2007; Van Lange, Rockenbach, and Yamagishi 2017).

Thus, questions of trust seem to arise when an individual is in a relationship that entails some risk of becoming vulnerable to the actions or decisions of another person (Levi and Stoker 2000). There are scholars who see a moral component in trust relationships. For instance, LaRue Hosmer (1995, 393) defines trust as

> the reliance by one person, group, or firm upon a voluntarily accepted duty on the part of another person, group, or firm to recognize and protect the rights and interests of all others engaged in a joint endeavor or economic exchange.

Within this second group, we agree with those scholars who highlight the vulnerability of the trustor. However, what prevents trustees from taking advantage of the vulnerability of the trustors? In other words, what is it that makes trustees trustworthy?

The *empowering theory of trust* suggests that by manifestly relying on another person B—by exercising trust—a person A may not only cause B to exercise their existing capacity for trust-responsiveness, but A may also cause B to *develop* that capacity, achieving a higher degree of dependability or durability. According to McGeer and Pettit (2017), three psychological effects contribute to what they call the "situational enhancement of dependability." The first is that when player A trusts player B, they display and communicate a belief in B's capacity to be motivated by A's manifest reliance, thereby encouraging B to prove reliable. The second is that when A trusts B to do something, A often makes a request, explicit or implicit, that B should do what is requested. And the third is that when A trusts B, A displays a good opinion of B's dependability, thereby giving B an extra esteem-based motive for not letting A down (see also Elster 2007). When player A decides to trust player B, this decision has a positive, empowering impact on B's psychology (Pettit 2002).

Thus, the mere fact of trusting—or declaring that one trusts—creates an obligation for the trustee to honor that trust, which (assuming some moral responsiveness to obligation on the part of the trustee) increases the probability that the trustee will demonstrate greater trustworthiness than originally expected. From an instrumental, utilitarian perspective, these reflections leave open questions on how to secure trustworthiness in the absence of transparency and control but in the presence of vulnerability and dependence. The

problem of discretion and lack of trustor's control over the trustee is well represented in the principal(trustor)–agent(trustee) model. The principal (e.g., a patient) has to delegate a task to an agent (e.g., a doctor) because the former lacks the ability to perform it. From a rationalistic, instrumental view, the principals can monitor the agents or create economic incentives to ensure they act in a trustworthy manner, that is, in the best interest of the principals.

The problem with the principal–agent framework is that it assumes the agent knows what is in the best interest of the principal. In other words, the theory assumes the agent's competence. It also assumes the principal has the power and the information to make the agent accountable. However, if we observe trust from the vantage point of the vulnerable—the newborn baby, the dependent elderly, the sick person—information loses its value and power is completely imbalanced. The newborn cannot assess its caregiver's intention or ability, even though survival depends on the caregiver's benevolence and competence. A capital of trust is handed over to the trustee as a blank check. The return of that investment will become visible in the long term with limited initial accountability. In the case presented here, we assume that the benevolence of the agents-trustees will motivate them to become competent and to act with integrity. Nonetheless, the principals-trustors' vulnerability prevents any meaningful expression of control through transparency on the other side of the relationship.

To sum up, we contend that the rationalistic instrumental paradigm offers an illusion of freedom and a denial of dependence, which are both dangerously misleading. The trust-as-control paradigm resolves any moral hazard problem by means of transparency, today achieved through digital surveillance. This vision generates widespread reliance on risk-based methodologies across different areas and a growing demand for data. We contend that mechanisms such as transparency cannot be effective in the presence of a high imbalance of power and that only agents on a level playing field can exercise meaningful mutual oversight. Furthermore, the trust-as-control paradigm offers no indication as to how to inscribe competence and moral principles into a trustee's identity. We stress the importance of benevolence in the trust relationship in the presence of vulnerable trustors. Benevolence in this scenario matters because it determines whether humans in power will decide to deceive other dependent and vulnerable humans or treat them with care and respect.

To better articulate these reflections, in the next section we propose an alternative characterization of the trustor–trustee relationship: the *caring one* (trustee) and the *vulnerable other* (trustor). This vision of trust as embracing the care of the vulnerable resembles the one adopted by Gus Hurwitz (2012), who takes trust to mean "reliance without recourse" in the context of online interactions. It also resonates with Annette Baier's (1986, 240) definition, which says that "[trust] is letting other persons (natural or artificial, such as

firms, nations, etc.) take care of something the trustor cares about, where such 'caring for' involves some exercise of discretionary powers."

The caring one (trustee) and the vulnerable other (trustor)

In the presence of vulnerable and dependent trustors, trustees need to demonstrate their competence, honesty, and benevolence in order to be considered trustworthy, that is, *able to meet the promise of care intrinsic to their role*. When trustors are vulnerable and dependent, trustees have to care for them in the absence of direct instructions on what the trustors need. The instrumental paradigm of trust-as-control offers the transparency of trustees' actions as a solution to any moral hazard or conflict of interests. However, this framework assumes trustees know how to act in the best interest of the trustors.

But, even assuming benevolence, how can trustees know what is beneficial for the trustors? We argue that the trustor needs not only the trustee's *dependability* but also their trustworthiness, that is, a mixture of *learned new knowledge* and *moral considerations* that will lead to some type of *wisdom*. The learning process leading to the creation of this knowledge base would start from a capital of affection that would make the trustee responsible for the wellbeing of the trustor. This capital, allocated without having previous knowledge of the trustworthiness of the trustee, would trigger a learning process that would lead the trustee to investigate a trustor's needs.

When the role is not defined by deep affection, duty of care principles could replace affection in guaranteeing effort is allocated to learning about a trustor's needs. Professional codes can instruct about the need to develop specific methodologies and about the necessity to embed empathy into trustees' professional identities (Kultgen 1988). Even though disciplinary methods can be applied to achieve transparency or to monitor professionalism (Fournier 1999), there are different domination and knowledge-generation dynamics at play in each case. In the trust-as-care scenario, norms of care are defined and voluntarily embraced by trustees within their epistemic communities (Haas 1992).

Mechanisms to foster professionalism differ from those transparency measures envisioned by supporters of the trust-as-control paradigm. Even though peer-pressure mechanisms may be present, it is the adoption of shared norms and mutual learning that makes trustees willing to become competent and that keeps them honest in the trust-as-care case. In other words, despite both being normative systems, the type of norms operative in the trust-as-control paradigm differs from that preached in the trust-as-care case. We next move this discussion to the implications of adopting the trust-as-care perspective in the fluid boundary where "the ordinary language systems terminate in the special sort of machine known as a human being" (Wiener 1988, 79).

From credible machines to dependable systems: drawing a distinction between dependability and trustworthiness

As machines are built by humans, we began by talking about the trustworthiness of those human beings acting in institutional or other organizational settings who create or operate technological systems. We now move to discuss the trustworthiness of the technical system itself; in the end we will reconcile the discussion about the trustworthiness of the machines and of their creators.

If we think about whether we trust computers, we will probably see them as reliable devices that enable us to perform daily activities such as reading emails, managing meetings, or editing and sharing documents. As noticed by Fogg and Tseng (1999), mass reliance on ICT would not be possible in a world where people were unwilling to trust credible computers. However, users' trust perceptions do not necessarily reflect trustworthiness attributes: malware or spear phishing attacks, for example, exploit systems' credibility to insert malicious code into the machines of their victims (Mitnick and Simon 2011).

The risks associated with the existence of malevolent agents, software, and untrusted hardware render trust a broad research topic, which spans areas as diverse as security and access control in computer networks, reliability in distributed systems, and policies for decision-making under uncertainty (Artz and Gil 2007). Even though the concept of trust in these different communities varies in how it is represented, computed, and used, overall we may say that "a trusted system or component is one whose failure can break the security policy, while a trustworthy system or component is one that won't fail" (Anderson 2008, 13). For instance, in the realm of the so-called Internet of Things (IoT), trust management implies ensuring that the physical perception layer made of sensors and actuators cooperates with the network layer, which transforms and processes sensed environment data, and with the application layer, which offers context-aware intelligent services (Sicari et al. 2015).

"Dependability is the system property that integrates such attributes as reliability, availability, safety, security, survivability, maintainability" (Avizienis, Laprie, and Randell 2001, 1). Dependable systems have integrity: they perform their intended functions in an unimpaired manner, free from deliberate or inadvertent unauthorized manipulation of the system (Greene 2014). To ensure the dependability of software and infrastructures, secure systems need to be able to operate within a context of adversity (Danezis 2014). Dependable systems are resilient[1]: they are able to resist and recover from disruptions and attacks.[2] Attackers can be passive or active, internal or external, and local or global with respect to the system they want to attack. A number of model-based evaluation techniques are available along with experimental red team-based approaches (Nicol, Sanders, and Trivedi 2004). In general, we may say that a trusted computing base (TCB) is a minimal set of components of a

system upon which the security of the entire system depends (Lysne 2018).[3] The objective of security engineering is "to design systems that are resilient in the face of malice, that degrade gracefully, and whose security can be recovered simply once the attack is past" (Anderson 2008, 212).

Authentic trustworthy trustees would be willing to draft security and privacy policies that ensure system dependability across all hardware and software layers. From the root of trust up to the automated decision support system, roles and responsibilities of human and machine trustees would become more visible and auditable by enabling algorithmic explainability and, hopefully, contestability (Vaccaro et al. 2019). We argue that trustworthiness and dependability represent distinct ideas, which need to be treated differently.[4] The distinction between trustworthiness and dependability reflects the difference between writing a policy and applying a policy. We expect trustworthy trustees to write the security policy on behalf of vulnerable trustors by taking into consideration both system owners' and end users' preferences. This distinction is important when it comes to discussing privacy/security–usability tradeoffs. If we think of digital platforms it is easy to see the conflict between platform surveillance capacity and end users' privacy. To ensure that privacy and security policies respond to trustors' needs, trustees need to be competent, honest, and benevolent toward all types of trustors. Extending trustees' benevolence to all trustors of a digital system requires the creation of governance mechanisms promoting ethics-of-care by design, professionalism, and integrity.

Trustworthy trustees writing information security and privacy policies

If we assume that those who have the ability to design and develop the system are the trustees, and that those who use or own the system are the trustors, we may explore their relationship in terms of dependence and vulnerability, to guarantee the respect of a duty of care in the development and application of security/privacy policies and procedures. Trustor-users, who do not design or deploy the system but still use it, tend to be dependent and vulnerable. Dependence derives from limited knowledge and a lack of convenient alternatives. The vulnerability and dependence of digital system end users are often discussed in the computer science literature. Under the famous "Why Johnny can't encrypt" lemma, several studies demonstrate users' reticence to adopt information security measures mostly because of the limited usability of available solutions (Whitten and Tygar 1999; Sheng et al. 2006; Ruoti et al. 2015). These problems affect individual users as well as entire industries, nation-states, and corporations, as pointed out by scholars working in the field of information security economics (Anderson and Moore 2006).

Widespread adoption of privacy-preserving measures is even more challenging. Privacy policies represent a good example of the reason why we claim

that under dependence and vulnerability, transparency is meaningless or even detrimental. Privacy policies are very long, obscure, and seldom read or understood by end users (McDonald and Cranor 2008; Vail, Earp, and Antón 2008). This implies that privacy policies do not help firms keep their privacy promises—which are viewed by consumers as not credible—or increase transparency and market efficiency (Farrell 2012). Despite all the efforts made to increase the readability and usability of these policies (Acquisti, Adjerid, and Brandimarte 2013), they still ineffectively communicate privacy risks and do not contribute to raising information security and privacy awareness.

Because trust is not interpreted as care but as control, corporations (trustees) have no intrinsic motivation or experience no peer pressure to protect their users' (trustors') privacy. Current available measures are designed to leverage data controllers' fear of losing their good reputation. An example of such mechanisms is the data breach notification provision present in the EU General Data Protection Regulation (GDPR), which relies on sanctions and negative publicity to force corporations to improve their information security procedures. Despite this measure being promising, we argue that information transparency is of limited use when the trustors are vulnerable—having no ability to technically engage with the system—yet still depend on the system. This implies that giving trustors more transparency over the decisions of trustees will not serve to increase the latter's trustworthiness.

By acknowledging the vulnerability of trustors, we implicitly admit how difficult it would be for this constituency to effectively negotiate security and privacy policies beneficial to them. An ethic of care, not utilitarianism, should inform and guide decisions taken by trustees on behalf of trustors—with the trustees being the programmers, standardization body members, scientists, and cryptographers, and the trustors being anyone who depends on the ICT system. We argue that the adoption of a vision of trust-as-care would foster the creation of other types of mechanisms. In the remaining part of the chapter, we try to sketch some proposals, after reviewing current mechanisms to establish the trustworthiness of the trustees.

On the authenticity of trustworthy trustees: authentication and authorization

"Whom do you trust?" and "for doing what?" are typical questions in conversations about trust. Are there identity traits or attributes that make someone naturally trustworthy? In the field of information security, the authenticity of one's identity—and, most importantly, the attributes of that identity—are taken as a given (or assumed as authentic) unless we suspect that we are dealing with a malicious entity that is lying about their identity to perform an attack. Authentication is a key element of information security. Through authentication, we assign information disclosure privileges, assess the reliability and integrity of a piece of information, authorize transactions, and

conduct audits. Multifactor authentication, which is required by the National Institute of Standards and Technology (NIST 2017) and compulsory for the Fintech sector in Europe,[5] is increasingly used to ensure proper authentication. Trust anchoring and oracles are other mechanisms widely applied in this domain. While trust anchoring involves the association of information about an object from reliable sources, oracles can be human beings or automated agents. These solutions can only be effective if we ensure the traceability and linkability of digital information with its original source. For instance, the main requirement in designing machine oracles is that the authenticity of the data must be publicly verifiable (van der Laan 2018).

During daily activities, the trustworthiness of another human being and the authenticity of their identity are established through face-to-face interactions. An example of how physical identities mutate into digital identities are *key signing parties*, which are get-togethers of people who use the PGP[6] encryption system. A Public Key Infrastructure (PKI)[7] is an arrangement that binds public keys with respective identities of entities (i.e., people or organizations). "Key signing" refers to the act of digitally signing a public key packet and a user ID packet; the aim is to verify that a given user ID and public key really belong to the entity that appears to own the key; in other words, to verify that the representation of identity in the user ID packet is valid. Usually, this means that the name on the PGP key matches the name on the identification that the person presents to you when asking that you sign their key.

In other words, physical, face-to-face contact is needed to assess the authenticity of one's identity. Bureaucratic systems also envisage analog entry points to establish the trustworthiness of the counterparty and the intermediary and to set up dispute resolution mechanisms (Werbach 2018a). The European eIDAS regulation (EU 2014), for instance, forces people to prove physical identity in front of an authority, which is assumed to be a trustworthy intermediary. The intervention of real humans is also necessary to set up dispute resolution mechanisms. For instance, the dream of blockchain as a disembodied trustless trust solution ended on June 17, 2016, when cryptocurrency worth USD 55 million was siphoned off by an anonymous user who exploited a loophole in the source code of the Ethereum Blockchain platform (Reyes 2019). The operation was legitimate from the perspective of the software, which could not distinguish a customer from a thief (Werbach 2018b). It was also technically irreversible and immutable, which implied that human intervention was needed to create a hard-fork, namely a bifurcation of the blockchain from the moment before the theft happened and a reimbursement to those affected by the illicit operation. Thus, human intervention was required to resolve the dispute triggered by the theft and to shape the history of the two parallel platforms, known as Ethereum and Ethereum Classic, which now exist.

Several authentication procedures exist to establish authenticity, that is, to establish that the being or thing that one is communicating with is

who or what they claim to be. No procedure, however, asks the entity to prove its competence, integrity, and benevolence. Here we argue for the need to establish the trustworthiness of the original source, namely, of the humans building and operating the system, that is, the trustees. We can imagine some sort of "artistic" irreversible signature left by the designer and administrators of the system that certifies their benevolence, competence, and integrity. Authorship mechanisms may help foster peer-review accountability among trustees, show their benevolence, and foster their trustworthiness.

Mechanisms to extend the roots of trust

Along the course of this chapter, we have rejected the trust-as-control paradigm and adopted a vision of trust as care in order to ask questions on how to distinguish trustworthy trustees from untrustworthy ones and how to build dependability and trustworthiness from the root of trust up to the interface. A trust-as-care vision of information security would expand the root of trust from the technical layer to the human component by reinforcing peer-review mechanisms among trustees who are designers and system administrators. New frameworks would see technical authentication mechanisms complemented by governance mechanisms designed to inscribe competence, honesty, and benevolence into the identities of the human trustees, who would guarantee the dependability of the system and the respect of policies. Technologists (trustees) need to unite in an epistemic community of practice informed by the highest ethical and professional standards to be able to generate the knowledge needed to produce next-generation trustworthy technology, so important especially in the case of AI-driven critical infrastructures. We argue that emerging technologies such as quantum computing demand the creation of new spaces of critical and constructive dialogue, enabling trustees to learn about trustors' needs.

Trustees' trustworthiness is generated by trustees' competence, which demands the leveraging of expert knowledge; integrity, which requires training and application of ethical codes of conduct; and benevolence, which demands that trustees learn about trustors' needs and openly discuss their corporate mission, business rationale, and technical and organizational methods with the needs of clients or users in mind. If the trustee has a duty of care toward the trustor, the respect of this duty of care should be guaranteed by other trustees within collegial bodies that underwrite codes of conduct and codes of principles, and through mentoring, training, and education (ECA 2019). Professionalism, knowledge generation, and peer review should be guaranteed and fostered through collegial bodies supporting the activities of, and decisions taken by, the trustees. Examples of collegial bodies are standardization authorities,[8] professional associations and forums, and the scholarly and scientific community.

Mechanisms to reinforce collaboration and mutual accountability among trustees can protect society against the risk of technological determinism and herding behavior in policy and R&D investment decisions. Technological determinism and herding behavior may lead policymakers to ignore certain policy stages, such as problem structuring and definition, as noticed by Veale (2019), or certain problems (assessing the usefulness of computing in any given context), while spending time and effort on issues related to economic competitiveness (e.g., increasing the availability or intensity of European AI). As competition may prevent beneficial exchanges of knowledge and expertise, the creation of nonmonetary social markets for auditability and accountability could facilitate the exchange of confidential information among trustees working in the security and digital surveillance domains. Of course, soft coordination mechanisms like these need to be anchored in other types of strong enforcement procedures in order to ensure prompt conflict resolution and intervention. Ben Wagner (2019, 89–99) suggests providing "a mechanism for external independent (not necessarily public) oversight" and "a clear statement on the relationship between the commitments made and existing legal or regulatory frameworks, in particular on what happens when the two are in conflict."

To foster a vision of security as a public good, new legal instruments and governance methods to facilitate security audits (see e.g., Sanchez-Gomez et al. 2018 in the domain of cloud storage) should be envisioned in order to facilitate the discovery of system vulnerabilities and other privacy and security issues. In the domain of machine and deep learning, "blind trust" mechanisms could be devised to enable algorithm auditing and the sharing of training datasets. Imagine a scenario in which the management of a company developing a predictive algorithm wants to understand the system's privacy and reidentification risks. Data and code could be anonymously sent to a Digital Blind Trust (DBT) with instructions on the tasks to be performed. The Trust would open a bid and assign the task to an anonymous research team, after controlling for potential conflicts of interest. The anonymous team would perform the analysis. Results would be sent to the client for rebuttal. The revised version of the study would be published on the trusted network and made public according to confidentiality agreements, which would balance individual and collective interests. This and similar types of systems could be designed to enable peer pressure and peer review among trustees.

The considerations and proposals made here are not meant to undermine the role of trustors in fostering the trust relationship. The High-Level Expert Group on Artificial Intelligence (AI-HLEG 2019b, 12), in its second report on "Policy and investment recommendations for trustworthy Artificial Intelligence," suggests

> [i]ntroduc[ing] a mandatory self-identification of AI systems … [Given that] there is a reasonable likelihood that end users could be led to believe

> that they are interacting with a human, deployers of AI systems should be attributed a general responsibility to disclose that in reality the system is non-human.

We want to clarify that a focus on the trustworthiness of the trustees does not preclude "[p]romoting the ability of individuals and society as a whole to understand and reflect critically in the information society," which is an important recommendation made by the Data Ethics Commission for the Federal Government's Strategy on Artificial Intelligence (DEK 2018, 1). If trustees have a duty of care toward trustors, they have an obligation to maintain a permanent dialogue with the trustors, understand their needs and demands, and increase their awareness and literacy. Furthermore, we suggest that trustors should retain some degree of skepticism in the form of *parrhesia* (Foucault 1983) to denounce untrustworthy trustees and wrongdoing. Trustors could also be willing to play *parrhesiastic games* to help trustees demonstrate their ability to listen and calibrate their actions in their best interest. Trustees should review each other's actions and decisions to help enhance their knowledge of how to better care for the trustors.

Conclusion

The problem at the core of this article is how we can ensure that we trust the trustworthy and distrust the untrustworthy when we are confronted with disembodiment and automated beings to which we cannot direct our gaze. Information technology introduces a conception of trust as dependability, reliability, or credibility compatible with visions of trust-as-control rooted in the rationalistic instrumental paradigm. However, as noted by Olav Lysne (2018, 18), "we should not make Hardin's kind of trust a basis for our security concerns about equipment in a country's critical infrastructure." While the necessity of shedding light on economic incentives and psychological biases that shape security policy decisions has been acknowledged (Anderson and Moore 2009, 2006), the role that ethics and moral principles should play in defining next-generation security policies has received little attention.

In this chapter we have challenged the underlying assumption, present in the rationalistic instrumental visions of trust, that the trustor enjoys the freedom not to trust the trustee. By presenting the caring-one and vulnerable-other dyad as an alternative to the utilitarian trustor–trustee dyad, we have argued for the need to embed an ethic of care, and not simply a logic of control, into the trustors. In the presence of dependence and vulnerability we argue that the logic of control, based on transparency, sanctions, and incentives, is useless, even detrimental. The issue then becomes how to foster the trustee's trustworthiness, beyond the trust-as-dependability currently pursued and enacted in the information security domain. Trustworthiness concerns the

confidence of the trustor that the trustee has attributes, such as competence or integrity, that serve the trustee in a beneficial manner (Gabarro 2014). We trust our doctor, or the pilot of the plane, to do their job in a professional manner; in other words, we expect professionals to perform their duties—that is, to follow certain established social norms by showing high levels of competence, integrity, and benevolence.

The emphasis on trustworthiness is meant to reconcile functional, privacy, and security requirements with multiparty-negotiated policies and foreground the pivotal role of coders' and operators' competence, integrity, and benevolence. We acknowledge the continuity between the trust-as-control and the trust-as-care models, and simply clarify that in the presence of highly vulnerable trustors a logic of trust-as-care should be preferred over a logic of trust-as-control, which is better suited for scenarios featuring low dependence and low vulnerability. We argue that in a scenario where the trustor has enough autonomy to exercise a certain degree of control over the trustee, all they need is the trustee's dependability. In the opposite case, when the trustor is highly vulnerable and depends on the trustee, with limited or no control or exit strategy, the trustee needs to demonstrate trustworthiness, that is, the ability to take care of the trustor in the absence of control, but in the presence of an ethic of care.

If we are truly moving toward a future in which computer code is the new law, the only chance we have to program sensible machines is to train a new generation of culturally, morally, and socially sophisticated coders able to confront the challenge and embrace the normativity and performativity of the system they are designing. From a security engineering perspective, the question "is the system trusted?" is underdefined unless we answer other related questions, such as "By whom? For which attributes? Against what adversary?" As in everyday reality, the question, "Do you trust them?" should be qualified with "trust them to do what?" to take into consideration the ability of the trustees to deliver on their promise of care.

Some commentators claim that cryptography has a role to play in keeping power in check,[9] whether in protecting those resisting authoritarian regimes or in bringing more transparency to democratic ones (Rogaway 2015). We hope that our reflections will help inspire new generations of coders (cryptographers and lawmakers) willing to cooperate in the name of human flourishing and security as a public good. We also hope that these coders will be inspired by new expressions of moral philosophy, different from those which replicate "uncaring forms of justice and unjust forms of care" (Clement 2018, 2) that amplify unfairness through the denial of basic human conditions, such as dependence and vulnerability and the need of care.[10] We hope that a vision of trust based on a philosophy of care could help us better reflect on the relationship between transparency and digital surveillance in new policy and technology terms.

Acknowledgments

This work was partially funded by the "TRESCA—Trustworthy, Reliable, and Engaging Scientific Communication Approaches" project, funded by the European Union's Horizon 2020 Research and Innovation Program under grant agreement no. 872855, and by the project "CYNAMON—Cybersecurity, Network Analysis, and Monitoring for the Next Generation Internet," sponsored by "Programas de Actividades de I+D entre grupos de investigación de la Comunidad de Madrid en tecnologías 2018" (P2018/TCS-4566), cofinanced with FSE and FEDER EU funds.

Notes

1 Dependability represents "the ability to deliver service that can justifiably be trusted," while resilience is "the persistence of service delivery that can justifiably be trusted, when facing changes" (Laprie 2008, 8).
2 Typical examples are: denial-of-service attacks, which limit or jeopardize data or system availability; man-in-the-middle attacks, which disrupt the confidentiality of communications; zero-day or SQL-injection attacks, which disrupt system integrity through vulnerability exploitation or code injection; and adversarial attacks on neural networks (deep learning) that compromise data integrity and system performance.
3 It is worth noticing that "[e]ach virtual machine presumes the correctness (integrity) of whatever virtual or real machines underlie its own operation" (Arbaugh et al. 1997, 1). In other words, a technical system is made of many interdependent layers; the security of each layer is dependent on assumptions made about the functioning of previous layers.
4 Of course, we are adopting a reductionist logic to produce binary categories and we acknowledge that reality is the gray zone which lies in-between these two extreme scenarios and that the two ideas need to coexist and complement each other.
5 Payment services (PSD 2)—Directive (EU) 2015/2366, URL: https://ec.europa.eu/info/law/payment-services-psd-2-directive-eu-2015-2366_en. NIST Special Publication 800-63B "Digital Identity Guidelines," URL: https://pages.nist.gov/800-63-3/sp800-63b.html
6 PGP stands for "pretty good privacy (data encryption)." Public key cryptography infrastructure (PKI) has two main implementations. One is done using certificates and certificate authorities (CAs) and is described in the X.509 standard. It is best suited for structured organizational hierarchies with an implicitly trusted authority that vouches for all issued certificates. It is the standard that is behind SSL/TLS and S/MIME email encryption. However, there is also another widely used standard for PKI, which was developed with the explicit intention of avoiding centralized certification authorities, and instead relies on trust relationships built between regular users. It was first implemented in the original PGP software back in 1991 and, since then, has developed into a robust open standard, known as OpenPGP (openpgp.org) for email encryption.
7 PKI is a set of protocols, standards, and procedures to manage public key encryption and digital certificates (Adams and Lloyd 1999).

8 E.g., ISO International Standards; the National Institute of Standards and Technology (NIST), part of the US Department of Commerce; "Bundesamt für Sicherheit in der Informationstechnik" (BSI).
9 For instance, Tor (www.torproject.org) has found considerable success as a censorship-circumvention tool.
10 Of course, engaging with ideas of care and control leads us to face two famous stereotypical constructions: womanhood (Clement 2018) and blackness (Mbembe 2017).

References

Acquisti, Alessandro, Idris Adjerid, and Laura Brandimarte. 2013. "Gone in 15 Seconds: The Limits of Privacy Transparency and Control." *IEEE Security & Privacy* 11(4): 72–4.

Adams, Carlisle, and Steve Lloyd. 1999. *Understanding Public-Key Infrastructure: Concepts, Standards, and Deployment Considerations*. Indianapolis: Sams Publishing.

AI-HLEG. 2019a. *Ethics Guidelines for Trustworthy AI.* High-Level Expert Group on Artificial Intelligence, European Commission. April 8, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

AI-HLEG. 2019b. *Policy and Investment Recommendations for Trustworthy Artificial Intelligence.* High-Level Expert Group on Artificial Intelligence, European Commission, https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

Anderson, Ross. 2008. *Security Engineering: A Guide to Building Dependable Distributed Systems*. Indianapolis: Wiley.

Anderson, Ross, and Tyler Moore. 2006. "The Economics of Information Security." *Science* 314(5799): 610–3.

Anderson, Ross, and Tyler Moore. 2009. "Information Security: Where Computer Science, Economics and Psychology Meet." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367(1898): 2717–27.

Arbaugh, William A., Angelos D. Keromytis, David J. Farber, and Jonathan M. Smith. 1997. "Automated Recovery in a Secure Bootstrap Process." *University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-97–13*.

Artz, Donovan, and Yolanda Gil. 2007. "A Survey of Trust in Computer Science and the Semantic Web." *Web Semantics: Science, Services and Agents on the World Wide Web* 5(2): 58–71.

Avizienis, Algirdas, Jean-Claude Laprie, and Brian Randell. 2001. Fundamental Concepts of Dependability. *UCLA CSD Report no. 010028; LAAS Report No. 01–145; Newcastle University Report No. CS-TR-739, 2001*, https://pld.ttu.ee/IAF0530/16/avi1.pdf.

Axelrod, Robert. 1997. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Vol. 3. Princeton: Princeton University Press.

Baier, Annette. 1986. "Trust and Antitrust." *Ethics* 96(2): 231–60.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10(1): 122–42.

Clement, Grace. 2018. *Care, Autonomy, and Justice: Feminism and the Ethic of Care*. New York: Routledge.

Colquitt, Jason A., Brent A. Scott, and Jeffery A. LePine. 2007. "Trust, Trustworthiness, and Trust Propensity: A Meta-Analytic Test of their Unique Relationships with Risk Taking and Job Performance." *Journal of Applied Psychology* 92(4): 909–27.

Danezis, George. 2014. "Trust as a Methodological Tool in Security Engineering." In *Trust, Computing and Society*, edited by Richard H. R. Harper, 68–91. Cambridge: Cambridge University Press.

De Filippi, Primavera, and Benjamin Loveluck. 2016. "The Invisible Politics of Bitcoin: Governance Crisis of a Decentralized Infrastructure." *Internet Policy Review* 5(3): 1–28.

Degli Esposti, Sara, Kirstie Ball, and Sally Dibb. 2021. "What's In It For Us? Benevolence, National Security, and Digital Surveillance." *Public Administration Review*: 1–12, doi.org/10.1111/puar.13362.

DEK. 2018. *Recommendations of the Data Ethics Commission for the Federal Government's Strategy on Artificial Intelligence.* Data Ethics Commission for the Federal Government's Strategy on Artificial Intelligence, www.bmjv.de/SharedDocs/Downloads/DE/Ministerium/ForschungUndWissenschaft/DEK_Empfehlungen_englisch.html?nn=11678512.

ECA. 2019. *Challenges to Effective EU Cybersecurity Policy: Briefing Paper.* European Court of Auditors, European Union, www.eca.europa.eu/Lists/ECADocuments/BRP_CYBERSECURITY/BRP_CYBERSECURITY_EN.pdf.

Elster, Jon. 2007. *Explaining Social Behavior. More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

EU. 2014. "Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on Electronic Identification and Trust Services for Electronic Transactions in the Internal Market and Repealing Directive 1999/93/EC." *Official Journal of the European Union*, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32014R0910.

Farrell, Joseph. 2012. "Can Privacy Be Just Another Good." *Journal on Telecommunications and High Technology Law* 10: 251–64.

Fetchenhauer, Detlef, David Dunning, and Thomas Shlösser. 2017. "The Mysteries of Trust. Trusting Too Little and Too Much at the Same Time." In *Trust in Social Dilemmas*, edited by Paul A. M. Van Lange, Bettina Rockenbach, and Toshio Yamagishi. Oxford: Oxford University Press.

Fogg, BJ, and Hsiang Tseng. 1999. "The Elements of Computer Credibility." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Pittsburgh, PA: Association for Computing Machinery (ACM).

Foucault, Michel. 1983. "Discourse and Truth: The Problematization of Parrhesia." Six Lectures Given at the University of California at Berkeley, Berkeley, October–November, https://foucault.info/parrhesia/.

Fournier, Valérie. 1999. "The Appeal to 'Professionalism' as a Disciplinary Mechanism." *The Sociological Review* 47(2): 280–307.

Gabarro, John J. 2014. "The Development of Working Relationships." In *Intellectual Teamwork: Social and Technological Foundations of Cooperative Work*, edited by Jolene Galegher, Robert E. Kraut, and Carmen Egido, 79–110. New York: Psychology Press.

Greene, Sari. 2014. *Security Program and Policies: Principles and Practices*. 2nd Edition. Indianapolis, IN: Pearson IT Certification.

Haas, Peter M. 1992. "Introduction: Epistemic Communities and International Policy Coordination." *International Organization* 46(1): 1–35.

Hardin, Russell. 2002. *Trust and Trustworthiness*. *The Russell Sage Foundation Series on Trust*. New York: Russell Sage Foundation Publications.

Hechter, Michael. 1992. "The Insufficiency of Game Theory for the Resolution of Real-World Collective Action Problems." *Rationality and Society* 4(1): 33–40.

Henrich, Joseph Patrick, Robert Boyd, Samuel Bowles, Ernst Fehr, Colin Camerer, and Herbert Gintis. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press on Demand.

Hosmer, LaRue Tone. 1995. "Trust: The Connecting Link Between Organizational Theory and Philosophical Ethics." *Academy of Management Review* 20(2): 379–403.

Hurwitz, Justin. 2012. "Trust and Online Interaction." *University of Pennsylvania Law Review* 161: 1579–622.

Jones, Stephen L., and Priti Pradhan Shah. 2016. "Diagnosing the Locus of Trust: A Temporal Perspective for Trustor, Trustee, and Dyadic Influences on Perceived Trustworthiness." *Journal of Applied Psychology* 101(3): 392–414.

Kramer, Roderick M. 2009. "Rethinking Trust." *Harvard Business Review*, June 2009, https://hbr.org/2009/06/rethinking-trust.

Kultgen, John H. 1988. *Ethics and Professionalism*. Philadelphia: University of Pennsylvania Press.

Laprie, Jean-Claude. 2008. "From Dependability to Resilience." 38th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Anchorage, USA.

Levi, Margaret, and Laura Stoker. 2000. "Political Trust and Trustworthiness." *Annual Review of Political Science* 3(1): 475–507.

Lysne, Olav. 2018. *The Huawei and Snowden Questions: Can Electronic Equipment from Untrusted Vendors Be Verified? Can an Untrusted Vendor Build Trust into Electronic Equipment?* Vol. 4. Cham, Switzerland: Springer.

Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. "An Integrative Model of Organizational Trust." *Academy of Management Review* 20(3): 709–34.

Mbembe, Achille. 2017. *Critique of Black Reason.* Translated and with an introduction by Laurent Dubois. Durham, NC: Duke University Press.

McDonald, Aleecia M., and Lorrie Faith Cranor. 2008. "The Cost of Reading Privacy Policies." *ISJLP* 4: 543–68.

McGeer, Victoria, and Philip Pettit. 2017. "The Empowering Theory of Trust." In *The Philosophy of Trust*, edited by Paul Faulkner and Thomas Simpson, 14–34. Oxford: Oxford University Press.

Mitnick, Kevin D., and William L. Simon. 2011. *The Art of Deception: Controlling the Human Element of Security*. Indianapolis: John Wiley.

Nakamoto, Satoshi. 2008. "Bitcoin: A Peer-to-Peer Electronic Cash System." Bitcoin Whitepaper, Satoshi Nakamoto Institute.

Nicol, David M., William H. Sanders, and Kishor S. Trivedi. 2004. "Model-Based Evaluation: From Dependability to Security." *IEEE Transactions on dependable and secure computing* 1(1): 48–65.

Nissenbaum, Helen. 2004. "Will Security Enhance Trust Online, or Supplant It?" In *Trust and Distrust Within Organizations: Emerging Perspectives, Enduring Questions*, edited by R. Kramer and K. Cook, 155–88. New York: Russell Sage Publications.

NIST. 2017. *Digital Identity Guidelines: Authentication and Lifecycle Management*. Gaithersburg, MD: National Institute of Standards and Technology (NIST), U.S. Department of Commerce.

O'Neill, Onora. 2017. "Trust, Trustworthiness and Transparency. Output of a Breakfast Briefing held on 24th January 2017." www.britac.ac.uk/sites/default/files/Trust%2C%20Trustworthiness%20And%20Transparency%20briefing%20note%2024%20January%202017.pdf.

Pettit, Philip. 2002. *Rules, Reasons, and Norms*. Oxford: Oxford University Press.

Reyes, Carla L. 2019. "If Rockefeller Were a Coder." *George Washington Law Review* 87: 373–429.

Rogaway, Phillip. 2015. "The Moral Character of Cryptographic Work." 2015 IACR Distinguished Lecture Given at Asiacrypt 2015 on December 2, 2015, in Auckland, New Zealand.

Ruoti, Scott, Jeff Andersen, Daniel Zappala, and Kent Seamons. 2015. "Why Johnny Still, Still Can't Encrypt: Evaluating the Usability of a Modern PGP Client." arXiv preprint arXiv:1510.08555.

Sanchez-Gomez, Alejandro, Jesus Diaz, Luis Hernandez-Encinas, and David Arroyo. 2018. "Review of the Main Security Threats and Challenges in Free-Access Public Cloud Storage Servers." In *Computer and Network Security Essentials*, edited by Kevin Daimi, 263–81, 263–81. Cham, Switzerland: Springer.

Sheng, Steve, Levi Broderick, Colleen Alison Koranda, and Jeremy J. Hyland. 2006. "Why Johnny Still Can't Encrypt: Evaluating the Usability of Email Encryption Software." Symposium on Usable Privacy and Security, Pittsburgh, PA, July 12–14.

Sicari, Sabrina, Alessandra Rizzardi, Luigi Alfredo Grieco, and Alberto Coen-Porisini. 2015. "Security, Privacy and Trust in Internet of Things: The Road Ahead." *Computer Networks* 76: 146–64.

Taylor, Shelley E. 2014. *The Tending Instinct: Women, Men, and the Biology of Nurturing*. New York: Times Books.

Vaccaro, Kristen, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. "Contestability in Algorithmic Systems." Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing, Austin, TX, USA.

Vail, Matthew W., Julia B. Earp, and Annie L. Antón. 2008. "An Empirical Study of Consumer Perceptions and Comprehension of Web Site Privacy Policies." *IEEE Transactions on Engineering Management* 55(3): 442–54.

van der Laan, Bjorn. 2018. "Publicly Verifiable Authenticity of Data from Multiple External Sources for Smart Contracts Using Aggregate Signatures." Master of Science in Computer Science, Department of Intelligent Systems, Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology.

Van Lange, Paul A. M., Bettina Rockenbach, and Toshio Yamagishi. 2017. *Trust in Social Dilemmas*. Oxford: Oxford University Press.

Veale, Michael. 2019. "A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence." *European Journal of Risk Regulation*: 1–10, doi.org/10.1017/err.2019.65.

Wagner, Ben. 2019. "Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping." In *Being Profiled: Cogitas Ergo Sum*, edited by Irina

Baraliuc, Emre Bayamlıoğlu, Mireille Hildebrandt, and Liisa Janssens, 84–90. Amsterdam: Amsterdam University Press.

Werbach, Kevin. 2018a. *The Blockchain and the New Architecture of Trust*. Cambridge: MIT Press.

Werbach, Kevin. 2018b. "Trust, but Verify: Why the Blockchain Needs the Law." *Berkeley Technology Law Journal* 33: 487–550.

Whitten, Alma, and J. Doug Tygar. 1999. "Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0." In *Security and Usability: Designing Secure Systems that People Can Use*, edited by Lorrie Faith Cranor and Simson Garfinkel, 679–702. Boston: O'Reilly.

Wiener, Norbert. 1988. *The Human Use of Human Beings: Cybernetics and Society*. New York: Da Capo Press.