Ethics of Trust--Inviting Digital Systems

Blockchain, Reputation--Based Platforms, and COVID-19 Tracing Technologies

Teng, Y.

DOI
10.4233/uuid:b5134895-8d70-4b2f-a416-662891b2aa0d

Publication date
2021

Document Version
Final published version

Citation (APA)
Teng, Y. (2021). *Ethics of Trust--Inviting Digital Systems: Blockchain, Reputation--Based Platforms, and COVID-19 Tracing Technologies*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:b5134895-8d70-4b2f-a416-662891b2aa0d

Ethics of Trust-Inviting Digital Systems

Blockchain, Reputation-Based Platforms, and COVID-19 Tracing Technologies

Ethics of Trust-Inviting Digital Systems

Blockchain, Reputation-Based Platforms, and COVID-19 Tracing Technologies

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Monday 6 December 2021 at 12:30 hours

by

Yan TENG

Master of Philosophy in Ethics
Dalian University of Technology, China,
born in Dalian, China.

This dissertation has been approved by the promotors.

Promotor: Prof.dr. M.J. van den Hoven
Copromotor: Dr. F. Santoni De Sio

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof.dr. M.J. van den Hoven	Delft University of Technology, promotor
Dr. F. Santoni De Sio	Delft University of Technology, copromotor
Independent members:	
Prof.dr. G. Wang	Fudan University
Prof.dr.ir. N. Bharosa	Delft University of Technology
Prof.dr.ir. I.R. van de Poel	Delft University of Technology
Dr. P.J. Nickel	Eindhoven University of Technology
Dr.ir. J.A. Pouwelse	Delft University of Technology

This work is supported by the Chinese Scholarship Council, grant number 201606060133.

Keywords: trust, trustworthiness, digital ethics, blockchain, reputation-based platforms, digital contact tracing technologies

Printed by: Ipskamp Printing, Enschede

ISBN 978-94-6384-282-2

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

We inhabit a climate of trust as we inhabit an atmosphere
and notice it as we notice air
only when it becomes scarce or polluted.

Annette Baier

Contents

1

Introduction

1.1. Trust in the digital context

Trust this computer? Similar windows pop up on your smartphone when you connect it with a laptop or other devices. As you are moving your hands to choose between yes or no, a string of interrelated questions may flash through your mind naturally,

a). Who or what am I trusting here? This question considers the targets of trust, which might be the computer as the direct target and/or the institutions (e.g., technology corporations and government agencies) and/or individuals (e.g., designers, engineers, and leaders of corporations) as the indirect targets relevant to this computer's life cycle.

b). What are the reasons for giving trust? This question asks why one needs to trust, considering the practical interest, benefits, and value of trusting the relevant targets, as well as epistemic reasons built on the actual trustworthiness of the associated trustees (those who receive trust).

c). What can I lose by giving trust? This question shows the inherent vulnerability of being a trustor (the one who gives trust) – i.e., the risk of losing the valued entrusted things and facing other knock-on effects.

The above inquiries roughly sketch the main general considerations of trust as a topic of philosophical interest, and the situations where trust becomes most relevant as a way to facilitate interactions under vulnerability and uncertainty (Becker 1996; Heimer 2001). Many valuable insights into the descriptive, explanatory, and normative aspects of these questions have already been achieved when the object

This chapter is based on the following article:

Teng, Y. (under review). Warranted trust in the context of trust-inviting digital systems. *Ethics and Information Technology.*

of trust is a particular person, which is widely acknowledged to be the original form of trust (Luhmann 1979; Yamagishi 2011). In the most general sense, interpersonal trust can be considered as a phenomenon generated within a relation that at least involves two parties: a trustor (x) and a trustee (y). By trusting, x is willing and feels optimistic to rely on y in certain domains while cannot monitoring, fully predicting, or controlling the behaviour of y, and thus x is vulnerable to y's discretionary power relevant to the fulfilment of the entrusted thing (Baier 1986; Gambetta 1988; Hardin 2001; Alfano 2016; Taddeo 2010). For this risky nature, trust has been described as a result of balancing between confidence and vulnerability of relying on a particular target (Werbach 2018), and is often distinguished from rational, strategic reliance by its unique explanatory role in enabling one to take a "leap of faith" and form interactions under uncertainty (Möllering 2006; Nickel 2013; Nickel 2020).

Given the inherent risk of placing trust, the philosophical question of when trust is warranted – i.e., plausible, justified, and well-grounded – is of central importance in trust practices and discourse. It requires to examine whether the conditions for trust to be relevant are met, whether displaying trust in a given context is justifiable from the perspectives of epistemology and a sense of value, and whether the trusted person is indeed trustworthy (McLeod 2015). While this makes the competence of the trustee a clear condition for warranted trust to exist and being a trustworthy person, many philosophers also argue for the importance of allowable non-cognitive, motivational factors (that can underlie one's trust-keeping behaviour) – such as goodwill (Baier 1986) and moral integrity (McLeod 2002), an attitude of trust-responsiveness (Jones 2012; Faulkner 2007), and responsibility for satisfying different norms and values in context (Walker 2006; Jones 2004) – for being a necessary component of trustworthiness and a distinctive feature of what we seek from trusting others.

In the digital age, the function of trust as a way to reduce complexity and encourage interaction and adoption is evidently crucial (Luhmann 1979; McKnight 2011; Bahmanziari et al. 2003; Choi and Ji 2015). However, new challenges and complexity are present to make warranted trust decisions and avoid being gullible, credulous, and naïve, especially considering the proliferation of socio-technical systems that take trust as part of the design goal. As a hybrid system constitutive of technical artefacts, human agents, and social institutions, a socio-technical system could be conceived as a complex people-containing system (Kroes et al. 2006; Franssen 2015). Given this hybrid nature, an overt aspect that makes trust in the digital age more complex is that, unlike a human trustee, trust directed to, or mediated by, a socio-technical system involves more parties – e.g., corporations, government agencies, professionals, and individual actors – that are pertinent to the design, development, and deployment of the system, whose interests and decisions can impact one's trust. Equally important, the peculiarities pertaining to complex digital systems, which are hard to understand for people without a technical background, exacerbate the power imbalance that already exists between trust givers and trust receivers.

To conclude, regarding the two-sided effects of placing trust, as Floridi (2016) and

Taddeo (2017) point out, it is significant to envision and enact correct strategies to trust, and harness the value of, digital technologies while protecting and preserving what we value of information societies. And, the difficulties brought by the above two aspects indicate that, for arriving at warranted trust in the digital age, we need to take a look at the distinctive features of trust against the backdrop of specific system engaged, particularly with respect to how we could deal with conflicting values of stakeholders and making good use of the system's peculiarities.

1.2. Ethics and philosophy of trust-inviting systems

While we could say that trust in the digital context is of general ethical and practical importance, this thesis focuses on discussing trust issues and the relevant ethical concerns arising in the context of socio-technical systems that are explicitly designed to foster and make use of the value of trust. In this thesis, this sort of systems is called *trust-inviting systems* and can be understood in two ways. In a narrower sense, trust-inviting systems refer to those systems that take the use and promotion of trusted interactions among network of peers, professionals, strangers, and authorities as the dominant function of the system. Think of Airbnb and credit-reporting agencies that build all the business on the idea of design for overcoming natural social bias and trusting people who are (near-)strangers (Gebbia 2016; Abrahao et al. 2017). In other words, the systems can barely function without the basis of trust provided. In a broader sense, trust-inviting systems refer widely to those systems that contain design objectives and cues of eliciting or nudging trust, such as open government initiatives (O'Neill 2004), anthropomorphic robots (Coeckelbergh 2012), intelligent user interfaces used in autonomous vehicles (Ruijten et al. 2018).

From a holistic view, this thesis holds that, due to the revolutionary features trust-inviting systems bring to trust relations, such systems can be regarded as a special sort of socio-technical system that raises special ethical issues deserving systematic reflection. Indeed, we have heard a lot about the trust revolutions made by systems like Airbnb and blockchains, but what essentially makes them revolutionary? One important precondition for x to feel optimistic about y's competence and commitment is familiarity filled with x's knowledge and previous experience related to y (Luhmann 1979). This explains why friends, family, small groups, and traditional villages are often seen as the collectives where cooperation both prevails and is confined within the insiders (Yamagishi 2011). The gap left by unfamiliarity or a society of strangers is exactly the place where reputation-based platforms seek to bridge. By aggregating reliable information and formalizing activities on both sides, these platforms play an intermediary role between participants, leading people from the natural tendency of trusting homophily and the rooted stranger-danger bias to connect and collaborate with individuals who are known – if at all – only by reputation. At a general level, reputation can be understood as a collection of direct or indirect experiences towards an entity (Thiessen 2013). It relies heavily on the information about the past behaviour of the entity and is often seen as an important signal of what that entity is likely to do in the future. As a matter of fact, systems curating

reputation information are changing the way we interact with each other on a scale incredible even a decade ago. Consider, for example, the wide wave of behemoth sharing-economy platforms – like TaskRabbit, Airbnb, Uber, BlaBlaCar, and Didi – and credit-reporting agencies – such as Equifax, Transunion, Experian, and Public Bank of China that both aim to facilitate trust-based interactions in certain areas. With these systems, trusted connections are able to emancipate from closed tight-knit communities to an open society, leading direct reciprocity to indirect reciprocity on which more sophisticated cooperation mechanisms can be built (van den Hoven et al. 2019). It is no exaggeration to say that effects of the trust revolution taken place in the landscape of the alike systems have already been infused into most citizens' lives, bringing together institutions, entrepreneurs, consumers, investors, and other individuals who would otherwise not engage with (Botsman 2017; Werbach 2018).

Furthermore, some other technological systems, exemplified by decentralized blockchains, seem to bring the trust revolution to a novel stage. As the decentralized database technology behind Bitcoin, blockchain is originally designed to replace the role of third-party authorities (e.g., banks) in facilitating online transactions between heterogeneous groups of participants. It achieves this by containing several important properties – immutability, accessibility, authentication, and verifiability – core to a record-keeping device provided traditionally and almost exclusively by centralized institutions. In blockchain-enabled interactions, what people rely upon is instead the system itself that functions as an institution-like, decentralized database and the network of peers and developers involved. From this single perspective, trust in blockchain systems underpinned by the systems' technical infrastructure might be considered more morally acceptable given that the systems retain the value of trust for the trustor while eliminating the discretion of the whims of bureaucrats and malicious users, which fundamentally reduce the relevant risk and uncertainty involved in the whole interaction.

In sum, from behemoth reputation-based and bureaucratic platforms to decentralized blockchains that take a further step by attempting to replace the role of giant institutions, trust-inviting systems are opening up innovative avenues that can facilitate trusted interactions in an open society with faceless commitments. These technological achievements resonate with Nickel's (2020) clarification of the two trust-related goals of engineers in design practices respectively. Namely, the goal of using technological design to shape trust between people and the goal of creating automated technologies that take over the need for humans when fulfilling certain tasks and are meant to elicit trust directly.

However, leveraging trust, especially at such a substantial level, can lead to negative moral effects if the design and implementation of these systems are shown to be problematic. This is not simply about the participants' inherent vulnerable position that might be taken advantage of. Nor simply about how the trust built through the system might be frustrated after the fact and result in potential loss. This concerns in particular the double-edged technological capacity enabling the revolutionary as-

pect of these systems. Ranging from the worrying fact of creating tech monopolies that in some cases people have little freedom of choice but have to rely on their services, to the privacy concern over the incredible amount of personal data controlled by these giant institutions. Also, such technological capacity is confronted with accusations that the measures of reputation adopted by such systems are able to steer participants' behaviour and decision-making in a profound but sometimes hidden and morally problematic way. This indicates how such systems might easily exploit and manipulate users' trust with respect to the specific entrusted things and the protection of personal data and other human rights.

After all, inviting trust is itself a morally sensitive issue given that to trust is to expose oneself to the discretion of others and the possibility of being harmed. And, to promote trust, even at the minimum level, is to deliberately promote others to give such discretion and put them in a vulnerable and imbalanced position in which they may otherwise not reside (Baier 1986). This feature of trust-inviting systems arguably multiplies the moral obligation of the relevant parties to take possible measures to envision, assess, regulate, improve, and signal the systems' trustworthiness and thus reduce the gap left by the uneven power distribution between the interacting parties in the first place. The above discussion makes clear the revolutionary aspects and the resulting ethical concerns related to trust-inviting systems with respect to the generation of secure trust, supporting the argument that trust-inviting systems should be regarded as a special sort of systems deserving systematic reflection.

Based on these considerations, this thesis focuses on the question – how can trust-inviting systems foster trust in an appropriate way? The question is of practical importance not only for individuals who hope to make wise trust decisions and those who yearn for earning and restoring trust at issue, but are also generally important for designers, entrepreneurs, regulators, policymakers, and the public to make collective efforts to help prevent undesirable consequences potentially brought by trusting flawed systems. With such a wider audience in mind, ethics and philosophy of trust-inviting systems should not just focus on the impacts and implications of the systems on those who perform as the direct trustor but also on the society at large as the relevant but indirect trustor who would be affected by the systems.

1.3. A categorization of trust-inviting systems

To study the above research question more systematically, this thesis classifies trust-inviting systems into three broad categories in terms of the different sorts of targets that the trust enabled by a given system is primarily directed to. The three categories are: systems inviting interpersonal trust, systems inviting institutional trust, and systems inviting technology trust. The main chapters will then use specific and representative cases of each category to take a closer look at issues of trust that are of ethical importance, with the ultimate aim of improving trustworthiness and making different types of trust more warranted. It is worth noting that these three situations are not mutually exclusive. In fact, they are often closely

intertwined, and hence a system can elicit, for example, both interpersonal trust and institutional trust. Nonetheless, these situations concern distinctive trust forms, and the ways that the systems use to foster different trust forms are also not the same.

In general, a categorization of these systems based on the trust forms they enable could help clarify and assess the specific issues emerging in context. Trust, as a rich and multi-faceted phenomenon, is highly relational and contextual, and thus the specific conception of trust used for analysis is largely contingent on the example chosen and the research perspective used (Simon 2013).[1] For example, it makes a difference to the reasons and vulnerabilities involved when we theorize and characterize trust in intimates and trust in strangers. While the former relationship is more motivated by affective factors such as love and goodwill, mutual expectations of distant relationships in an open society seem to be more normative, holding that people ought to behave in accord with different norms and standards required by the context while weakening the emphasis on their specific motivations (Simpson 2012; Walker 2006). This normative expectation is one kind of logic behind the first category of trust-inviting systems – systems used to mediate *interpersonal trust*. Typified by reputation-based platforms canvassed above, this sort of systems is commonly applied to cases when participants are not familiar with or do not trust each other at the outset. Here the main targets of trust are individual participants of the systems. But mere normative expectations do not make one's trust secure. Another important logic that underlies these platforms is thus to improve the predictability of the future action of participants by providing more evidence of their past behaviour as well as introducing more constraints for potential violation behaviour, which both bring the systems into continuous services.

Secondly, trust-inviting systems can also mediate *trust in institutions* "behind" the systems. This is not limited to the institutional trust invited by reputation-based platforms, which is often regarded as a precondition for participants to use their services.[2] More typical cases of this category are websites, apps, and other digital products of governments and corporations that take trust as an explicit design goal, such as open government initiatives and websites with a TRUSTe seal.[3] Unlike the interpersonal trust discussed above, trust in these cases is not obviously or necessarily placed in a particular person but in unnamed role holders and social groups relevant to the design and implementation of the systems. This makes

[1] Different disciplines often study trust from distinct perspectives and emphasize different explanatory aspects of this concept. For example, economists study trust with an eye to improving overall economic functioning (as measured by, for example, per capita GDP). They tend to focus on the calculative and consequential dimensions of trust. Sociologists view social trust as a structural feature of a community, while psychologists tend to examine trustingness and gullibility as traits of individual persons. And ethicists care more about the conceptualization, moral and prudential value, and epistemology of trust and trustworthiness. These perspectives can be used as complementary approaches to assess the trust relation in question.

[2] This point is opened up in Section 1.4.

[3] TRUSTe is a privacy assurance and certification program that helps organizations to demonstrate regulatory expectations. https://trustarc.com/consumer-info/privacy-certification-standards/.

characteristics and the associated mechanisms of institutional trust both different from-, and sharing some similarities to, trust between individuals. On the one hand, as Hardin (2002) points out, trust in institutions and officials are not analogous to trust in a particular person since knowledge demanded by interpersonal trust – such as motivations and fame of the potential trustee and their past actions with others – are usually unavailable to regular people. On the other hand, people do care and act on the basis of their trust and distrust (perhaps more often) in institutions in the senses relevant to organizational interests, responsibility, and generalized personal benefits of power/role holders, as well as the reliability of the framework that regulates how institutions and their representatives work. With the mediation of digital systems as an indirect way to communicate with the public, impacts and implications of all the above factors can be enlarged and obscured, which may lead to unwarranted trust as well as unwarranted mistrust. Such specificities make technology-mediated institutional trust a distinctive trust form in the digital age.

Lastly, as discussed earlier, blockchain systems neither fully fall into the category of systems inviting interpersonal trust nor the category of systems inviting institutional trust given that blockchain-enabled trust revolution is based on a rather different idea that seeks to render malicious actions of individuals infeasible and the role of centralized institutions unnecessary by a set of technical peculiarities. The design of the original blockchain thus makes it the third sort of trust-inviting systems – systems that primarily invite *technology trust*. Unlike interpersonal and institutional trust, the idea of trust in technological artefacts is traditionally controversial due to their lack of agency and consciousness required by most philosophical accounts for trust to be a plausible and meaningful, rather than a trivial vernacular, concept. (Friedman et al. 2000; Pitt 2010; Cook 2010).[4] However, blockchains' potential for performing as a self-sufficient database without the need for any third-party authority indicates the systems' mixed role as both a technological system for achieving functional services and an institution-like entity that can organize relatively stable patterns of social practices (see Chapter 5). This special role of blockchains, on the one hand, shows the importance and possibility of grasping blockchain trust in terms of what we understand of institutional trust that the systems intend to replace. On the other hand, it implies that blockchain trust, though grounded in the mechanism of institutional trust, goes far beyond trust facilitated by institutions by removing the involvement of third parties while retaining the value of trust for the trustor. Given these considerations, blockchain-based technology trust is arguably a unique and intriguing form of trust.

The categorization above sketches the main types of trust that trust-inviting systems are enabling. Simultaneously, the different conceptions of trust and application examples involved show that the idea of "trust-inviting system" is inherently neutral and generic. Such an idea is flexible enough to accommodate different types of trustees and different understanding of trust that are both context-dependent while at the same time being precise enough to express the systems' common design ob-

[4]More detailed discussion about this claim is provided in Chapter 4.1 and Chapter 4.4.

jective of fostering trust. In other words, rather than assuming and relying on a definite conception, the idea of trust-inviting systems relatives what we understand of trust by making the trust conceptualization process context- and relation-based. Before going into a detailed discussion about the specific cases and conceptions of trust we use to explore detailed issues under these trust types – including China's Social Credit System (SCS, for interpersonal trust), digital contact tracing technologies (for institutional trust), and blockchain systems (for technology trust), in what follows, an explanation of the analytical approaches used throughout this thesis for pursuing an improvement of trustworthiness and proper trust is provided.

1.4. Justifying proper trust in context

While the development and adoption of advanced digital systems are unlocking the value, and broadening the circle, of trust, trust encouraged by trust-inviting systems should not be unreflective and uncritical. To avoid undesirable consequences made by misplaced trust, standards for evaluating others' trustworthiness have to be justified. More importantly, individuals, organizations, and corporations who seek to earn trust should reliably justify whether they are fostering trust in an appropriate manner and be trust-deserving. But how can we achieve such proper trust? Notions like trust and trustworthiness derive their meaning from different given contexts (van den Hoven 1997). Therefore, there exists no overarching strategy ready to be applied to different trust cases to understand what is at stake and to inform the design and development processes accordingly. Nevertheless, there are at least three intertwined approaches to be used as a starting point for a reflection on warranted trust, including insights from computer ethics, conditions for trust created by prescribed rules and procedures, and context-sensitive views derived from given contexts. As the last approach receives relatively less attention in the discourse and the impact it has on addressing concrete trust issues is significant, this thesis emphasizes its role in approaching proper trust in the digital age.

The first approach is based on the idea that "trust is built on competence and ethics", which is also the headline for Edelman Trust Barometer for 2020. According to the firm's long-lasting investigation on trust, effectiveness (e.g., whether an institution is generally good at what it does) and ethical conduct are considered as the two core elements vital to any trusting relation. They use sub-indicators, including driven purpose, honesty, vision of future, and fairness, to define and measure the ethical conduct of institutions in the age of technologies (Edelman 2020). Overall, this idea of trust sits well with the approach to warranted trust discussed in this thesis, consisting of both cognitive and non-cognitive aspects. However, the Edelman Trust Barometer lacks a nuanced explanation of the ethical use of technologies. This shortage can be alleviated with the aid of computer ethics.

Computer ethics emphasises invisibility as one general factor of computers bringing extra vulnerability and moral complexity. As Moor already pointed out in 1985, the invisibility of computers is a blessing for improving efficiency by saving incredible time and energy of human operators, but it is of crucially moral significance, in

three ways. First, using invisible operations for nefarious purposes makes people open to abuse of power and capacity that is often difficult to be detected but can cause real harm. Second, the presence of opaque, implicit, and flawed normative assumptions infused into the operation of computer systems, either intentionally or unintentionally, can generate far-reaching impacts upon individuals and society. Third, decisions made by computer calculations, though sometimes are less fallible and less dangerous than those made by power holders – such as in the case of deciding when to launch a nuclear weapon, are not always apprehensible by human intelligence but can bring negative consequences once they are shown to be wrong. These three ethical dimensions of computer technology show the importance of the justification for the purposes, assumptions, and methods of processing used by computer systems. Trusting computers, as Moor concludes, should be grounded in the formulation of policies that could help apply our moral opinions to guide the use of these efficient yet invisible systems and thus protect, preserve, promote what we consider important in life.

The first approach grounded in competence and ethics, seems to suggest that proper trust may be achieved via justified purposes, assumptions, and methods of processing and the pursuit of more transparency and reliability. From the perspective of this approach, these aspects could be regarded as the basis and evidence for trust to thrive. Indeed, such an approach is helpful to identify common issues shared by computer-enabled systems as well as to define some ways to address these issues. Consider, for example, the transparency agenda energetically pursued to overcome the intangible functioning of computers and the vagueness and complexity of institutional procedurals (O'Neill 2004). However, a call for more justification and transparency is not sufficient to grant warranted trust, to the extent that this is not compatible with high stakes and complex interactions ubiquitous in modern societies. More robust measures to ensure socially and morally desirable outcomes that can be trusted by our society are required.

A second, alternative approach would be that of imposing more precise conditions to make trust within a particular environment less dangerous and vulnerable (Cook 2001). This approach is currently already realised by institutional structures which facilitate interactions and cooperation – e.g., government-issued laws, bureaucratic rules, contracts, insurance, and compensation. Most philosophers interpret these mechanisms as alternatives to trust since they are able to bring about cooperative behaviour while lowering the demand of trust. Thompson (2000) and Kerasidou (2016), for example, refer to these measures as an audit agenda that seeks to utilize a higher level of control, monitoring, and inspection to guarantee compliance and accountability. Trust, in this case, is seen as a by-product of a good economic system or an unnecessary attitude that one should not be bothered (Elster and Moene 1988).

It is fair to say that the above approach is probably the most standard and realistic way to address the imbalanced power distribution between the interacting parties so as to facilitate cooperation in modern society (Gambetta 1988). However, there are

at least three cases where it may not be possible to follow this approach: (1) where these prescribed procedures and practices are not applicable or not compatible, such as the vacuum of law and policy in some cutting-edge fields and the gap left by insurance and compensation that both intend to reduce the loss caused by deviation behaviour to monetary level, (2) where finding enough evidence and information about these measures requires so much time, energy, and resources that make such behaviour costly, difficult, and less attractive especially under time pressure (Nickel 2013; Gambetta 1988), (3) where these arrangements simply fall short of facilitating adoption and cooperation. The requirement for "accepting the privacy and cookies policy" of a new website before browsing is a simple case where the rules of the agreement are transparent and abundant but seem not doing much for promoting connections. Instead of checking the rather lengthy and obscure details, most of us may just click "accept" routinely or even close some websites by seeing this notification as a privacy alert. Rather than restoring trust, as argues O'Neill (2004), increased demands for accountability and benchmarks often "economize on trust" while escalating the atmosphere of distrust they intend to bring about in the first place.

These first two approaches, coupled together, seem to represent the mainstream understanding of trustworthiness. Consider, for example, the High-Level Expert Group's definition for trustworthy AI as a confluence of lawfulness, ethics, and robustness (European Commission 2019). A critical shortage of this definition is that they seem to jump from the trust-establishment process – i.e., to understand what constitutes trust in different contexts – to the trust-evaluation process (Gille et al. 2020; Marsh et al. 2020; Åm 2011). But trust should neither be conflated with trustworthiness, nor is it simply a by-product. Trust as a distinctive concept is infused with subjective, relational, and contextual factors that may consider concrete social dynamics beyond the scope of any presupposed framework. Thus, something that is, or appears to be, trustworthy under a top-down, authorities-based institutional framework does not mean that it will be trusted and adopted in reality by different people. Also, past evidence cannot eliminate future risk and contingencies, and there might be other competitive options at the user's disposal. The effect of trust deficit is especially apparent and troublesome when collective effort is urgently required for the implementation of public policies in extreme situations. During the COVID-19 pandemic, the effect of the gap between trust and trustworthiness has been apparent. Many people have expressed their distrust and privacy anxiety towards the so-called privacy-preserving contact tracing technologies and the governmental agencies and private companies behind them, even though many privacy experts and healthcare authorities have endorsed the apps' privacy-friendly settings (O'Halloran 2020).

Based on these considerations, this thesis holds the position that debates on warranted trust in the digital age should include a third approach that addresses a systematic focus on understanding the meaning and value of proper trust within different given contexts. This context-sensitive conceptual approach of trust is a pragmatic endeavour and considers, first and foremost, the social dynamics of a

particular situation. It is thus an inquiry into what the trust relation at issue is essentially about i.e., why and how a particular entity might be established as a plausible target of trust and where distrust may appear. To be more specific, we need to know what the potential trustee has on offer; why trust is relevant and valuable for certain groups of people in the first place; Are there any conceptual and related practical issues that may create implications negative to the intended trust relation? How can these issues be addressed, for example, by a recalibration of how we understand the trust relations in question?

Reflections on these issues contribute to identifying, and moving stakeholders to address, what is at stake in the current, detailed design and implementation processes of the systems from the perspective of philosophy of trust. Using the trust construct in this way means that we will not try to develop an overarching understanding of trust and apply it once and for all. Instead, with the situational and contextual nature of trust in mind, different assumptions and aspects of trust are arguably most useful for analysis when they are utilized in accord with the trust relation enabled by a particular system in context. Also, they should be constantly recalibrated in terms of the changing of what is considered valuable for the relation, with proper trust as a touchpoint, manifesting what is needed for a suitable definition of trust in context and shaping the relation at issue towards moral acceptability and societal desirability. While the other two approaches discussed remain necessary and important in the discourse, this context-sensitive approach is the main methodology insight emphasized by this dissertation when doing ethics and philosophy of trust in the digital age.

1.5. Overview of dissertation

Following the categorization of trust-inviting systems presented above, this thesis focus on discussing specific cases of systems inviting interpersonal trust (chapter 2), institutional trust (chapter 3), and technology trust (chapters 4 and 5). By reflecting on several present cases of trust-inviting systems that are experiencing great tension of trust, including China's SCS, digital contact tracing technologies, and blockchain technology, this thesis argues that (1) trust-inviting systems essentially attempt to interpret, translate, and ultimately institutionalize the idea of trustworthiness in given contexts; (2) however, the ways that trust-inviting systems are using to institutionalize the characteristics of trustworthy persons, institutions, and technologies should not be accepted without scrutiny. For each case analysed by this thesis, a discrepancy between the intention to improve trust and trustworthiness and the means that are adopted to facilitate them is shown. Such cleavage is argued to be primarily caused by flawed understanding of the trust concepts and the resulting ill-suited design choices, as well as problems emerged from the implementation process; finally, (3) these issues are proposed to be ameliorated by a recalibration of the understanding of the trust concepts, which has the potential for remedying shortcomings of the current design and development of the systems with forward-looking strategies taking into account a wide range of needs, societal values, and technical properties of the systems. In a word, it is argued

that trust-inviting digital systems should be designed, developed, and deployed in ways that are aligned with the essence of the trust relation in context, in order to achieve proper trust and trustworthy systems. As such, the pitfalls identified in each case are used as perspectives that contribute to building affordances that foster warranted trust and foreclosing affordances that would undermine warranted trust.

Chapter 2 focuses on interpersonal trust, and the guiding question of this chapter is: can the means currently adopted by China's SCS achieve the moral ends set by the overall project with respect to the promotion of social trust and trustworthiness? The chapter provides a critical look at the SCS and its goals to foster social trust and people's trustworthy behaviour with its three pillars including the credit reporting system, the joint punishment and reward system, and local implementation of the SCS. This national project has raised the issue of interpersonal trust shaped by trust-inviting systems. Essentially, the SCS could be seen as a special type of reputation-based platform that helps bridge the gap between two parties who are (near-)strangers by curating the participants' reputation information in certain aspects.

Similar to other reputation-based platforms, on the one hand, the SCS provides ways to judge a stranger's capability to predict that person's likely future behaviour based on the exposed past records before making a trust decision. On the other hand, standards adopted by the SCS navigate its participants to behave in accordance with the system's preferences. This means benchmarks adopted by reputation-based platforms have the power to define a trustworthy X (i.e., a debtor, a host, a restaurant, a driver, etc.), and consequently, within the platforms, high-scoring users are almost always considered good and more trustworthy than those with lower scores. Hence, if participants want to pursue higher scores and benefit from the systems, their online and/or offline actions will be directly steered by the systems' standards. This makes clear the importance of reliable systems and benchmarks as a precondition for eliciting secure trust since people might be led in the wrong direction if the standards turn out to be irrelevant or inappropriate. In this respect, it can be said that although participants of the SCS and some analogous systems (e.g., Equifax and TransUnion, to some extent) ultimately interact with each other, people often first place their trust in the credibility of the corporations, financial institutions, and government agencies that create those platforms. In cases when participants have little room to choose whom to engage with at the first place, like Didi and Uber, it seems fair to say that people trust the platforms more than the interacted individuals. In such cases, what is relied upon by the participants are mostly the formal endorsements and indemnity provided by these giant platforms, which could to varying degrees protect both sides' interests and vulnerability beyond the actual performance of the ultimate interacted parties.

The particularities of the SCS lie in its unprecedented comprehensiveness and state-driven invasiveness. In short, the SCS could be understood as a governance approach that attempts to improve moral trust relations by institutionalizing the rough

idea of "being a trustworthy citizen." Along with the proliferation of local implementations, however, the rules and principles adopted for identifying "trust-breaking" actions under the SCS are fraught: the domain of application seems to be almost unlimited, and there is little coherence to the set of behaviors and omissions that are included. This chapter takes a critical look at the SCS with a focus on whether its current initiatives can achieve the moral ends set by the overall project. We explore this question at two levels: assumptions and applications. First, based on an analysis of the implicit assumptions about trust underlying the SCS, we argue that the current SCS primarily facilitate the instrumental and prudential aspects of trust, showing a discrepancy between the moral objective intended to be achieved and the ways adopted to approach it. We then focus our reflection on three pilot cities' scoring systems built on the rational trust assumptions, further clarifying the detailed ethical issues associated with the design and audit of these applications. These issues arguably lie at the core of ethical discourse on the SCS and may lead the applications to deviate from the overall moral goal. As such, trust issues in the context of the SCS are discussed from the perspective of current initiatives' trustworthiness, with the project's moral goal as a touchstone, manifesting what is needed to move the initiatives to build moral trust relations.

Chapter 3 focuses on institutional trust, and the guiding question of this section is: what are impacts and implications of institutional trust on privacy issues of digital contact tracing technologies? The chapter offers a constructive analysis of the importance and implications of institutional trust in the context of digital contact tracing technologies to control the spreading of SARS-CoV-2. As a supplement to conventional tracing measures, digital technologies based on instant signals of smartphones promise to improve the efficacy of the tracing processes by minimizing the time to find, notify, and quarantine the contacts at risk, contributing to improving the effectiveness of the so-called "test-tracing-isolate" strategy for exiting lockdowns (Abueg et al. 2020; Hinch et al. 2020; Panovska-Griffiths et al. 2020; Ferretti et al. 2020). However, proper implementation of this information-based solution should deal with participants' privacy vulnerability and the uncertainty from the relevant institutions' side, among others. This chapter proposes to understand the current approaches for preserving privacy, referred to as privacy by legislation and privacy by technological design, as distrusting strategies that primarily work to reduce participants' vulnerability by specifying and implementing privacy standards related to this digital solution. It points out that mere distrusting strategies are insufficient for ethically appropriate development of this digital solution, nor can they eliminate the need for institutional trust that plays an essential role in fostering voluntary support for this solution. To reach well-grounded trust in both an ethical and epistemological sense, this chapter argues that trust in institutions concerning personal data protection in the case of digital contact tracing ought to be built on the relevant institutions and individuals' goodwill towards the public and their competence in improving the actual effectiveness of this solution. It concludes by clarifying three aspects, including the purpose, procedure, and outcome, where the relevant trustees can work to signal and justify their intentions and increase

their trustworthiness. Given the complementary qualities shown by the distrusting strategies and trusting strategies, a combined strategy including both sorts seems closer to what we expect from the responsible implementation of this digital solution, which could also improve the effectiveness of this institutional response.

Chapter 4 and 5 focus on technology trust, and the respective guiding questions of these two sections are: what does it mean to trust blockchain technology? And how should we understand trust engaged with blockchain-enabled virtual institutions? They take blockchain technology as the case study and provide a systematic reflection on how blockchain trust could be established, evaluated, and improved. Compared to the previous chapters, these two chapters take a step further to understand technology-mediated trust in the context where the arising of trust does not rely on authoritative, reliable institutions or goodwill of individuals. As such, a comprehensive ethical reflection on three typical forms of trust shaped by trust-inviting systems is provided with detailed perspectives and forward-looking suggestions from which the trustworthiness of the systems studied is able to be improved.

Blockchain came to prominence as the distributed database technology behind Bitcoin that enables continuous transitions of system states between participants without the need for a trusted intermediary (e.g., a bank). For this reason, there is a widespread belief that user-blockchain interactions is trust-free or trustless. Chapter 4 argues that such a belief is inaccurate and misleading since it not merely overlooks the vital role of trust in facilitating interactions especially when one lacks knowledge and control but also conceals the moral and normative relevance of relying on blockchain applications. It reaches this argument by providing a comprehensive analysis of the phenomenon referred to as trust in blockchain technology or blockchain trust, clarifying the trustor group, the structure, the normatively loaded nature, and the risk of this form of trust relation. To understand the rich expectations people hold toward blockchain-based systems, the crucial role played by the appropriateness granted to the normative ideas built into the system is highlighted. Given the vulnerable position of the trustor, the actual trustworthiness of blockchain applications for realizing these values should be scrutinized before the placement of trust. With such concern, the chapter ends by critically reflecting on two of the most promising values that can invite users' trust in blockchain technology, arguing that trust built on these values is risky and unjustifiable due to the moral and technical limits involved in current blockchain applications.

Following the broad idea that blockchain trust is normatively loaded, chapter 5 proposes an in-depth analysis of the nature of blockchain trust based on an analogy with trust placed in institutions. In support of the analysis, a detailed investigation of institutional trust is provided, which is then used as the basis for capturing the nature and ethical limits of blockchain trust. Two interrelated arguments are presented. First, given blockchains' capacity for being institution-like entities by inviting expectations similar to those invited by traditional institutions, blockchain trust is argued to be best conceptualized as a specialized form of trust in institutions. Keeping only the core functionality and certain normative ideas of institutions, this

technology broadens our understanding of trust by removing the need for third parties while retaining the value of trust for the trustor. Second, the chapter argues that blockchains' decentralized nature and the implications and effects of this decentralization on trust issues are double-edged. With the erasure of central points, the systems simultaneously crowd out the pivotal role played by traditional institutions and a cadre of representatives in meeting their assigned obligations and securing the functional systems' trustworthy performances. As such, blockchain is positioned as a technology containing both disruptive features that can be embedded with meaningful normative values and inherent ethical limits that pose a direct challenge to the actual trustworthiness of blockchain implementations. Such limits are proposed to be ameliorated by facilitating a shift of responsibility to the groups of people directly associated with the engendering of trust in the blockchain context.

References

Abrahao, B., Parigi, P., Gupta, A., & Cook, K. S. (2017). Reputation offsets trust judgments based on social biases among Airbnb users. Proceedings of the National Academy of Sciences, 114(37), 9848-9853.

Abueg, M., Hinch, R., Wu, N.et al.(2020). Modeling the combined effect of digital exposure notification and non-pharmaceutical interventions on the COVID-19 epidemic in Washington state. medRxiv. Accessed 10 October 2020. https://www.medrxiv.org/content/10.1101/ 2020.08.29.20184135v1.

Alfano, M. (2016). The Topology of Communities of Trust. Russian Sociological Review 15(4): 30-56.

Åm, T. G. (2011). Trust in nanotechnology? On trust as analytical tool in social research on emerging technologies. NanoEthics, 5(1), 15-28.

Bahmanziari, T., Pearson, J. M., & Crosby, L. (2003). Is trust important in technology adoption? A policy capturing approach. Journal of Computer Information Systems, 43(4), 46-54.

Baier, A. (1986). Trust and antitrust. Ethics, 96(2): 231-260.

Botsman, R. (2017). Who can you trust: How technology brought us together and why it might drive us apart. Hachette UK.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. International Journal of Human-Computer Interaction, 31(10), 692-702.

Coeckelbergh, M. (2012). Can we trust robots? Ethics and information technology, 14(1), 53-60.

Cook, K. (2001). Introduction. In K. S. Cook (Ed.), Trust in Society. New York: Russell Sage Foundation.

Edelman. (2020). Edelman Trust Barometer 2020. Accessed December 17, 2020. https://cdn2.hubspot.net/hubfs/440941/Trust

Elster, J., & Moene, K. O. (Eds.). (1989). Alternatives to capitalism. Cambridge University Press.

European Commission (2019). Ethics Guidelines for Trustworthy AI. Accessed June 6, 2020. https://ec.europa.eu/futurium/en/ai-alliance-consultation.

Faulkner, P. (2007). On telling and trusting. Mind, 116(464), 875-902.

Floridi, L. (2016). Tolerant paternalism: Pro-ethical design as a resolution of the dilemma of toleration. Science and Engineering Ethics, 22(6), 1669-1688.

Franssen, M. (2015). Design for values and operator roles in sociotechnical systems. In J van den Hoven, PE Vermaas, I van de Poel (Eds.), Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains, 117-149.

Gambetta, D. (1988). Can we trust trust? In Gambetta, Diego (ed.), Trust: Making and breaking cooperative relations, 213, 214. Oxford: Basil Blackwell.

Gille, F., Jobin, A., & Ienca, M. (2020). What we talk about when we talk about trust: Theory of trust for AI in healthcare. Intelligence-Based Medicine, 1, 100001.

Halloran, J. (2020). Consistent trust gap in contact-tracing apps in US, Europe. Computer Weekly. Accessed 4 September 2020. https://www.computerweekly.com/news/252486058/Consistent-trust-gap-in-contact-tracing-apps-in-US-Europe.

Hardin, R. (2001). Conceptions and explanations of trust. In K. S. Cook (Ed.), Trust in Society (pp. 3–39). New York: Russell Sage Foundation.

Hardin, R. (2002). Trust and trustworthiness. Russell Sage Foundation.

Jones, K. (2004). Trust and terror. In P. DesAutels & M. U. Walker (Eds.), Moral psychology: Feminist ethics and social theory (pp.3-18). Maryland: Rowman & Littlefield.

Jones, K. (2012). Trustworthiness. Ethics, 123(1), 61-85.

Kerasidou, A. (2016). Trust me, I'ma researcher!: The role of trust in biomedical research. Medicine, Health Care and Philosophy, 20(1), 43-50.

Kroes, P., Franssen, M., Poel, I. V. D., & Ottens, M. (2006). Treating sociotechnical systems as engineering systems: some conceptual problems. Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research, 23(6), 803-814.

Luhmann, Niklas (1979) Trust and power. Chichester: John Wiley.

Marsh, S., Atele-Williams, T., Basu, A., Dwyer, N., Lewis, P. R., Miller-Bakewell, H., & Pitt, J. (2020). Thinking about trust: People, process, and place. Patterns, 1(3), 100039.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. ACM Transactions on management information systems (TMIS), 2(2), 1-25.

Mcleod, C. (2015). Trust. Accessed 1 November 2018. http://plato.stanford.edu/archives/fall2015/entries/trust/.

McLeod, C. (2002). Self-Trust and Reproductive Autonomy, Cambridge, MA: MIT.

Möllering, G. (2006). Trust: Reason, routine, reflexivity. Amsterdam: Elsevier.

Nickel, P. J. (2013). Trust in technological systems. In M. J. de Vries, S. O. Hansson & A. W. M. Meijers (Eds.), Norms in technology, philosophy of engineering

and technology (pp. 223-237). Dordrecht: Springer.

Nickel, P. J. (2020). Trust in engineering. In D.P. Michelfelder & N. Doorn, (Eds.), Routledge companion to philosophy of engineering.

O'Neill, O. (2004). Autonomy and Trust in Bioethics. Cambridge: Cambridge University Press.

Panovska-Griffiths, J., Kerr, C. C., Stuart, R. M., Mistry, D., Klein, D. J., Viner, R. M., & Bonell, C. (2020). Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second COVID-19 epidemic wave in the UK: A modelling study. The Lancet Child & Adolescent Health.

Ruijten, P. A., Terken, J., & Chandramouli, S. N. (2018). Enhancing trust in autonomous vehicles through intelligent user interfaces that mimic human behavior. Multimodal Technologies and Interaction, 2(4), 62.

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. (2017). Trusting digital technologies correctly. Minds and Machines, 27(4), 565-568.

Thiessen A. (2013) Reputation/Reputation Management. In: Idowu S.O., Capaldi N., Zu L., Gupta A.D. (eds) Encyclopedia of Corporate Social Responsibility. Springer, Berlin, Heidelberg.

Thompson, J. B. (2000). Political Scandal: Power and Visibility in the Media Age, Polity, 253-4.

van den Hoven, J. (1997). Computer ethics and moral methodology. Metaphilosophy, 28(3), 234-248.

van den Hoven, J., Pouwelse, J., Helbing, D., & Klauser, S. (2019). The blockchain age: Awareness, empowerment and coordination. In D. Helbing (Ed.), Towards digital enlightenment (pp. 163-166). Springer, Cham.

Walker, M. U. (2006). Moral repair. New York: Cambridge University Press.

Werbach, K. (2018). The blockchain and the new architecture of trust. Cambridge: MIT Press.

2

Can Social Credit System Promote Social Trust and Trustworthiness?

The Chinese Social Credit System (SCS) is a governance approach that contains a broad moral goal to foster virtues related to xinyong by institutionalizing the rough idea of "being a trustworthy citizen." However, in the absence of a coherent, upper-layer framework for understanding the trust concepts (i.e., trust and trustworthiness) and implementing the project, local initiatives adopt multifarious rules that are fraught with ethical and practical challenges. In this chapter, we critically engage with the SCS with an emphasis on whether its current implementations can achieve the moral end set by the overall project. We explore this question at two levels: assumptions and applications. First, based on an analysis of the implicit assumptions about trust underlying the SCS, we argue that the current SCS primarily facilitate the instrumental and prudential aspects of trust, showing a discrepancy between the moral objective intended to be achieved and the ways adopted to approach it. We then focus our reflection on three pilot cities' scoring systems built on the rational trust assumptions, further clarifying the detailed ethical issues associated with the design and audit of these applications. These issues arguably lie at the core of ethical discourse on the SCS and may lead the applications to deviate from the overall moral goal. Thus, trust issues in the

This chapter is based on the following article:

Teng, Y., & Alfano, M. (under review). Can Social Credit System Promote Social Trust and Trustworthiness? *Philosophy & Technology.*

Author contributions: YT structured this chapter and took the lead in the writing of the manuscript. MA heavily revised Section 2.1 and 2.3.1. Both authors have done multiple revisions of the draft chapter.

context of the SCS are discussed from the perspective of current initiatives' trustworthiness, with the project's moral goal as a touchstone, manifesting what is needed to move the initiatives to build moral trust relations.

2.1. Introduction

Social trust is the subspecies of the trust relationship that exists between distantly connected individuals (Banu 2019). It is most relevant in anonymous, one-off interactions with strangers who are known – if at all – only by reputation. In Giddens' (1990, 80) terms, social trust involves "faceless" commitments rather than "face-work" commitments. Despite the apparent fragility of social trust, it is arguably essential to expanding from closed tight-knit communities to an open society where more complicated forms of cooperation and coordination can be facilitated (Yamagishi 2011, 37-57). Unlike the mechanisms that account for trust in close, ongoing relationships with people one knows and shares important similarities, the mechanisms that account for social trust are more diffuse, generic, and distal, and they often rely on a complex and opaque system of contracts, laws, norms, institutions, and conventions.

In recent years, increasing studies have shown that high reputation-based trust can counteract people's natural tendency to place trust guided by the heuristic of homophily by providing information about a partner's intentions and capabilities (Abrahao et al. 2017). As a result, reputation systems are widely regarded as promising ways of encouraging social interaction within a community of (near-)strangers. Private and public initiatives around the world have been undertaken to make good on this promise. On the private side, behemoth reputation-based platforms such as Uber, Lyft, AirBnB, eBay, and others curate the reputations of both consumers and goods- and service-providers to facilitate social trust in economic interactions. Private credit-reporting agencies in the United States such as Equifax, Transunion, and Experian likewise track and provide reports on the creditworthiness of financial consumers that inform not only home and car loan decisions but also hiring and other high-stakes decisions.

In 2014, the government of China instituted a state-managed reputation system: the social credit system (SCS), known domestically as "*shehui xinyong tixi*". Before proceeding, we pause to note that the English term "credit" is an imperfect translation of the Chinese term "*xinyong*" (Dai 2018). The reasons are primarily two, both of which indicate the broad moral goal embedded into this national project. The first is about a semantic distinction. According to the Cambridge English Dictionary (n.d.), credit is "a method of payment for goods or services at a later time," which is mainly associated with financial and monetary behaviors or capabilities. In comparison, the meaning and scope of *xinyong* are much broader. Apart from the commercial meaning, in a broad sense, *xinyong* can be used to refer to the willingness and capacity of individuals and organizations to comply with social commitments (National Standardization Administration 2017). The term thus implies not only to the ability to make repayments but also to a rich array of qualities and moral virtues such as promise-keeping, sincerity, integrity, honesty, and self-discipline, depicting both dispositional and relational trustworthy behaviors in almost all social interactions. In this regard, it may be more accurate to interpret the Chinese SCS as a "social trustworthiness system."

The second reason that credit is not sufficient for understanding *xinyong* has to do with the actual implementation of this project. Indeed, the original shape of the SCS, which was already proposed in the early 1990s, is China's credit-reporting systems serving purposes similar to the private credit-reporting systems that already exist worldwide (Han 2005). As an effective way to ameliorate information asymmetry and risk in economic activities between strangers, such credit systems find their particular usefulness in the context of Chinese Confucianism where, as Feng et al. (2016) note, the understanding of trust contains a strong inclination to trust family members rather than strangers and outsiders. Over the years of the development of China's market economy, the government has noticed the advantages brought by the credit-reporting systems and intended to extend credit thinking to larger fields of social life. Thus, the SCS has gradually evolved to be an umbrella category that includes a cluster of initiatives going far beyond its original shape as an economic tool (Knight 2020). This also makes the term "credit" no longer enough for covering the assemblage of ideas and applications under the framework of the SCS. In particular, we have witnessed a goal shift of the SCS, from countering against financial risk to raising "the honest mentality and the level of *xinyong* of the entire society" (SCPRC 2014).[1] An important reason for this shift derives from the critical social need for remedying the ongoing trust crisis, which manifests in fraudulent behaviors, corruption, and professional malpractices rife in many social and commercial realms (Dai 2018). From this perspective, it seems that the expansion of the SCS not just incorporates but also emphasizes a significant moral element that is expected to be used to guide people's trustworthy behavior in general and increase social trust accordingly.

Based on the semantic clarification and the goal shift of the SCS, it seems reasonable to say that the project is seeking to foster not merely citizens' creditworthiness in economic interactions, but also *moral* trust relations grounded in the rough idea of "being a trustworthy citizen". This is comprehensible in the sense that virtue is always considered central for governance and self-governance in the Confucian tradition (Creemers 2018). Thus, if the SCS succeeds in its aims, it would promote a virtuous feedback loop between trust-keeping behavior and trust-placing commitment, facilitating both trustworthiness and an ecosystem of social trust.[2] In this chapter, we provide a philosophical reflection on the SCS project with a focus on whether its current initiatives can achieve the broad moral end set by the project. Based on an analysis of how local initiatives are theorized and implemented today, we argue that the systems primarily facilitate the instrumental aspects rather than the virtuous aspects of trust. In the light of the main issues discussed, we also provide several suggestions that can help foster virtues-based trust relations,including the cultivation of benign motivations and the guarantee of proper design and audit of the systems.

[1] SCPRC is short for the State Council of the People's Republic of China.

[2] A broader and more complex question, which we do not address in this chapter, is whether partially- or fully-automated credit systems are defensible full stop. This chapter should thus be seen as exploring a range of necessary (though not necessarily sufficient) conditions that the SCS would have to meet in order to be defensible. We then assess existing implementations against these criteria.

The rest of the chapter is organized as follows. First, we provide a thumbnail sketch of the contours of the SCS and point out the conceptual confusion involved in the implementation process. We then turn to the discussion about the conceptions of trust and trustworthiness assumed by the current initiatives, which are arguably not completely consistent with its overall objective of facilitating moral trust relations. Next, we examine three pilot cities' scoring systems as case studies to further explicate the detailed ethical issues shown by the design and audit of these applications. As such, social trust and trustworthiness, as two core issues of the SCS, are analyzed with the practical goals of cultivating virtues related to *xinyong* and informing the design, development, and audit of the initiatives.

2.2. Understanding the SCS: The conceptual confusion and the resulting discrepancy

At a general level, the SCS refers to "a new mode of data-driven governance" that builds on a series of social arrangements ranging from the construction of relevant institutions, incentive mechanisms, technical infrastructure, and the credit market to educational and cultural developments (Backer 2018). While analogous credit systems have existed for decades in many countries, the unprecedented comprehensiveness and government-driven invasiveness of the SCS have received extensive attention, especially given the fact that the number of potential subjects of this system is 1.4 billion people.

Currently, the SCS is supported by three pillars:

(1) *The credit-reporting systems*. Issued by credit-reporting agencies, the stated principle of the credit-reporting system is to provide objective information describing historical financial records of debtors. For the sake of reducing risk of the creditor, credit reports for individuals and enterprises as final outputs of credit-reporting agencies have been described as a means of "self-defense" by the Public Bank of China (PBoC 2014). As of January 2021, apart from the PBoC, there are 2 licensed institutions that can issue credit reports for individuals (i.e., Baihang Credit and Pudao Credit) and 131 institutions that are issuing credit reports for enterprises.[3]

(2) *The joint punishment and reward systems, including blacklists and redlists*. These lists are proactively published by social credit authorities and publicly

[3]The well-known Sesame Credit from Alibaba is now part of Baihang Credit, together with other seven institutions. But it should be noted that the Sesame Score (like the scoring systems created by other tech giants that are part of Baihang) is different from the official credit reports issued by Baihang Credit. While the former primarily functions as a comprehensive data-driven tool for "pay-later" services (e.g., borrowing a mobile power supply or a bike without deposit) within its own ecosystem (i.e., Alipay and Taobao), the latter can be used legally by financial institutions to inform debts and loans. In fact, after failing to get an independent license for issuing credit reports for individuals, Sesame has gradually terminated all financial services for individuals. A full list of these credit-reporting agencies can be found: http://www.creditsoso.org/content.asp?ID=4479 (in Chinese). More information about the relationship between Sesame and Baihang can be found: https://baijiahao.baidu.com/s?id=1648987655033975415wfr=spiderfor=pc.

accessible on the website of Credit China and local websites of the SCS. The term "trust-breakers" derives from the Supreme Court, originally referring to people and organizations that face a series of joint sanctions due to their law- and/or regulation-breaking behavior and malicious refusal to respond to valid court decisions (e.g., by paying fines). Empowered by the SCS project, other central government sectors and local governments are also eligible to identify trust-breakers, as well as trust-keepers who are seen honorable from the SCS's perspective.[4] Likewise, while people and organizations on trust-breaking blacklists face a series of sanctions, people and organizations on trust-keeping redlists can obtain certain rewards. During the time of crisis caused by COVID-19, several local governments have used the blacklist system as an extra punishment for those who violated public policies (e.g., concealment of epidemic information) made for fighting against the pandemic, as well as the redlists systems to honor those who contribute to the fight (Knight and Creemers 2021).[5]

(3) *Local implementations of the SCS*. As of 2018, 43 pilot cities, at different administrative levels, have implemented local SCSs (NDRC 2018).[6] As the oft-cited guidance documents – "2014 Planning Outline for the Construction of a Social Credit System" (Hereafter 2014 Planning) and "Guiding Opinions on Strengthening the Construction of Personal Credit System" (Hereafter 2016 Opinions) does not specify how the SCS should be implemented in practice, pilot cities implement the idea in various ways. The main forms of these implementations include: rating and scoring systems for individuals, focus groups, corporations, industries, and government sectors; portals for information disclosure, e.g., the disclosure of credit information about universities, corporations, and industries, the disclosure of individual certificates of lawyers, judicial appraisers, teachers, and accountants, and the disclosure of business commitments made by corporations; and *Xinyi+*, different ways to reward trust-keepers (e.g., simplified procedures of public services and discounts on rents and tickets for tourist attractions).[7] Although the above two documents never mention a social credit score for individuals, local scoring systems are currently experimenting in many pilot cities. As a result, the identifications of trust-keeping and trust-breaking acts are depicted in multifarious ways that lack a coherent logic with respect to how the score-related items are formulated and how the points are related to the severity of trust-breaking acts (SCPRC 2016; 2014).

[4]Public servants working in these sectors can also be blacklisted if they break related rules (e.g., fraudulent purchase and corruption)

[5]For more information about how blacklists are used for pandemic-related situations, see http://xy.fujian.gov.cn/133/10836.html (in Chinese).

[6]Local blacklists and redlists systems are also local implementations of the SCS, but we ascribe them to the second pillar for the sake of convenience. NDRC is short for National Development and Reform Commission.

[7]This does not mean that all pilot cities implement all of the listed measures. For example, many cities do not include a scoring system. Also, pilot cities may have implementations other than these categories.

On the one hand, the appropriate use of the three systems is able to contribute to solving otherwise-intractable social issues. The credit-reporting systems, as mentioned earlier, can ameliorate the acute problem of information asymmetry in the financial market, reducing the occurrence of adverse selection and moral hazard. The joint sanction and reward systems, which can be largely traced to the judicial system's enforcement predicament given the lack of individual bankruptcy law in China, can contribute to enhancing the efficacy of the court system and other derivative areas (Chen 2019). Regarding the local implementations, one important positive aspect suggested by van' t Klooster (2019) is that the systems might bring justice to market economies through rewarding moral virtues that are not compulsory yet still praiseworthy.

On the other hand, the different ways local government agencies use to build their systems cause overt confusion about how the trust concepts should be understood and applied in the context of the SCS. One primary reason for this confusion is the fact that neither the 2014 Planning nor any other upper-layer guidance document has stipulated a precise definition of "trust-breaking acts". What counts as a trust-breaking act? The answer provided by the 2014 Planning is a proliferation of examples, without a guiding principle to unify them. According to the Planning, the SCS is supposed to tackle social pathologies that occur in governmental affairs, as well as economic, social, and judicial fields, ranging from the violation of laws such as "grave production safety accidents, food and drug security incidents, commercial swindles, tax evasion…" to the breaking of moral standards and social norms such as "academic impropriety and professional courtesy".[8] Likewise, in Credit—General Vocabulary, "discredit" and "faith-breaking" are identified as broadly as the credit subject's failure to perform the promised act imposed by laws and regulations, contract terms, and other socially reasonable expectations (National Standardization Administration 2017).

In the absence of full-fledged laws and regulations that can guide and oversee the SCS in practice, contemporary and traditional understandings of the promised act and local implementations of the SCS tend to fill in the granular details (Romele 2019). As Dumbrava (2019) argues, by means of various incentivization mechanisms, local SCSs "transgresses the boundary between legality and morality", engaging with a wide variety of normative structures, including social norms, economic obligations, market dynamics, and governmental regulations. Overlaps and mutual impacts among these systems can impose repeated punishments for the same offense, running counter to the basic legal principle that the same fault should not be punished twice (De Filippi 2019). Nevertheless, this is not to say that the meaning of trust-breaking cannot be applied to violations arising in these areas. What is controversial is whether violations of such disparate rules and principles should all be perceived as trust-breaking acts. Arguably, trust-breaking should neither fully equate to law-breaking nor equate to morality-breaking, considering that the former

[8]While the scope of the SCS is broad and also covers trust in business and governmental agencies, the context of discussion of this chapter is primarily trust between individuals.

disregards the principle of rule-of-law societies and the latter disrespects individual rights and one's own personal value system within the ambit of legality. After all, there are so many ways to do or be evil, but not all of them are trust-breaking acts and thus not all of them should be incorporated in the initiatives.

Aside from the relevance of the detailed standards to trustworthiness, a more fundamental problem with the basic design of most local implementations is that these systems tend to rush into assigning point values to various acts and result in a single score that is meant to represent an individual's overall level of trustworthiness. While such an approach may make it easy to calculate and extend the influence of individual reputations, an amalgam of trustworthy characteristics without sub-indicators makes it difficult to assess the correlations between specific component items and results (Colquitt et al. 2007). The lack of such assessments for understanding the correlations between different items further impedes the examination and optimization of the models adopted by local initiatives. Furthermore, considering the wide range of behaviors and domains covered by the systems, applying a score resulted from one's behavior in one life domain (e.g. commercial) to other domains (e.g., public health) seems not just statistically but also morally indefensible. Doing so breaks the boundaries of social spheres that, according to Michael Walzer's (1983) influential theory of justice, are grounded in different internal goods and distinct sorts of distributive logic. Sphere transgression that converts one's advantages or disadvantages in sphere A (e.g., financial sphere) into sphere B (e.g. public health) thus squeezes out the significance of the particular distributive logic of sphere B and goes against the principle of justice.

To conclude, given the lack of a coherent, upper-layer framework for understanding the trust concepts in the context of the SCS, local initiatives interpret and apply the rough ideas of facilitating 'trustworthy citizens' in various ways. Gradually, a gap appears to emerge between the moral objective set by the overall project and the ways local initiatives implement the project. More specifically, there is a discrepancy between the conceptions of trust and trustworthiness that should be theorized and applied in order to achieve the moral objectives of the project and the implicit assumptions of the concepts we can reason from how the systems are implemented today. In other words, current initiatives built on flawed assumptions can hardly be considered as a justifiable means of achieving the project's moral ends – i.e., to create a virtuous feedback loop between trust-keeping behavior and trust-placing commitment. In the next section, we explicate this argument by providing a close examination of the conceptual muddles related to the trust and trustworthiness in the SCS.

2.3. Conceptions of trust and trustworthiness associated with the SCS

Different assumptions about the nature of trust can generate distinct normative and practical consequences (Jacobs 2020). In the context of the SCS, the trust concepts

might be conceptualized in a way that can help develop practical standards fostering *xinyong* properly, but they might also be poorly conceptualized such that they lead to distrust or an atmosphere that economizes on moral trust relations. A trust relation is commonly thought to have three places: a trustor, a trustee, and an entrusted thing of value (Horsburgh 1960; Baier 1986). Moral trust relations can thus be understood from the moral significance of the trustor's trust and the trustee's trustworthiness that are both built on the relation connected by the entrusted thing. In this section, by scrutinizing how reputation systems in general and the SCS in particular work to facilitate trust, we contend that the underlying assumptions of the trust concepts shown by the current initiatives of the SCS are trust as rational-choice and trustworthiness as a prudential strategy, which seem not entirely in compliance with the project's moral aim. Taking into account the rational assumptions of trust adopted, we then discuss the significance of building trustworthy implementations to shape the moral effects of the project.

2.3.1. Trust as rational-choice and trustworthiness as a prudential strategy

Perhaps in stark contrast to the frequent use of the term "trust", what mental construct trust is remains a highly contentious issue in contemporary discourse. For example, whether it is a belief, an attitude, an emotion, or a judgment. But one thing that is clear is that trust gets its richness and specific meaning in specific contexts. Applying the three-place trust structure to reputation systems, a potential trustor is the subject who intends to reach a decision about whether to trust a (near-)stranger based on the information offered by reputation systems; a potential trustee is the subject about whom there is information in the reputation systems; and the entrusted thing of value depends on the trustor's specific needs and interests.

Following this structure, a basic precondition on the success of a reputation system such as the SCS is that users of the system make decisions rationally in terms of the specific information that has been generated, harvested, and released by various sub-systems of the SCS (see the three pillars canvassed above for details). In this sense, it might be said that the SCS is designed to be best suitable for rational actors to make rational decisions. This approach to trust is essentially calculative, accentuating trust's role as an instrument for maximizing personal utility without attributing or relying upon extraordinary virtue in the one-trusted (Coleman 1994). According to this approach, trust can be seen as a particular expectation the trustor has with respect to the likely future behaviour of the trustee (Gembetta 1988). In trust theories, a trust decision reached through this mechanism is often referred to as rational-choice or risk-assessment accounts, hinging on the weighing-up of the prospects for gain and loss.

However, trust as rational-choice falls short of reaching the SCS's goal of forming moral trust relations as a narrowly rational account of trust does not link the trustee's action to the motivation behind the action; and thus the decisions made accordingly

do not address a focus on what motivates another to display certain actions. In philosophical research, a trust relation is thought to be moral and distinctive typically because it is grounded in some morally defensible motives. For example, when the trustee cares about or at least does not bear ill will towards the trustor (Baier 1986), when the trustee is committed to common decency and has moral integrity (McLeod 2002), and when the trustor believes that the trustee is obligated to do certain things (Nickel 2007). Such motives allow the trustor to rely on the trustee to fulfill the entrusted task while simultaneously accepting the vulnerability and risk that come with placing trust. The idea of using prescribed, institutional structures such as contracts, inspection, monitoring, and legislation to secure the trustworthiness of the trustee, instead, simply replaces the moral grounds and uncertainty involved in trust relations with more formal procedures and makes genuine trust less necessary (O'Neill 2004).

Unlike trust as a mental state, trustworthiness is often understood as a characteristic (O'Neill 2004). In the context of reputation systems, it is arguable that the implicit conception of trustworthiness assumed and enabled by the initiatives is trustworthiness as a prudential strategy rather than a moral disposition. To clarify, to say that values are prudentially acceptable means that they are good for someone's well-being but not necessarily morally good (Tiberius 2006; Ferrario et al. 2019). In this case, one may be motivated by mere self-interest in order to maintain *potential* relationships shaped by the systems. This account might be seen as an extension of Hardin's (2001) "encapsulated interest" view that explains trustworthy action as a result of the trustee's interest to maintain the relationship with the trustor. The difference between trustworthiness as a prudential strategy and trustworthiness as a moral disposition can thus be reflected by the answer to a simple question: can the incentivized and rewardable "trust-keeping behavior" reveal and advance one's moral virtues of honesty, promise-keeping, and sincerity? In *Marx's notebook comments on James Mill*, Marx (1844) claims that the credit system is in fact a higher-level and mature mode of existence of the money system, the object of which "is no longer commodity, metal, paper, but man's moral existence, man's social existence, the inmost depths of his heart, and because under the appearance of man's trust in man it is the height of *distrust* and complete estrangement." The alienation view of credit regards all the social virtues of human beings as the insurance of interest and repayment. If this is right, then in the context of credit systems, a "trusted" or a "good" person merely means "a person who is able to pay or otherwise live up to what is expected of them" (Marx 1844). This makes it unclear whether the underlying logic of different sorts of credit systems that exist around the world genuinely address the virtues related to repayment at all.

From the perspective of alienation, it seems that the whole meaning of one's social virtues in credit systems is to be translated into economically calculable credit. And, once this translation is finished, social virtues will become meaningless to the system. An example can be found in the principle of credit reporting that often does not distinguish "malicious in arrears" from "unintentionally in arrears".[9] Under the

[9]For example, see PBOC's identification of arrears: http://www.pbccrc.org.cn/crc/kffw/201310/9bc082c1

mask of the credit industry, the meaning of trust might be reduced to a behavioral level and become phony and passive as human initiative and selectivity are gradually replaced by standards, contracts, and institutions. Risk aversion and social rewards squeeze virtue incentives and contextual factors out of the motivation list, repeatedly polishing the socially desirable understanding of trust and trustworthiness. As a result, a person's trustworthiness appears to be entirely incarnated by their overt, measurable, algorithmically- or bureaucratically-evaluable behavior.

Through the lens of scoring systems, a trustworthy person is meant to be the one who has a higher score in the system. However, it seems that virtuous people are not necessarily those who have higher scores and people who have higher scores are not necessarily virtuous. In other words, there are reasons to expect both false positives (people rated as trustworthy who do not embody the relevant virtue) and false negatives (people rated as untrustworthy who are in fact trustworthy) in the system as it is currently implemented. Regarding the false negatives, genuine virtues are typically understood as deep features of someone's character. If this is right, then they should not be overly responsive to instrumental concerns such as reputation and money nor be prescribed as blind conformity to fixed social conventions (Reijers 2019). On the contrary, people should act in accordance with the reasons to which their virtues are responsive, and not necessarily the incentivized mechanisms and prescribed benchmarks that would help them score points. In this regard, it makes sense that such people may not have a high score, but this should not follow the indication that they are untrustworthy.

Regarding false positives, malicious users can disguise themselves by deceiving and exploiting the system via system gaming. For some local initiatives of the SCS, for example, users can accumulate points by deliberately doing certain score-increasing things requiring little effort (e.g., small donations). In that case, not only are they able to reinstate points deducted for minor infractions, but they could also leverage this mechanism as a means or shield for doing unscrupulous things. Such problematic behavior directs us back to the importance of correct motivations for performing socially desirable actions; otherwise it might go right against the ultimate goal set by the overall project. To achieve a virtuous feedback loop between trust-keeping behavior and trust-placing commitment, then, more efforts should be made regarding both education and cultivation of citizens' good motivations for displaying real virtuous behavior.

It should be noted that the instrumental view of trust and trustworthiness analyzed here is not exceptional for the SCS. As mentioned, it can be applied more broadly to reputation systems and credit-reporting systems widely adopted by public and private initiatives worldwide, and especially to those systems involving scoring and rating mechanisms. Nonetheless, issues entailed by this narrow, rational fashion become particularly contentious against the backdrop of the moral goal of the SCS since it creates a discrepancy between how the trust concepts should be charac-

96ca4 308bd5886b9fab5b2e5.shtml.

terized and how they are currently conceptualized.

2.3.2. Implementing the rational assumptions: The importance of justifiable rules

The above discussion makes clear the discrepancy, but it does not necessarily mean that applications built on the rational assumptions cannot protect, preserve, and promote morally salient features of trust. The moral appropriateness and moral effects of a reputation system, in this case, relies heavily on the design and audit of the system. If a system is well-designed and well-audited in terms of appropriate and justifiable rules, then citizens' compliance actions might promote appropriate patterns of behavior and ameliorate severe social problems. This can be analogous to corporations' compliance with stringent data-protection regulations where trust and trustworthiness might be brought about through legal certainty and transparency. Accordingly, it seems reasonable to consider the rule-following acts informed by the system as a sort of second-best, which is preceded by virtuous behaviors motivated by self-awareness and self-government.

For example, the development of credit-reporting systems, though based on objective, financial-related information and prescribed rules, can help cultivate people's repayment and promise-keeping habits, facilitating a safer and more responsible atmosphere within which people can reasonably expect others' behaviour, even if only out of prudence rather than moral virtue. Interactions are thus facilitated by the system itself that is designed and regulated with the good of society in mind, especially when it comes to helping individuals build tentative trust relations via a transparent and effective application. Also, recording important information that is relevant to social trust – such as refusals to pay public utility fees and carry out court decisions after being notified sufficiently – helps address the enforcement predicament mentioned earlier and protect the creditors' vulnerabilities. Such information is currently included not only in many pilot cities' systems but also in other countries' credit systems.[10]

By contrast, if the benchmarks of a system are not defensible from the perspective of the ethics of trust, then it remains unclear whether citizen's rule-following behaviors induced by, or subjected to, the system would contribute to fostering the morally laden aspects of trust. It should be noted that using algorithms to measure attributes that are not directly observable is always a thorny problem given that the process of mapping from the attribute of interest to observable proxies is limited and fallible (Friedler et al. 2016). One significant way of coping with this difficulty is to reconsider the relevance of the attributes being changed and close the distance between attributes making a difference and attributes that should make a difference (Venkatasubramanian and Alfano 2020). This requires rule-makers to remove

[10] For example, these rules are included in Shanghai's SCS (http://www.spcsc.sh.cn/n1939/n2440/n3898/u1ai149901.html) and Guangdong's SCS (http://www.rd.gd.cn/zxfb/202103/t20210325_183314.html). Also, they are included in the United States' credit system: https://www.consumer.ftc.gov/articles/getting-utility-services-why-your-credit-matters.

irrelevant and inappropriate attributes (to the extent possible) in order to make the outcome of algorithmic decision-making more normatively acceptable.

Such a rule-making process highlights the importance of establishing reliable systems before talking about and promoting people's trust and trustworthiness. Relatedly and more fundamentally, this emphasizes the dominant role played by rule-makers of reputation systems who directly determine the specific proxies chosen to reflect individuals' trustworthiness and guide the meaning of trusted interactions. The privileged position of the SCS agencies as rule-makers creates an indirect but "ultimate trustor" to whom individuals, groups, and organizations should be or appear trustworthy. Everyone within the reach of the systems is thus obliged to comply with the stipulated benchmarks considered societally good (or bad) on pain of reputational loss (and the knock-on effects of such reputational loss, such as loss of opportunities for personal, business, and governmental cooperation). This inherently rational mechanism, as discussed, links only indirectly to the moral motives of participants, and should be built on a well-designed and -audited system if morally desirable consequences on social trust remain one of the main goals of the project.

In sum, we have argued that the implicit assumptions of trust and trustworthiness made by current initiatives of the SCS are primarily rational, which fall short of providing a justifiable way of approaching moral trust relations. We have also pointed out the significance of incorporating justifiable rules into the systems built on the rational trust assumptions to help shape the moral effects of the systems and bring about proper trust and trustworthy behaviour. To take a closer look at how these assumptions are embedded in and embodied by current initiatives of the SCS, in what follows, we take three pilot cities' scoring systems as case studies and scrutinize whether they can be reliable premises of the promotion of social trust and trustworthiness.

2.4. A reflection on current local implementations of the SCS: Design and audit

Since the publishing of the 2014 Planning, the domains covered by local scoring systems seem to be almost unlimited, and there is little coherence to the set of behaviors and omissions that are included (Chen 2019). Given the above analysis of the importance of establishing trustworthy systems, detailed reflection on the relevance and appropriateness of the rules and specific items built into current applications is urgently needed, in order to gradually approach an optimal, fine-tuned amount of governance that does not overstep the purpose of the SCS. In this section, we focus our ethical reflection on three pilot cities' scoring and rating systems that, as discussed, are deeply rational and mainly invite prudential behaviour. The primary aim here is to demonstrate whether current implementations grounded in the instrumental view of trust are fit for the moral purpose and properly audited. We begin by introducing these systems, and then investigate two essential ele-

ments of the design of these systems – i.e., the content and the point values of the rules. To clarify, we first take a close look at local SCS agencies' discretion over the identification of trust-breaking acts. Then, drawing on an inquiry into the equivalences built into these systems, we come to discuss the ethical issues inherent in the underlying logic of the scoring systems, which may lead the systems to deviate from the intended moral goal. We conclude with a brief discussion about the audit of the SCS, which has made effective efforts to govern the initiatives as well as left some remaining concerns.

2.4.1. Discretion over the identification of trust-breaking acts

The discretion at issue here is local governments' right or ability to determine the enactment, judgment, and implementation of locally defined standards for trust-breaking acts. As discussed earlier, since there is no nationally unified standard with respect to the construction of scoring systems and rating approaches, pilot cities can construct local standards and data platforms with a certain amount of discretion. The initial points of local SCSs normally range from 100 to 1000, with simple addition and subtraction of score-related standards and direct rating judgments (i.e., serious violations) as the major methods of calculating points and ratings.

Here we adopt the initiatives of Suqian, Rongcheng, and Weihai as case studies.[11] The initial points of the three cities are the same—1000, and they all include a rating mechanism that applies a qualitative rating to each subject. Despite being developed with allegedly the same aim of assessing and promoting individual trustworthiness, the rating levels, score sections, and specific calculating standards adopted by the three cities vary significantly (see Table 2.1).

Table 2.1: Basic information about the SCSs in Suqian, Rongcheng, and Weihai

	Rating levels	Score (X) section for the lowest and highest level	Score-deducting items	Score-increasing items	Discretion level
A. Suqian	8	X≤599; X≥1250	59	21	Low
B. Rongcheng	7	X≤599; X≥1050	570	150	High
C. Weihai	6	X≤800; X≥1150	2900	240	Middle

With only 59 score-deducting and 21 score-increasing items, Suqian's system has the fewest benchmarks. Explicit items in Suqian's system address loans, public utilities, tax arrears, cheating on examinations, judicial information, and disciplinary punishments within the administration. Focusing on particular fields, these trust-breaking items are mostly violations of conventional social commitments and obligations that are fairly uncontroversial in the social and cultural context. Regardless

[11] Choosing these three cities instead of others is because their implementations of the SCS contain relatively integrated content and those official documents are accessible online. Also, these three cities are all rewarded as exemplary cities for social credit system construction in 2018. For Suqian's SCS (in Chinese), see https://cxsq.suqian.gov.cn/xysq/xc/content/f4959e26-f442-4341-b237-647e493e1025.html. For Rongcheng's SCS (in Chinese), see https://www.chinalawtranslate.com/rongcheng-municipal-personal-credit-appraisal-standards/. For Weihai's SCS (in Chinese), see http://credit.weihai.gov.cn/ueditor_upload/file/20181114/1542160367090027562.pdf.

of how these rules are calculated, this system might constitute an acceptable governance supplement to existing legal systems without many transgressions of other normative structures due to the clear boundary and limited number of items included. In comparison, the number of standards of Weihai's system is quite high. With 2900 score-deducting and 240 score-increasing items, Weihai's SCS initiative covers almost all previously isolated information generated within the ambit of local government sectors. But what significantly distinguishes Weihai's identification of trust-breaking acts from Suqian's are the numerous overlaps between Weihai's standards and existing laws and regulations that contribute to the vast majority of the standards. From this perspective, it can be said that Weihai's SCS leans heavily on double jeopardy discussed earlier, which is worrying from both moral and legal perspectives.

The number of rules defined by Rongcheng's system lies in between the above two cities, which is 570 score-deducting items and 150 score-increasing items. But it seems that the explicit items of this system manifest an overwhelming governance attempt for behavioral control and collective management. Particularly in the field of social management, clues indicating the high discretionary power possessed by the local government over the formulation and execution of the standards are easily found. For instance, citizens can be penalized for "unreasonable refusal to demolition (-100 points)", "failure to perform one's filial duty (-50 points)", "extravagant weddings and funerals (-10 points)", and "acute conflicts between neighbors (-5 points)". These benchmarks indicate the local government's high discretionary power in two respects. First, unlike the other two cities, the normative structures adopted by Rongcheng's SCS largely transgress different sorts of social expectations, some of which are quite contentious with respect to their relevance to the quality of trustworthiness. As Backer (2018) points out, a fundamental concern with the SCS is the difficulty of separating the SCS's role as a rational solution to social pathologies from its role in promoting a wide variety of governance preferences. In the absence of efficacious constraints, Western observers often link such items to an inclination for social control echoing the plot of the "Nosedive" episode of Black Mirror. Second, the degree-dependence terminology (i.e., unreasonable, failure, extravagant, bad, and acute) used to describe the rules is ambiguous. In this regard, the power over judging different situations is left to the discretion of government officials on a case-by-case basis, which might cause unfairness between different cases.

Based on the above investigation into the design of the three systems, three rough routes of the construction of local scoring and rating systems are shown. (1) The route of Suqian: incorporating rules related merely to violations of conventional social commitments and obligations that are pertinent to social trust and trustworthiness but do not overlap much with the legal system. (2) The route of Weihai: incorporating violations related to not merely social trust and trustworthiness but also existing laws and regulations. (3) The route of Rongcheng: incorporating violations not merely associated with social trust and trustworthiness but also a wide variety of governance preferences. To the extent that the three local governments'

power practices penetrate into determining trust-breaking acts, the discretion levels of the three local governments can thus be ranked as, from low to high, Suqian, Weihai, Rongcheng. Such a comparison study gives us a sense of how the SCS is applied differently in pilot cities and the nuances among the design of these systems. It also lays a foundation for our analysis of more detailed ethical problems associated with the equivalences built into these scoring systems.

2.4.2. Discretion over equivalences built into the scoring systems

Apart from the content of the rules, another important element of the scoring systems concerns the point values of the rules. Points assigned to specific items can be viewed as an index implying the severity of different violations; nonetheless, the making of these rules often lacks scientific explanations for how equivalence is created among different items. Questions about why certain points are assigned to certain items and why the points of item A are equivalent to the points of item B have rarely been addressed. These ambiguities are especially troublesome in three situations.

First, when the compared items are from different areas, does that mean heterogeneous values are comparable and commensurable by one universally calculable standard? As mentioned, whereas trustworthiness is often understood as a multi-dimensional construct in contemporary discourse, the scoring systems seem to presuppose that trustworthiness is a uni-dimensional concept where different characteristics of trustworthiness are amalgamated, and equivalences can be built among component items. To make these items commensurable for comparison, a common measure is needed (Nien-hê 2016). From the scoring systems' perspective, it follows that all the score-related items are presupposed homogeneously in their nature, which allows them to be then quantified by a common measure so that the items are comparable with respect to different levels of intensity. By the same token, values assigned to the scoring systems seem to be assumed at the ratio level of measurement through which attributes could not merely be rank-ordered in terms of higher or lower values as well as ordinal scales, but also the distance between two attributes is meaningful and the zero point is a true zero. For example, think of Y and Z as two residents living in the same city, whose personal scores are 1000 and 500 respectively. Following indications of ratio scales, the system might interpret that Y's chance of keeping trust is twice as much as that of Z. Such a holistic approach for calculating different values on a large-scale population appears to sit well with utilitarianism proposed by Jeremy Bentham (1996) and John Stuart Mill (1895), which seeks to maximize utility for the greatest number of a population. However, if the above discussion is on the right track, this approach cannot get rid of the typical criticisms against classical utilitarianism – not only because of the inability to measure and quantify varying degrees of trustworthiness in such a precise and scalable method, but also because some values constituting trustworthiness are inherently incalculable and incommensurable.

Second, when the compared items owning the same content are allocated diverse points across different cities' systems, does that mean values should be judged with geographical differences? There are two possible ways of understanding the geographical differences questioned. One considers the relational element of well-being, indicating that a value is supposed to be good to the extent that it fits a particular subject's needs and interests (Rodogno 2015). The other considers regional differences in judging the nature of the relevant values. In the case of the SCS where different cities maximize their utility in terms of local interests and priories of the relevant values, it seems that the divergence of points-assignment can be attributed more to the former rather than the latter category. For instance, it is possible that a city striving for economic development would reward more points for business investments while a city encountering serious environmental problems would impose more sanctions on industrial pollution. Nevertheless, such a divergence manifests the fact that this national experiment is currently more a patchwork of local initiatives, which lacks a feasible upper-level plan for the interoperation and integration of local systems. This partly explains why Dai (2018) predicts that, compared with more feasible projects (e.g., the nationwide credit-reporting system), a nationwide trust-scoring system is largely "in the nature of vacuous and propaganda projects".

While till now we have introduced pilot cities' different interests and the confusion made by the lack of a coherent, upper-layer framework for understanding the basic trust concepts in the SCS context, it is important to acknowledge the fragmented and decentralized governance model used by China's bureaucracy for decades (Zeng 2020). As Lieberthal (1992) argues, in this fragmented model, authority below the very peak of the system is allocated to different levels of government with certain discretion. This allows lower administrative levels of power to implement policies of the center in a way that takes into account local specificities while furthering interactive processes between the top and the bottom. After conducting repeated local trials and error revision, the central government may reach more suitable approaches that can then be extended at the regional or national level. In the context of the SCS, likewise, it can be said that the central government is rolling out the broad idea of "being a trustworthy citizen" via this fragmented model, which gives organizational agencies a certain amount of discretionary latitude and monitors their implementations at the same time (Knight 2020). Understanding this strengthens the argument that, as "an ecosystem broadly sharing a similar underlying logic" (Creemers 2018), the SCS is likely to remain fragmented in the future.

Third, when one of the compared items results in score deduction and the other leads to score augmentation, does that mean people can remedy small crimes via donations and investments? Score-increasing items provide ways of encouraging and rewarding citizens who exhibit prosocial behavior benefiting others or society. These items can be utilized as normal ways of accumulating points or remedying points deducted by certain violations.[12]However, a predicament of scoring systems,

[12]It should be noted that not all violations can be remedied by performing score-increased actions. Low

as mentioned earlier, is that the systems might invite gaming. Cunning users might leverage score-increasing rules as shields for doing unscrupulous things, particularly considering that some items are partial to the rich. In this way, it is likely that the new system would replicate the distributive concerns in real-world society and engender more unfairness. As a response to the fairness conundrum, Finland introduces a "progressive punishment" approach through which speeding fines are calculated in proportion to the offender's daily salary, but not based on a fixed number applying to all.[13]It might be acceptable to apply this approach to money-related items of the SCS (e.g., tax fines) within which what should be equalized is the motivation of citizens rather than the absolute amount of money they give. The ways of measuring rewards/sanctions might thus be linked in proportion to the subject's disposable income. But the scope to which this approach applies should be very limited and deserves systematic discussion before coming into play.

To conclude, it is questionable to what extent the extant local scoring systems can be trusted to reflect one's overall trustworthiness and to what extent a consequent score can be trusted when applied to other areas as a reference predicting one's future behavior. The three questions discussed above all imply the design flaws of the initiatives built on the rational trust assumptions, which are not compatible with the moral objective set by the overall project. Troubles spelled by the above inquiries are not exceptional for an instrumental view of social trust and trustworthiness, but more deliberations are needed when this view is adopted by a governance tool for promoting trustworthiness that is expected to be praiseworthy from both social and moral points of view. All these concerns refer to a more trustworthy SCS before we can talk about building our trust on the basis of it.

2.4.3. The protection of credit information: Efforts and concerns

Let us turn to the audit of the SCS especially the protection of credit information. Typically, data and information processed by the SCS are known as "public credit information (PCI)" or "social credit information", widely referring to the credit information generated, collected, and aggregated by government agencies. In this part, we briefly discuss the efforts made by the initiatives used to audit the processing of PCI – i.e., to examine and make sure that PCI has been processed correctly, as well as the remaining concerns over PCI and the implementation of the SCS.

As Chen and Cheung (2017) point out, legislation on PCI in China remains largely insufficient and fragmented. Although citizens are granted basic access and correction rights to their social credit scores, only limited restrictions are imposed on the collection and secondary usage of PCI. For example, for Rongcheng's system, eligible citizens can get access to their scores instantly after identity authentication

points resulting from severe violations that lead to direct rating deductions (e.g., from A to B or from A to D) commonly cannot be remedied before certain conditions are satisfied.

[13]Information about Finland's speeding fines is available at www.weforum.org/agenda/2018/06/in-finland-speeding-tickets-are-linked-to-your-income/.

via the system's WeChat mini-program or certain places offline, and they can also check the reports elaborating their violations and rewarded actions. This is somewhat tricky since it seems that all information related to the system's benchmarks is collected and possessed before the user's consent.

The good news is that things might be changed after the implementation of a series of legal provisions related to the SCS. The first is the strict "Personal Information Protection Law" that was passed on August 2021 and will take effect from November 2021 (Creemers and Webster 2021). Centered around principles including legality, propriety, and necessity, the law emphasizes data subjects' rights of informed consent and the withdraw of consent, as well as tech companies' and government sectors' obligations to ensure the quality of users' consent (e.g., free, specific, and prior). Also, the law specifies individuals' right to refuse data processors' decisions made merely by automated decision making and the right to request an explanation from the processors. This is especially important for regulating the initiatives grounded in artificial intelligence (AI) – such as the Sesame Credit – that can learn and make decisions autonomously (Dai 2018). But this is not sufficient, considering the potential risks AI might entail. Due to AI's specific characteristics – such as opacity, complexity, and unpredictability, for example, AI applications may display various sorts of discrimination based on gender, age, and ethnic origins, leading to unfair distribution of public services and financial resources (European Commission 2020). Thus, more specific and nuanced regulatory frameworks regarding AI applications in the Chinese context are urgently required to avoid undesirable consequences engendered by AI.

The second legal effort is about regulating the categories of PCI. Perhaps as a direct result of the center's monitoring of local initiatives under the fragmented governance model, in July 2021, the NDRC published a brief, updated version of "the Basic Catalogue of PCI (Consultation Paper)." In this document, the NDRC numerates some contentious information – such as those related to "complaint-reporting, garbage-sorting, uncivilized dog-keeping, blood donation…" – and explicitly stipulates that such information should not be included in the SCS. This brief document might be seen as a formal response to the ill-suited design of local initiatives. It helps not only to limit local organizational sectors' discretionary power over the identification of trust-breaking acts but also to crack down on system gamers who intend to leverage the points related to blood donation. In particular, this document highlights the importance of malicious intention when it comes to deciding whether to include small crimes such as jaywalking and arrears of property fees to PCI. This also meets the discussion of the importance of motivations provided earlier.

The introduction of these legal efforts shows the dominant role of the central government as the ultimate trustor who always monitors and supervises the innovations revolving around the SCS. It's certain that these legal efforts will soon be translated into the design and implementation of public and private initiatives and contribute to shaping the moral effects of the overall project grounded in the rational trust assumptions. Nonetheless, the above documents do not mention how local initiatives

should be built on the basis of the currently allowable PCI, nor is there any content about local scoring systems, both of which make the ethical inquires presented in the above subsection ongoing concerns.

2.5. Conclusion

In this chapter, we critically engage with the question of whether the Chinese SCS can foster moral trust relations via its current implementations, as well as some logic behind reputation systems in general. To this end, we provide a close philosophical reflection on the normative assumptions of trust and trustworthiness made implicitly by the initiatives of the SCS. We contend that these underlying assumptions primarily foster trust relations in an instrumental and prudential sense, showing a discrepancy between the moral objective of the overall project and the current ways of approaching it. To scrutinize the moral effects shaped by the design of current initiatives built on the rational assumptions, we provide an ethical inquiry into three pilot cities' scoring systems within which more detailed statistical and moral issues in relation to the measure and use of citizens' trustworthiness begin to emerge. Furthermore, a brief discussion about the efforts and concerns associated with the audit of the SCS is provided. In sum, the analysis of both the underlying assumptions and the current applications of this project all make clear the need for developing more trustworthy systems before talking about promoting trust and trustworthiness in terms of the systems' rules.

References

Abrahao, B., Parigi, P., Gupta, A., and Cook, K. S. (2017). Reputation offsets trust judgments based on social biases among Airbnb users. Proceedings of the National Academy of Sciences, 114(37), 9848-9853.

Backer, L. C. (2018). Next generation law: Data-driven governance and accoun tability-based regulatory systems in the West, and social credit regimes in China. Law &: Southern California Interdisciplinary Law Journal 28(1):123-172.

Baier, A. (1986). Trust and antitrust. Ethics, 96(2): 231-260.

Banu, M. (2019). Why do we trust strangers? Social trust, moral reasoning and identity. Annals of the University of Bucharest-Philosophy Series, 67(2), 39-66.

Bentham, J. (1996). The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation. Oxford: Clarendon Press.

Cambridge Dictionary. (n.d.) "Credit". Accessed June 6, 2019. https://dictionary.cambridge.org/dictionary/english/credit?q=credit.

Chen, J. H. (2019). Putting 'good citizens' in 'The Good Place'? VerfBlog, https://verfassungsblog.de/putting-good-citizens-in-the-good-place/. Accessed July 2, 2019.

Chen, Y., and Cheung, A. S. (2017). The transparent self under big data profiling: Privacy and Chinese legislation on the social credit system. Journal of Com-

parative Law, 12(2), 356-378.

Coleman, J. S. (1994). Foundations of social theory. Cambridge: Harvard University Press.

Colquitt, J. A., Scott, B. A., and LePine, J. A. (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.

Creemers, R. (2018). China's social credit system: An evolving practice of control. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3175792. Acce ssed May 28, 2019.

Creemers, R., and Webster, G. (2021). Translation: Personal Information Protection Law of the People's Republic of China (Effective Nov. 1, 2021). https://digich ina.stanford.edu/news/translation-personal-information-protection-law-peoples-re public-china-effective-nov-1-2021. Accessed Aug 25, 2021.

Dai, X. (2018). Toward a reputation state: The social credit system project of China. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3193577. Acc essed June 2, 2019.

De Filippi, P. (2019). The social credit system as a new regulatory approach: From 'code-based' to 'market-based' regulation. VerfBlog. https://verfassungsblog. de/the-social-credit-system-as-a-new-regulatory-approach -from-code-based-to-m arket-based-regulation/. Accessed July 2, 2019.

Dumbrava, C. (2019). The citizen, the tyrant, and the tyranny of patterns, Verf-Blog. https://verfassungsblog.de/the-citizen-the-tyrant-and-the-tyranny-of-patterns/. Accessed July 2, 2019.

European Commission. (2020). White Paper on Artificial Intelligence – A European Approach to Excellence and Trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en. Accessed August 8, 2021.

Feng, Z., Vlachantoni, A., Liu, X., and Jones, K. (2016). Social trust, interpersonal trust and self-rated health in China: A multi-level study. International Journal for Equity in Health, 15(1), 1-11.

Ferrario, A., Loi, M., and Viganò, E. (2019). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions, Philosophy & Technology, 1-17.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016). On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236.

Gembetta, D. (2000). Can we trust trust? in D. Gembetta (ed.) Trust: Making and Breaking Cooperative Relations (pp 213-237).

Giddens, A. (1990). The consequences of modernity. California: Stanford University Press.

Han, B. (2005). The economic analysis of social credit evolution, PhD diss., Jilin University.

Hardin, R. (2001). Conceptions and explanations of trust. In Karen S. Cook (Ed), Trust in Society. New York: Russell Sage Foundation, 3-39.

Horsburgh, H. J. N. (1960). The ethics of trust. The Philosophical Quarterly, 10(41), 343-354.

Jacobs, M. (2020). How implicit assumptions on the nature of trust shape the understanding of the blockchain technology. Philosophy & Technology, 1-15.

Knight, A. (2020). Technologies of risk and discipline in China's Social Credit System. In Creemers, R., and Trevaskes, S. (eds), Law and the Party in China: Ideology and Organi-sation. Cambridge: Cambridge University Press.

Knight, A., and Creemers, R. (2021). Going viral: The Social Credit System and COVID-19. SSRN. http://sci-hub.tw/10.2139/ssrn.3770208. Accessed Aug 20, 2021.

Lieberthal, K. G. (1992). Introduction: The "fragmented authoritarianism" model and its limita-tions. In Kenneth G. Lieberthal and David M. Lampton (eds.), Bureaucracy, Politics, and Decision Making in Post-Mao China, 1-30.

Marx, K. (1844). Comments on James Mill. In Marx and Engels collected works, Vol 3. London: Lawrence & Wishart.

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

Mill, J. S. (1895). Utilitarianism. London: Longmans.

National Development and Reform Commission. (2018). The Announcement of the List of the First Batch of Demonstration Cities for Social Credit System Construction, https://www.ndrc.gov.cn/xwzx/xwfb/201801/t20180109_873409.html.

National Development and Reform Commission. (2021). The Basic Catalogue of PCI (Consultation Paper). https://www.ndrc.gov.cn/yjzxDownload/20210713fj1.pdf. Accessed August 15, 2021.

National Standardization Administration. (2017). Credit – General Vocabulary. http://www.zggov.cn/article.php?id=48. Accessed September 4, 2020.

Nien-hê, H. (2016). Incommensurable values, The Stanford Encyclopedia of Philosophy, pla-to.stanford.edu/entries/value-incommensurable/.

Nickel, P.J. (2007). Trust and Obligation-Ascription. Ethic Theory Moral Prac 10, 309–319.

O'Neill, O. (2004). Autonomy and Trust in Bioethics. Cambridge: Cambridge University Press.

Public Bank of China. (2014). The main concepts of credit reporting. https://www.pbccrc.org.cn/zxzx/zxzs/201401/87814073fac f4b9795480d40fd626467.shtml. Accessed July 14, 2019.

Reijers, W. (2019). How to make the perfect citizen?, VerfBlog, https://verfassungsblog.de/how-to-make-the-perfect-citizen/.

Rodogno, R. (2015). Prudential value or well-being. In Brosch, T., and Sander, D. (Eds.). Handbook of value: Perspectives from economics, neuroscience, philosophy, psychology and sociology. Oxford: Oxford University Press.

Romele, A. (2019). An illusion of Western democracies, VerfBlog. https://verfassungsblog.de/an-illusion-of-western-democracies/. Accessed July 2, 2019.

The State Council of the People's Republic of China. Guiding Opinions of the General Office of the State Council on Strengthening the Building of the Personal Honesty System, December 23, 2016, chinacopyrightandmedia.wordpress.com/2016/12/23/guiding-opinions-concerning-strengthening-the-construction-of-a-personal-sincerity-system/; and the State Council of the People's Republic of China. Planning

Outline for the Construction of a Social Credit System (2014-2020), June 14, 2014, chinacopyrightandmedia.wordpress.com/2014/06/14/planning-outline-for-the-construction-of-a-social-credit-system-2014-2020/.

Tiberius, V. (2006). Well-being: Psychological research for philosophers. Philosophy Compass, 1(5), 493-505.

van't Klooster, J. (2019). Rewarding virtuous citizens, VerfBlog. https://verfassungsblog.de/rewarding-virtuous-citizens/. Accessed July 2, 2019.

Venkatasubramanian, S., and Alfano, M. (2020) The philosophical basis of algorithmic recourse. ACM Confer-ence on Fairness, Accountability, and Transparency. Barcelona, Spain.

Yamagishi, T. (2011). Trust: The evolutionary game of mind and society. New York: Springer Science + Business Media.

Walzer, M. (1983). Spheres of justice: A defense of pluralism and equality. New York: Basic books.

Zeng, J. (2020). Artificial intelligence and China's authoritarian governance. International Affairs, 96(6), 1441-1459.

Zhang, C. (2020). Governing (through) trustworthiness: technologies of power and subjectification in China's Social Credit System. Critical Asian Studies, 52(4), 565-588.

3

Beyond legislation and technological design: The importance and implications of institutional trust for privacy issues of digital contact tracing

For proper implementation of digital contact tracing technologies for fighting against SARS-CoV-2, participants' privacy vulnerability and the uncertainty from the relevant institutions' side could be seen as two core elements that should be dealt with, among others. In this chapter, we propose to understand the current approaches for preserving privacy, referred to as privacy by legislation and privacy by technological design, as distrusting strategies that primarily work to reduce participants' vulnerability by specifying and implementing privacy standards related to this digital solution. We point out that mere distrusting strategies are insufficient for ethically appropriate develop-

This chapter is based on the following article:

Teng, Y.,& Song, Y. (under review) Beyond legislation and technological design: the importance and implications of institutional trust for privacy issues of digital contact tracing. *Science and Engineering Ethics.*

Author contributions: YT structured this chapter and took the lead in the writing of the manuscript. YS wrote Section 3.3.2 and contributed to Section 3.2.1. Both authors have done multiple revisions of the draft chapter.

ment of this digital solution, nor can they eliminate the need for institutional trust that plays an essential role in fostering voluntary support for this solution. To reach well-grounded trust in both ethical and epistemological sense, we argue that trust in institutions concerning personal data protection in the case of digital contact tracing ought to be built on the relevant institutions and individuals' goodwill towards the public and their competence in improving the actual effectiveness of this solution. We conclude by clarifying three aspects, including the purpose, procedure, and outcome, where the relevant trustees can work to signal and justify their intentions and increase their trustworthiness. Given the complementary qualities shown by the distrusting strategies and trusting strategies, a combined strategy including both sorts seems closer to what we expect from the responsible implementation of this digital solution, which could also improve the effectiveness of this institutional response.

3.1. Introduction: The deficit of institutional trust as part of the privacy issues

A great deal of research has shown that digital contact tracing technologies, as a supplement to conventional tracing measures, can play a positive role in strategies for easing intense lockdown measures against coronavirus disease 2019 (COVID-19) (Abueg et al. 2020; Hinch et al. 2020). While contact tracing is a crucial way to break the chain of transmission by finding and notifying people who have been in close proximity with symptomatic patients to take further measures (e.g., test and self-quarantine), the large number of pre-symptomatic infections and fast speed of SARS-CoV-2 transmission have posed grave difficulties in doing this manually. By using instant signals of smartphones, digital technologies promise to improve the efficacy of the tracing processes by minimizing the time to find, notify, and quarantine the contacts at risk, contributing to breaking the overall transmission chain (Ferretti et al. 2020).

Along with the prominence achieved by the efficacy promise of digital contact tracing, a considerable effort has also been made to address the privacy issues of this information-based solution. Given the current stage of using technological design and legislation to protect personal data, it is fair to say that the consensus on developing opt-in, privacy-preserving tracing apps has almost been reached in democratic societies (Bengio et al., 2020). However, the low uptake of many privacy-preserving initiatives shows an intractable issue of this digital solution: citizens' general lack of trust in institutions with respect to personal data protection (de la Garza 2020; Sim and Lim 2020).

According to Ogury's research, more than half of the respondents found in the US, France, and the UK state that they do not trust their government to protect any data they share through the tracing apps (O'Halloran 2020). As the apps developed and used in these countries contain distinct settings (e.g., decentralized or centralized databases), this study concludes that users' trust in government needs to be rebuilt no matter which basic technique a tracing app is built on. It won't be surprising that people are more apprehensive about digital tracing when private, commercial companies are stepping in and functioning as a separate "data controller" from, or even a "gatekeeper" for public agencies, even if some of these companies' products (e.g., Apple and Google's exposure notification service) contain higher privacy requirements than nation-state-based initiatives. Considering their profit-minded shareholders and notorious records of data breaches, as Bradford et al. (2020) put forward, any uncertainty health providers and patients hold towards the future use of the medical data shared through or issued by third parties could discourage the uptake of the apps, making people constantly sceptical about the privacy promises made by these institutions on their tracing apps.

The fact that current privacy-preserving apps fall short of fostering residents' participation delivers a deeper public concern over institutions' intentions of promoting the apps. Namely, people are anxious that the relevant institutions and power holders

are promoting the tracing technologies primarily for business and political interests instead of putting the citizens' best at heart. Such a worry can lead to a serious challenge to the tracing apps, namely, people may hardly trust that the institutions will protect their data even though many high privacy requirements and promises of data protection have been made publicly. As Kreps et al. (2020) point out, in the absence of institutional trust, citizens may just *not* perceive the privacy-preserving apps as privacy-preserving, and thus obviate the apps to control unwanted privacy losses. Combined with the significant role of trust in impacting the adoption of new technologies shown by various empirical studies (Bahmanziari et al. 2003; Dhagarra et al. 2020; Choi and Ji 2015), the low uptake of those tracing initiatives with a lower level of trust seems to be a coherent result.

Taking the above impacts of trust into consideration, we provide a novel perspective to examine privacy issues related to contact tracing technologies by viewing the lack of institutional trust as part of, rather than an additional issue that is separate from, users' privacy anxieties associated with this institutional response. A holistic understanding of distinct strategies and approaches that help address users' privacy concerns is provided, with a particular focus on clarifying the importance and implications of moral trust relations based on goodwill. As such, rather than asking general questions of trust, we start from the descriptive aspect of institutional trust (i.e., its impacts on technology adoption) and then delve into the discussion on its normative aspect (i.e., what makes justified trust decisions), seeking to explore what is at stake morally in the relation between citizens and institutions beyond the privacy issues on the surface, and the potential ways that can help release such tension.

3.2. Exiting lockdowns: Why or why not digital contact tracing?

While it is clear that, without mass vaccinations and specific therapeutics, any measure used for easing intense lockdowns (e.g., the closure of business and movement restrictions) should be assessed cautiously, it is unclear which set of measures is most effective. One formula that seems successful when used in Singapore, China, and South Korea is the so-called "test-trace-isolate" strategy. The purpose of this strategy is to allow the gradual reopening of economic and social activities in the prevention of overwhelming the health care system and a potential next COVID-19 wave (The Economist 2020).[1] Each step included in this strategy is considered crucial and indispensable. According to Aleta et al. (2020), for example, a resurgence of the epidemic can be prevented when 50% of symptomatic cases are identified by tests, 40% of their contacts are traced, and all of the contacts are then quarantined for two weeks. In a comparative study done by Panovska-Griffiths et al. (2020), these figures are 59-87%, 40%, and 75% respectively.

[1] It is worth noting that this chapter was first finished in May, 2020. As the situations of the pandemic, including the digital technologies used, policies, and the vaccination campaign might be changed, some information can be outdated.

3.2.1. Digital contact tracing and its role in the overall strategy

Considering the intractable features SARS-CoV-2, while the conventional way of tracing relying on tracers remains necessary in the fight, it is shown to be insufficient to find enough contacts without delays (Ferretti et al. 2020). This means that other supplement tools for tracing are needed in order to make the overall strategy useful (Kretzschmar et al. 2020). As mentioned, digital technologies based on smartphones might play a role here. Despite that the more citizens use the apps, the more effective the apps might perform, the efficacy of digital tracing apps is not a binary off-on switch. Research has shown that digital tracing combined with other containment measures can contribute to the reduction of infections, deaths, and hospitalizations at almost any level of uptake rate (O'Neill 2020; Hinch et al. 2020). For example, even only 15% of people use the apps, according to Abuge et al. (2020)'s model, they can reduce around 8% of infections and 6% of deaths. Thus, the goals of digital tracing are quite clear: (1) to find the contacts being overlooked in specific instances by traditional tracing and contribute to the total number of contacts required by the overall strategy to be useful; and (2) to provide rapid notification of exposures to reduce delays occurring between individuals being exposed and being tested or quarantined.

Considering the efficacy of digital tracing, by March 16, 2020, around 49 countries have launched their mobile-assisted tracing apps.[2] Most of these apps implement digital tracing by two technologies: Global Positioning System (GPS) and Bluetooth (low energy mode). Apps that use GPS technology collect users' location data and use a central server to analyse whether the location information of the app users overlays with the spots of those positively tested patients at a similar time (Gaur 2020). The apps will then alert the direct or indirect contacts accordingly. Apps that use Bluetooth seek to achieve similar goal of alerting potentially infected people, but by swapping anonymous codes with other app users when they are nearby at a certain distance (e.g., 3 meters) for a certain period (e.g., 15 minutes) (Kelion 2020). Based on the code switch history, users will get a notification when their contacts upload positive diagnosis. In terms of the general goals of digital tracing discussed, both technologies can contribute to finding more exposed people and shortening the time of notifying and isolating these people and their contacts.[3]

3.2.2. Privacy concerns over early tracing apps

However, concerns over these technical solutions have also been voiced. Many researchers have debated on the practical issues related to these apps and their countermeasures, such as the situations that can cause false-alarms and unnecessary panic, civil compliance to voluntary self-quarantine, and the reliance on high-quality mobile devices (Servick 2020; Chandler 2020; Kumar Radcliffe 2020). For the interest of this chapter, in this subsection, we mainly take a look at the privacy issues associated with early tracing initiatives. To discuss these issues in good or-

[2]The source of the figure: https://public.flourish.studio/visualisation/2241702/?utm_source=showcaseu tm_campaign=visualisation/2241702.

[3]It should be noted that the apps may contain other purposes that hinge on different interests.

der, here we use the clarification of the privacy concept provided by Warnier et al. (2015) as a simple framework to structure our discussion. While there are different conceptions of the privacy concept in philosophy, the three interconnected aspects of privacy they propose seem to nicely capture the most intractable issues faced by the poorly designed apps.

Consider first freedom from intrusion. Although none of the apps are compulsory to be downloaded, some are strictly linked to other aspects of human life, such as travel and entering public spaces. For instance, Health QR Codes were widely used as an electronic certificate for public transportation and activity permits inside some countries (Bonsall et al. 2020). Similarly, tourists to South Korea are required to install Self-Check when purchasing tickets and report their health condition through the app for 14 days after arriving (Kim 2020). By binding app installation with the permission to social activities, both cases are in tension with privacy as an effort striving for freedom from external constraints and render the apps de-facto mandatory (Ranisch et al. 2020).

Consider second the control of personal data. Having control over information concerns the restriction of information flow and whether it flows properly (Nissenbaum 2015). Tracing initiatives, such as Singapore's TraceTogether and Norway's early Smittestopp, apply central servers to store and analyse the uploaded anonymous data, which enable the authorities to gain more insight into epidemic responses. Nevertheless, data aggregation not only contains the risk of being hacked and divulged but also threatens users' right to control over the flow of personal data and increases the risk of "function creep" since the authorities might abuse their power and illegitimately use the contact tracing data for other purposes such as law enforcement (East and Africa 2020).

Consider third freedom from surveillance. Data gathered by central servers might also be used for surveillance purposes, particularly considering those initiatives that collect vast location data and unnecessary personal information, such as gender, age, and profession (Clarance 2020; Johnson 2020). The comprehensive information collection makes it possible for the authorities to produce big-data-driven policies to mitigate or suppress the contagion. However, a combination of the behaviour-related information (e.g., locations and payment history) and identity information can be illegitimately used to not only track, watch, and follow a specific person's movement and travel history but also analyse implicit information linked to other characteristics and inner lives of the data subject (e.g., sexual orientation).

In times of public health crisis, while it is clear that measures that could contribute to "flattening the curve" are urgently required, it remains unclear how much privacy should be traded off in the name of community needs and to what extent governments' expansion of surveillance power can be justified (Sharon 2020). Such trade-offs are inextricably linked to the social-political contexts to which the apps are applied. Nevertheless, some obvious privacy flaws, such as the collection of unnecessary information and the analysis of behaviour-related information, should

be avoided by any tracing initiative for the sake of reducing unnecessary privacy costs. The pragmatic and epistemic weakness of citizens arguably creates an obligation of institutions to ameliorate the imbalanced situation and prevent from taking more advantage of the participants. In the next section, we begin by introducing two sorts of strategies related to trust that can help assuage the tension between citizens and institutions caused by the adoption of the apps. With this structure, we then take a closer look at the prevalent approaches for addressing the privacy concerns, setting the stage for analysing the value and implications of institutional trust.

3.3. Distrusting strategies: Current approaches for reducing vulnerability

Essentially, the privacy issues discussed above concern two main elements: users' vulnerability related to personal data and the uncertainty about how the relevant institutions may manage users' data. Relations that involve these two elements are exactly the situations where trust becomes most relevant (Nickel 2015; Becker 1996; Luhmann 1979). As an attitude of the trustor (X), trust typically develops in situations where X has the need or interest to rely on a trustee (Y) with respect to the fulfilment of a particular entrusted thing (Z), but X cannot fully control or predict the behaviour of Y (McLeod 2015). Here Z and other potential losses of X caused by Y's behaviour can be seen as the vulnerability of X, and the essential reason for X's vulnerable position is that X is uncertain about Y's real trustworthiness. These two commonalities indicate that the case of digital contact tracing is a plausible situation where citizens' trust in institutions can be relevant and cause real effects on app adoption.

3.3.1. Two sorts of strategies related to trust

As Heimer (2001) clarifies, there are two sorts of strategies that are particularly useful for facilitating more reliable interactions under conditions of vulnerability and uncertainty. The first is *trusting strategies* that seek to find more information about Y's competence and intentions to decrease uncertainty about Y's trustworthiness. If the information at hand suggests that Y is competent and bears goodwill towards X, X will likely trust Y to protect rather than harm the thing X cares about. The conception of trust used here assumes the trustee's goodwill as a basic characteristic of trust relations, which essentially distinguishes trust from reliance by justifying feelings of betrayal and the expectation that Y will take X's vulnerability into account favourably (Baier 1986; Jones 1996). In this case, the reduction of X's vulnerability is less necessary and perhaps undesirable since X believes that Y has the moral capacity to be responsive to X's considerations.

Conversely, if finding enough information is not available or costs too much time, energy, and resources, people might opt for *distrusting strategies* that strive to limit others' untrustworthy actions and reduce the vulnerability of themselves, for example, by making contracts, more specific market access standards, and terms

for sanctions and compensation. These measures, when serving for the purpose of limiting improper actions, provide warranties and guarantees to participants who have a stake in the interaction, leading to compliance and reliance that are often used as alternative or complementary approaches to trustworthiness and trust (Kerasidou 2016).[4]

In the context of digital tracing, getting sufficient information about the relevant institutions and individuals' trustworthiness seems not easy for ordinary people. This is because many citizens lack the knowledge and capability to rationally assess the relevant entities' competence, nor can they easily find ways to be aware of the actual intentions of these entities. In most cases, ordinary people cannot even find someone to whom their uncertainty can be directed due to the complex division of labour in such a nation-state-based or transnational solution. This also explains why in modern societies, strict measures, like legislation, contracts, and insurance, that do not rely on one's familiarities of another's intentions and competence are used more often between strangers (O'Neill 2002).

We argue that the prevalent approaches adopted to address the privacy issues, referred roughly to as privacy by legislation and privacy by technology (as we will discuss below), are closer to distrusting strategies rather than trusting strategies. The essential idea of these two approaches is to utilize legal and technical means to specify and implement a complex set of privacy requirements, such as data parsimony and data anonymization, formalizing the way that users' vulnerability can be reduced in the context of digital contact tracing. Here users' vulnerability is the direct and indirect information-related risk engendered by using the apps, including harms, injustice, and inequalities caused by the disclosure of diagnosis information or other data issued by the apps. As the question of what personal data might be at stake is largely determined by the kind of underlying technologies chosen by different apps and a complex set of criteria applied to regulate the life cycle of the apps, these two approaches can be crucial ways to ameliorate users' vulnerability.

3.3.2. Privacy by legislation and privacy by technological design

Privacy by legislation refers to the idea of protecting participants' vulnerability by the enactment, enforcement, and optimization of data protection laws and regulations. While poorly designed digital tracing apps pose serious threats to users' personal data, stringent privacy laws and regulations make app developers, data controllers, data issuers, and other relevant entities to be legally bound to create privacy-preserving apps to avoid lawsuits, fines, fees, and the loss of reputation (Watts 2020; Gasser et al. 2020).

[4]It is worth noting that this statement does not imply that regulations and industry standards are presented as solely distrusting strategies since they also provide a good starting point and relatively safe environment for cultivating trust.

Table 3.1: A comparison of different apps on their basic technology (√: Applied; ×: Not applied)

		Health QR Code	Aarogya Setu	Smittestopp (Previous)	HaMagen	Trace Together	Stop Covid	Corona Melder	Corona Warn App	NHS Covid-19
Country		China	India	Norway	Israel	Singapore	France	the Netherlands	Germany	the UK
Digital tracing technology	GPS	√	√	√	√	×	×	×	×	×
	Blue-tooth	×	√	√	√	√	√	√	√	√
Contact history storing	Central server	√	√	√	×	√	√	×	×	×
	Local phones	×	√	×	√	√	×	√	√	√

In the EU context, for example, digital tracing falls into General Data Protection Regulation's (GDPR) comprehensive scope that requires system design of digital tracing to demonstrate: lawfulness, fairness, and transparency; purpose limitation; data minimization; accuracy; storage limitation; integrity and confidentiality; and accountability (GDPR Art. 5). The regulation's expansive scope and principle-based approach, as Bradford et al. (2020) argue, offer a ready-made and flexible functional guideline for creating new technology applications that protect basic human rights. The Pan-European Privacy-Preserving Proximity Tracing is a fundamental effort to translate GDPR's general rules into more detailed technical standards for guiding the design and development of the tracing apps in the EU context (Abeler et al., 2020).

To some extent, privacy by technological design can be seen as a means of implementing privacy laws, but it is more than that since design can also be used to incorporate various norms and values into the product (van den Hoven 2008). Can we make the design of the tracing apps more ethically appropriate beyond what is required by laws and regulations? Based on our previous introduction of the core underlying technologies that enable a tracing app, it can be said that Bluetooth plus locally data storage embroil less privacy costs than other options since the former set collects almost no identifiable data, except for positive diagnosis that is already known by public health authorities. To provide more information, Table 3.1 provides a review of different apps on their basic technical settings.[5]

To be more specific, while the appropriate use of location data collected by GPS-based apps relies heavily on legal constraints, industrial standards, and central authorities' responsibility for processing data in a lawful and secure manner, Bluetooth-based apps weaken the identifiability at the technology level without the aid of legislation and bureaucratic structures. For this reason, Bluetooth-based apps could be considered as a product of both privacy by legislation and technological design. Likewise, while data protection in centralized servers relies heavily on privacy laws and regulations, decentralized databases that keep the exchanged identifiers merely on users' phones can reduce the reliance on centralized organizations and bureaucratic structures to protect data, which could also be regarded as a product

[5]For the technical settings of COVID-tracing apps, see https://craiedl.ca/gpaw/.

of both privacy by legislation and privacy by technology.

Granted, many privacy concerns over the violation of users' freedom from intrusion, control of information, and freedom from surveillance present by early tracing initiatives have been addressed by opt-in, Bluetooth-, and decentralization-based contact tracing apps together with other institutional privacy assurances. The uptake rate of such a privacy-preserving solution seems not as high as expected, even though it has already been higher than that of more intrusive solutions.[6] Many citizens' privacy anxieties still exist, despite the apps' vulnerability-reducing settings and the recommendation of participation appealed by public health authorities, governments, and privacy experts.

3.4. Dealing with uncertainties: Institutional trust and trustworthy institutions

While the current approaches discussed are necessary for providing a good starting point for proper implementation of digital contact tracing, they are not in themselves sufficient to facilitate the uptake of contact tracing technologies, nor can they eliminate the role of trust in fostering or impeding widespread voluntary adoption of digital contact tracing technologies. Besides, there is a danger that the situation of trust deficit may be exacerbated by the distrusting strategies adopted. As O'Neill (2002) and Thompson (2013) point out, trying to increase uptake merely by using regulatory approaches to limit some untrustworthy conduct shows the very idea of "economizing on trust", which may squeeze out the role played by trust, including its positive correlation with technology adoption.

An important reason for the insufficiency of current approaches is that although many privacy standards have been set and put into practice, there is a wide variety of nuanced, implicit, and unforeseen situations that may engender privacy risks but have yet to be covered by regulatory measures and technological solutions. For example, once diagnosis information is divulged, a broad sense of social avoidance, discrimination, bias, and other information-based harms might be imposed on the infectious, and some of these harms can neither be fully addressed nor equally compensated by the above measures. This means, even though the measures discussed can mitigate the power imbalance between citizens and power holders by specifying and regulating the latter's actions, being participants still directly points to people's privacy-related vulnerabilities that they would otherwise not take.

Following Heimer (2001) and Kerasidou (2016)'s identification discussed, the uncertainties of the institutions' side involved in the case of digital contact tracing are just the places where warranted institutional trust can play a role in encouraging uptake. That is to say, other things being equal, people will likely only choose to participate and cooperate with those institutions that favourably take into account their vulnerabilities and act as counted on; namely, those institutions that they think

[6]For uptake of contact tracing initiatives in 2020, see https://craiedl.ca/gpaw/.

are trustworthy and can really trust.

Understanding the value of trust provides an important step towards the establishment of healthy citizen-to-institution relations. Citizens' trust is a good thing for institutions to implement pandemic responses, but it should be clear that trust is not something that can be enforced or demanded. The best device to gain trust is by improving the potential trustee's trustworthiness, which makes trust easier to flourish (Hardin 2001). From the trustor's perspective, although citizens generally have the need for being protected at the collective level (Falcone et al. 2020), being too trusting comes with a considerable risk of generating false expectations and losing the entrusted things. As Devine et al. (2020) state, people may naively believe that institutions are doing the right thing or doing things in the right way when they are not. Due to the moral sensitivity of the entrusted things (e.g., illness history) and the irreversible harm that might be inflicted on data subjects once trust is frustrated, trust in the case of digital contact tracing should not be seen as something that can be unreflectively developed. Instead, a critical, ethical view of trust should be supposed.

Considering the above bilateral need for trust in the pandemic context, the normative question of what makes an institution trustworthy is of importance for both parties. To answer this question, we need to explicate what elements well-grounded trust ought to concern in the case of digital tracing and how such opinions can be applied to the improvement of institutions' trustworthiness, with the practical goals of making trust more warranted and the outcome of contact tracing technologies more morally desirable.

The first element, that is also the core one, is the associated trustees' *motives* for privacy protection in the case of digital contact tracing. In moral philosophy, trust is often considered as a distinctive concept assuming that the trustee bears goodwill towards the trustor and would like to take the trustor's vulnerability and dependence as compelling reasons for acting responsively, whereas reliance does not require so and is seen as a mere rational-decision based on the result of risk-benefit assessment (Baier 1986; Jones 1999; Jones 2012). In the case of digital tracing, a distinction can thus be made between preserving privacy as an instrument for achieving other ends set by the relevant institutions and preserving privacy out of genuine care that sees individuals' privacy rights as part of the desired end. Viewing privacy as something intrinsically valuable and worthy of promoting and preserving fundamentally explains why by trusting, people (X) feel optimistic that the associated institution and individuals (Y) are committed to protecting their personal data (Z) issued by the app related to Y even in situations out of the protection of current legal and technical solutions, since they believe that the trustees will take their vulnerabilities into account favorably and act as counted on.

From this perspective, creating law-compliance tracing apps out of some morally controversial reasons, such as self-interest, fear of sanctions or opprobrium, and force of social constraints, seems not sufficient to guarantee the institutions' future

trustworthy conduct. Reliable actions with motivation open to different contexts might be enough for citizens to interact with institutions in regular situations; however, in the pandemic context where people are already worried and anxious about the surroundings, more benign motives are arguably needed to improve the predictability of the outcome and comfort the sense of insecurity caused by the turbulence. Furthermore, motives governed by business practices and market thinking, when applied to other social spheres (e.g., public health), may jeopardize or simply crowd out non-material social good and moral values internal to those particular spheres (Sharon 2020; van den Hoven 2008), resulting in a violation of justice and equality that further washes away the desirable grounds for building trust. Likewise, politicians and government leaders are criticized for putting political interests ahead of what the public cares about (Schmitt 2020; Dimock 2020). As Floridi (2020) points out, in some cases, the development of the apps is not motivated by a public health standpoint, but it is rather a mere political solution that signals to the public that power holders have tried everything they can and should not be blamed for not trying.

These commercial and political opportunisms can raise the public's fear that the privacy promises and the actions that appear to be trustworthy are just means for achieving other ends of those power holders, which might be broken at a certain point. In fact, scandals of data breaches have been witnessed several times in the case of digital proximity tracings, such as the North Dakota's tracing app where studies find that personal data has been sent to Google and other service providers with the app's privacy promises being ignored (Melendez 2020). Similarly, Israel's national security agency is reported to have the power to access the database of Israel's tracing app HaMagen for surveillance purposes despite the app's promise that users' data will not be transmitted to third parties (Winer and Staff 2020). Such promise-breaking incidents may further undermine the public's image of the tracing apps in general. The moral apprehension about what motivates one to make a privacy promise is real. Such a concern, combined with the gradually strict rules adopted to regulate untrustworthy actions, may create a circular, self-reinforcing atmosphere of distrust that leaves little space for trust to thrive.

The second element constitutive of well-grounded trust concerns the awareness of the relevant trustees' *competence*. The evaluation of such competence mainly includes two aspects: whether users' personal data is well protected by a privacy-preserving app and whether the app is effective in achieving the predefined functional goals. While it is difficult for normal users to detect privacy problems until experts find loopholes or the spread of data breach news, the latter ultimately concerns whether developers and policymakers can sufficiently justify the effectiveness of, and the societal need for, the contact tracing apps. It is important to note that the discussion about the function and role of digital tracing technologies provided in early sections is more about the app's efficacy – i.e., how well an app works in a controlled environment, instead of its effectiveness that considers how well the same app will work when it is released in a real-world situation. As Floridi (2020) points out, the privacy issues and effectiveness of the apps, together with other eth-

ical difficulties, need to be carefully assessed by a clear deadline so that we could determine how this digital project ought to be improved, renewed, or terminated. Meanwhile, the relational and situational nature of trust indicates that very often the goods of trust are not inextricably linked to a particular trustee or a particular means used by that trustee (Teng 2021). For this reason, proper justification of the need for the apps should also include a comparison result between a contact tracing app with other alternatives contingent on different contexts and new opportunities.

Based on the will-centred account of institutional trust discussed above, for participants to trust an institution and the associated app, it means that they believe that the institution (1) does care about users' health and privacy right and develop the digital project as a means to improve citizens' well-being; and (2) would like to take possible steps to justify the need for, and improve the effectiveness of, the contact tracing app. Understanding institutional trust in this way does not lead to the fact that this trust is fully warranted given that trust is never fully warranted. Rather, this interpretation sketches the main value and meaning of trust as a complementary approach to legal frameworks and technological solutions. It captures the general expectation we have about what is appropriate for others to do and our shared sense of insecurity about others' motivation that is multiplied by the public health crisis.

3.5. Implications of trusting strategies for digital contact tracing

Till now, we have discussed two sorts of strategies that can be used to reduce the privacy issues related to digital contact tracing and help facilitate interactions between citizens and institutions. To enhance readability, a framework of how these strategies are applied to this digital project is provided in Table 3.2. While it is clear that the choice of strategies largely depends on the context to which they will be applied and probably no countries purely use one kind of the strategies, trusting strategies emphasizing institutions and individuals' intentions have received much less attention than the other type. In this section, we discuss how our moral opinions about institutional trust can be applied to the case of digital contact tracing to restore trust gradually through an improvement of institutions' trustworthiness. Combined with the distrusting approaches articulated, a combined strategy based on all the useful embodiments related to the two strategies seems closer to what we expect from the responsible implementation of this digital solution.

We propose that there are three aspects where institutions and individuals can apply the will-centred trust account to increase their trustworthiness with respect to data protection by signalling and justifying their goodwill towards the public. The first is the purpose aspect. The public's moral apprehension about what drives institutions to foster this digital project urges government and corporate leaders, employees, and app developers and maintainers to be willing and able reliably to show their intentions. To show care towards participants and the society at large,

Table 3.2: A framework for the two sorts of strategies in the case of digital contact tracing

	Central idea	Embodiments	Institutions	The public
Distrusting strategies	Reducing users' vulnerability	Privacy by legislation	Comply with privacy policies, laws, and regulations	Be aware of one's legal rights
		Privacy by technological design	Reduce privacy risks by technological innovations	Be aware of privacy implications made by different technologies
Trusting strategies	Reducing uncertainty about institutions' trustworthiness	Intentions	Display genuine care towards public health and privacy rights	Get information about the potential trustee's intentions
		Competence	Justify the effectiveness of and societal need for the apps	Get information about the potential trustee's competence

the relevant entities need to answer why the development, deployment, and use of contact tracing apps can be considered as a collective effort that can bring positive impacts on pandemic mitigation, as well as how the apps could improve the well-being of individual participants without improper intrusion into their right to be left alone. Answering these questions justifiably would require the relevant entities to provide reliable and understandable information to the public, and demonstrate how their benign intentions will be used to inform the operation processes explicitly, particularly in cases when conflicting interests and alternative solutions are involved.

The second is the procedure aspect. Procedural values – such as transparency, fairness, proportionality, accountability – that are often linked to good institutional responses are considered valuable for developing trust in the context of digital contact tracing (Ranishch et al. 2020; Woodhams 2020). Arguably, what makes the will-centred account of trust distinctive is its emphasis on the show of willingness to negotiate, compromise, and cooperate during the decision-making process. Public trust is not generated in an environment where the public's voice is not heard, even though that environment contains well-established legal frameworks and institutional procedures. Mechanisms built on the willingness to negotiate directly facilitate communication by shifting a certain level of control from power holders to those who are less powerful, enabling the latter to relieve some burden and anxieties of the former's discretionary power over the actual result of trust. Some governments that secure trust successfully during the epidemic have already shown the usefulness of such trusting strategies. For example, public agencies of Taiwan have built multiple platforms that allow citizens to participate in the enactment of public policies, such as the distribution of medical supplies (Chang 2020). The inclusive and interactive ideas involved not just deliver that the authorities do care about citizens' interests and would like to implement the policy responses in a responsive manner, they also inspire civic-mindedness and engagement that are considered crucial for fighting against the pandemic.

The third is the outcome aspect. Probably the most straightforward way to justify the trustee's goodwill towards the trustor is to make the entrusted thing or task warranted, to honor rather than break the privacy promises that invited trust, to show

honesty, empathy, and accountability by taking real actions, to improve the welfare of participants instead of making troubles by sending false alerts and misinformation. The outcome of trust can thus be seen as an important evaluation standard of trust, which directly impacts whether one would like to continue or stop trusting. Nevertheless, this seems to indicate the difficulty of initiating a trust relation, which somehow comes back to the usefulness of distrusting strategies in facilitating tentative interactions by providing a relatively safe route for individuals to depend upon others while gathering information about others' real trustworthiness.

That is to say, in terms of how to deal with public health recommendations and governmental policy responses made for achieving collective goals, it might be useful for citizens to start tentatively from distrusting strategies. For example, one may start by understanding the privacy implications of distinct technological settings used for contact tracing purposes, and by being aware of whether a given initiative defers to data-protection laws and industry self-regulation in advance. Meanwhile, institutions and their representatives should continue to reduce participants' vulnerability as well as signal and justify their intentions and competence, seeking to augment trustworthiness and decrease participants' uncertainties about the overall interaction. Later on, if that participant has sufficient successful experience with the interacted institution that also gains a fine reputation from the society, their distrust might turn into trust that can lead to more effective group functioning and productive social activities.

3.6. Conclusion

While trust, together with institutional procedures, technical settings, and market techniques, form the bedrock of cooperation in modern society, the absence of trust could create considerable difficulties in the execution of any public policy. During the pandemic, we have witnessed a fracturing of trust in many institutions worldwide, but a gradual recognition of the value of trust and the urgency of restoring trust. In this chapter, we have critically engaged with the topic of trust in institutions within the framework of the two sorts of strategies discussed. Distrusting strategies and trusting strategies, considering their central ideas, embodiments, and detailed implications for institutions and citizens, are not merely complementary to each other, but also both considered indispensable for proper and effective implementation of contact tracing technologies. Despite that there are no easy ways to fix trust in a short time, institutions should understand how trust works and work to explicitly improve their trustworthiness.

References

Abeler, J., Bäcker, M., Buermeyer, U., and Zillessen, H. (2020). COVID-19 contact tracing and data protection can go together. JMIR mHealth and uHealth,

8(4).

Abueg, M., Hinch, R., Wu, N., Liu, L., Probert, W. J., Wu, A., ... and Cheng, Z. (2020). Modeling the combined effect of digital exposure notification and non-pharmaceutical interventions on the COVID-19 epidemic in Washington state. medR xiv. Accessed 10 October 2020. https://www.medrxiv.org/content/10.1101/2020.08.29.20184135v1.

Aleta, A., Martín-Corral, D., y Piontti, A. P., Ajelli, M., Litvinova, M., Chinazzi, M., ... and Pentland, A. (2020). Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. Nature Human Behaviour, 4(9), 964-971.

Bahmanziari, T., Pearson, J. M., and Crosby, L. (2003). Is trust important in technology adoption? A policy capturing approach. Journal of Computer Information Systems, 43(4), 46-54.

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

Becker, L. C. (1996). Trust as noncognitive security about motives. Ethics, 107(1), 43-61.

Bengio, Y., Jandan, R., Yu, Y.W., Ippolito, D., Jarvie, M., Pilat, D., Struck, B., Krastev, S. and Sharma, A. (2020). The need for privacy with public digital contact tracing during the COVID-19 pandemic. The Lancet Digital Health, 7(2), 342-344.

Bonsall, D., Parker, M., and Fraser C. (2020). Sustainable containment of COVID-19 using smartphones in China: Scientific and ethical underpinnings for implementation of similar approaches in other settings. Accessed 20 May 2020. https://int.nyt.
com/data/documenthelper/6825-coronavirus-app-proposal-UK/76650ed3f 249bf888 f1e/optimized/full.pdf.

Bradford, L. R., Aboy, M., and Liddell, K. (2020). COVID-19 Contact tracing apps: A stress test for privacy, the GDPR and data protection regimes. Journal of Law and the Biosciences, 7(1).

Chandler, S. (2020). New ultrasonic contact-tracing app promises better accuracy than bluetooth alternatives. Forbes. Accessed 5 October 2020. https://www.forbes.com/sites/simonchandler/2020/05/26/new-ultrasonic-contact-tracing-app-promises-better-accuracy-than-bluetooth-alternatives/2f6cfba62122.

Chang, D. (2020). What coronavirus success of Taiwan and Iceland has in common. The Conversation. Accessed 9 September 2020. https://theconversation.com/what-coronavirus-success-of-taiwan-and-iceland-has-in-common-140455.

Choi, J. K., and Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. International Journal of Human-Computer Interaction, 31(10), 692-702.

Clarance, A. (2020). Aarogya Setu: Why India's Covid-19 contact tracing app is controversial. BBC. Accessed 20 May 2020 from https://www.bbc.com/news/world-asia-india-52659520/.

de la Garza, A. (2020). Contact tracing apps were big tech's best idea for fighting COVID-19. Why haven't they helped? Time. Accessed 8 December 2020. https://time.com/5905772/covid-19-contact-tracing-apps/.

Devine, D., Gaskell, J., Jennings, W., and Stoker, G. (2020). Trust and the

Coronavirus Pandemic: What are the consequences of and for trust? An early review of the literature. Political Studies Review, 1-12.

Dhagarra, D., Goswami, M., and Kumar, G. (2020). Impact of trust and privacy concerns on technology acceptance in healthcare: An Indian perspective. International journal of medical informatics, 141, 104164.

Dimock, M. (2020). How Americans view trust, facts, and democracy today. Assessed 10 December. https://www.pewtrusts.org/en/trust/archive/winter-2020/how-americans-view-trust-facts-and-democracy-today.

East, M. and Africa, N. (2020). Bahrain, Kuwait and Norway contact tracing apps among most dangerous for privacy. Amnesty International. Accessed 5 October 2020. https://www.amnesty.org/en/latest/news/ 2020/06/bahrain-kuwait-norway-contact-tracing-apps-danger-for-privacy/.

Falcone, R., Colì, E., Felletti, S., Sapienza, A., Castelfranchi, C., and Paglieri, F. (2020). All we need is trust: How the COVID-19 outbreak reconfigured trust in Italian public institutions. Frontiers in Psychology, 11.

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., ... and Fraser, C. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science.

Floridi, L. (2020). Mind the app – considerations on the ethical risks of COVID-19 apps. Accessed 10 November 2020. https://thephilosophyofinformation.blogspot.com/2020/04/mind-app-considerations-on-ethical.html.

Gasser, U., Ienca, M., Scheibner, J., Sleigh, J., and Vayena, E. (2020). Digital tools against COVID-19: taxonomy, ethical challenges, and navigation aid. The Lancet Digital Health, 8(2): e425-e434.

Gaur, A. (2020). Can Aarogya Setu beat the virus? Accessed 5 October 2020. https://timesofindia.indiatimes.com/india/can-aarogya-setu-beat-the-virus/articleshow/75314677.cms.

GDPR. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Assessed 6 May 2020. https://gdpr-info.eu/.

Halloran, J. (2020). Consistent trust gap in contact-tracing apps in US, Europe. Computer Weekly. Accessed 4 September 2020. https://www.computerweekly.com/news/252486058/Consistent-trust-gap-in-contact-tracing-apps-in-US-Europe.

Hardin, R. (2001). Conceptions and explanations of trust. In K. S. Cook (Ed.), Trust in Society (pp. 3–39). New York: Russell Sage Foundation.

Heimer, C. A. (2001). Solving the problem of trust. In K. S. Cook (Ed.), Trust in Society (pp. 40-88). New York: Russell Sage Foundation.

Hinch, R., Probert, W., Nurtay, A., Kendall, M., Wymant, C., Hall, M., and Fraser, C. (2020). Effective configurations of a digital contact tracing app: A report to NHSX. Accessed 10 September 2020. https://cdn.theconversation.com/static_files/files/1009/Report_-_Effective_App_Configurations.pdf? 1587531217.

Johnson, K. (2020). What privacy-preserving coronavirus tracing apps need to succeed. VentureBest. https://venturebeat.com/2020/04/13/what-privacy-preserving-coronavirus-tracing-apps-need-to-succeed/.

Jones, K. (1996). Trust as an affective attitude. Ethics, 107(1), 4-25.

Jones, K. (1999). Second-hand moral knowledge. The Journal of Philosophy, 96(2), 55-7.

Jones, K. (2012). Trustworthiness. Ethics, 123(1), 61-85.

Kelion, L. (2020). Coronavirus: Apple and France in stand-off over contact-tracing app. BBC. Accessed 5 October 2020. https://www.bbc.com/news/technology-52366129.

Kerasidou, A. (2016). Trust me, I'm a researcher!: The role of trust in biomedical research. Medicine, Health Care and Philosophy, 20(1), 43-50.

Kim, M.S. (2020). South Korea is watching quarantined citizens with a smartphone app. MIT Technology Review. Accessed 5 October 2020. https://www.technologyreview.com/2020/03/06/905459/coronavirus-south-korea-smartphone-app-quarantine/.

Kreps, S., Mcmurry, N., and Zhang B, B. (2020). Americans don't trust contact tracing apps. Here's how we can fix that. Fortune. Accessed 1 September 2020. https://fortune.com/2020/08/17/contact-tracing-privacy-coronavirus-google-apple/.

Kretzschmar, M. E. , G.Rozhnova, M. C.Bootsma, M.van Boven, J. H.van de Wijgert, and M. J.Bonten. (2020). Impact of delays on effectiveness of contact tracing strategies for COVID-19: A modelling study. The Lancet Public Health 5 (8): e452–e459.

Kumar, D., and Radcliffe, P. (2020). False positives, false negatives: It's hard to say if the COVIDSafe app can overcome its shortcomings. The Conversation. Accessed 5 October 2020. https://theconversation.com/false-positives-false-negatives-its-ha
rd-to-say-if-the-covidsafe-app-can-overcome-its-shortcomings-138129.

Luhmann, N. (1979). Trust and Power. Chichester: John Wiley.

McLeod, C. (2015). Trust. Accessed 1 September 2020. http://plato.stanford.edu/archives/fall2015/entries/trust/.

Melendez, S. (2020). North Dakota's COVID-19 app has been sending data to Foursquare and Google. FastCompany. Accessed 2 September 2020. https://www.fastcompany.com/90508044/north-dakotas-covid-19-app-has-been-sending-data-to-foursquare-and-google.

Nickel, P. J. (2015). Design for the value of trust. In J. van den Hoven, P. E. Vermaas, and I. van de Poel (Eds.), Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains, 551-567.

Nissenbaum, H. (2015). Respect for context as a benchmark for privacy online: What it is and isn't. In B. Roessler and D. Mokrosinska (Eds.), Social Dimensions of Privacy: Interdisciplinary Perspectives (pp. 278–302). Cambridge: Cambridge University Press.

O'Neill, O. (2002). Autonomy and Trust in Bioethics. Cambridge: Cambridge University Press.

O'Neill, P. H. (2020). No, coronavirus apps don't need 60

Panovska-Griffiths, J., Kerr, C. C., Stuart, R. M., Mistry, D., Klein, D. J., Viner, R. M., and Bonell, C. (2020). Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second COVID-19 epidemic wave in the UK: A modelling study. The Lancet Child and

Adolescent Health.

Ranisch, R., Nijsingh, N., Ballantyne, A., van Bergen, A., Buyx, A., Friedrich, O., ... and Wild, V. (2020). Digital contact tracing and exposure notification: Ethical guidance for trustworthy pandemic management. Ethics and Information Technology, 1-10.

Schmitt, M. (2020). In the wake of its COVID-19 failure, how do we restore trust in government? New American. Accessed 10 September 2020. https://www.newamerica.org/political-reform/reports/politics-policymaking/in-the-wake-of-its-covid-19-failure-how-do-we-restore-trust-in-government/.

Servick, K. (2020). COVID-19 contact tracing apps are coming to a phone near you. How will we know whether they work? Accessed 5 October 2020. https://www.sciencemag.org/news/2020/05/countries-around-world-are-rolling-out-contact-tracing-apps-contain-coronavirus-how.

Sharon, T. (2020). Blind-sided by privacy? Digital contact tracing, the Apple/Google API and big tech's newfound role as global health policy makers. Ethics and Information Technology.

Sim, D., and Lim, K. (2020). Coronavirus: why aren't Singapore residents using the TraceTogether contact-tracing app? The Coronavirus Pandemic. Accessed 9 September 2020.https://lkyspp.nus.edu.sg/docs/default-source/ips/scmp_coronavirus-why-arent-singapore-residents-using-the-tracetogether-contact-tracing-app_180520.pdf.

Teng, Y. (2021). Towards trustworthy blockchains: Normative reflections on blockchain-enabled virtual institutions. Ethics and Information Technology, 1-13.

The Economist. 2020. The right way to leave lockdown. The Economist. Accessed 9 September 2020. https://www.economist.com/leaders/2020/04/18/fumbling-for-the-exit-strategy.

Thompson, J. B. (2013). Political Scandal: Power and Visibility in the Media Age. John Wiley and Sons.

van den Hoven, J. (2008). Information technology, privacy, and the protection of personal data. In J. van den Hoven and J. Weckert (Eds.), Information Technology and Moral Philosophy, 301-321.

Warnier, M., Dechesne, F., and Brazier, F. (2015). Design for the value of privacy. In J. van den Hoven, P. E. Vermaas, and I. van de Poel (Eds.), Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains. Springer, Dordrecht, 431-445.

Watts, D. (2020). COVIDSafe, Australia's digital contact tracing app: The legal issues. SSRN. Accessed 5 October 2020. https://ssrn.com/abstract=3591622.

Winer, S., and Staff, T. (2020). High Court: Shin Bet surveillance of virus carriers must be enshrined in law. The Times of Israel. Accessed 15 May. https://www.timesofisrael.com/high-court-shin-bet-surveillance-of-virus-carriers-must-be-enshrined-in-law/.

Woodhams, S. (2020). COVID-19 digital rights tracker. Accessed 12 December 2020. https://www.top10vpn.com/research/investigations/covid-19-digital-rights-tracker/.

4

What does it mean to trust blockchain technology?

This chapter argues that the widespread belief that user-blockchain interactions are trust-free is inaccurate and misleading as it not only overlooks the vital role played by trust in the lack of knowledge and control but also conceals the moral and normative relevance of relying on blockchain applications. The chapter reaches this argument by providing a close philosophical examination of the concept referred to as trust in blockchain technology, clarifying the trustor group, the structure, the normatively loaded nature, and the risk of this form of trust relation. To understand the rich expectations that people hold toward blockchain-based systems, the crucial role played by the appropriateness granted to the normative values built into the system is highlighted. Given the vulnerable position of the trustor, the actual trustworthiness of blockchain applications for realizing these values should be scrutinized before the placement of trust. With such concern, the chapter ends by critically reflecting on two of the most promising values that can invite users' trust in blockchain technology, arguing that trust built on these values is risky due to the moral and technical limits involved in current blockchain applications.

This chapter is based on the following article:

Teng, Y. (under review). What does it mean to trust blockchain technology? *MetaPhilosophy.*

4.1. Introduction

If we think of traditional institutions as the trusted intermediaries that help human cooperation progress from direct reciprocity (i.e., an eye for an eye) to indirect reciprocity on which more sophisticated cooperation mechanisms can be built, then blockchain technology appears to be a possible means to establish a new basis of truth and trust without the need for any third party (van den Hoven et al. 2019). Traditionally, online interactions between heterogeneous participants are facilitated by trusted third-party authorities, such as financial institutions and legal branches. As the distributed database technology behind Bitcoin (i.e., a cryptocurrency), blockchain technology came to prominence as a decentralized solution that relies instead on consensus algorithms and rules to ensure the validity and immutability of transactions processed by the peer-to-peer network (Nakamoto 2008). With such decentralized nature, the original blockchain can perform as a virtual institution that users can directly rely upon and interact with, which may significantly reduce the risk, uncertainty, and cost caused by trusting third parties.

While the blockchain's designed attempt for eliminating the need for trusting third parties is distinct, the ways of characterizing the role of trust played in blockchain-based interactions are fairly controversial in literature (Jacobs 2020). As we see that some describe blockchain technology as trustless or trust-free (Nakamoto 2008; Glaser 2017), some capture the change of trust enabled by the technology as a shift of trust from third parties to the system and its underlying algorithms (Simser 2015; Velasco 2017; van Lier 2017), and others depict the change as trust distributed among developers and miners (Werbach 2018; Kasireddy 2018).[1] While each of the efforts partially captures the idea of how blockchains change the way we trust, they fall short of structuring a relatively complete picture of the blockchain-enabled trust revolution. The ambiguity involved not only presents difficulties for developers, regulators, users, and the public to reflect on the value, the targets, and the corresponding risk of talking about trust in the context of blockchain technology but also conceals the moral and normative relevance of blockchain-powered solutions. A systematic analysis teasing out the intertwined relationship between trust and blockchains is needed.

To this end, it is important to first make clear what causes such a divergence in understanding the blockchain-enabled trust revolution. Two reasons appear to be germane when we examine this issue from the perspective of the trust phenomenon: (1) one shortcoming of the current discussion is the absence of a clearly defined group of trustors (or the persons who give trust). Due to the complex relationship between knowledge and trust, the lack of a clearly defined trustor group can confuse the understanding of the relationship between humans and technologies

[1] It should be noted that as blockchain is an umbrella technology that can be implemented in various ways, the focus of this chapter is the original setup (i.e., the public, permissionless blockchain) that has the decentralization property and does not rely on any central authority to execute the protocol. This is also the only type of blockchain that can sometimes be considered trustless.

since such a relationship might be interpreted differently – e.g., with or without the need for trust – across diverse communities owning different levels of blockchain knowledge. Thus, clarifying the main trustor group of blockchain technology is the first task and contribution of this chapter. (2) The other reason for the divergent views on the role played by trust in blockchain-based interactions, as Jacobs (2020) argues, roots in different assumptions scholars make under the term 'trust.' For less demanding accounts, trust is understood as predictive expectations a rational actor holds toward the performance of a trustee (or the person who receives trust) (Gambetta 1988; Coleman 1990). In trust theories, these accounts are often labelled rational-choice accounts that regard trust as a result of weighing all risks and benefits of potential options in context, which is used not much different from the term 'reliance' (Simon 2013). Following these accounts, it seems that trust is applicable to not just humans but also things since they do not require the trustee to have any other condition other than reliability for developing a trust relation. This enables people to talk about trust in blockchain context as a shift of trust from traditional third parties to the algorithms, developers, miners, and markets (e.g., exchanges and online markets) in relation to blockchain technology.

In contrast, for more demanding accounts, trust is construed as a distinctive concept that differs from reliance in that the generation of trust also involves moral, normative, or affective beliefs about the trustee or certain aspect of the trustee (Baier 1986; O'Neill 2002; Hollis 1998; Holton 1994; Weckert 2005). For these accounts, to trust someone is to rely on them in a morally, normatively, or affectively loaded manner. Given such conditions, these accounts are often considered not applicable to entities having no mental state, which directly results in the understanding that blockchain-based systems are trust-free or trustless. However, other research has shown that, apart from physical persons, we also normatively expect from professionals (Jones 2004), institutions (Walker 2006), and technological systems (Nickel 2013). From this perspective, it seems that the depiction of blockchain as a trust-free technology ignores the rich array of expectations invited by the manifold normative values embedded in blockchain's basic infrastructure, and the negative attitudes (e.g., disappointment, anger, and a feeling of being betrayed) when one's trust is frustrated after the fact. Both aspects are seen as important cues for a normatively loaded trust relation going beyond mere reliance. Thus, the second task of this chapter is to explore a rich conception of blockchain trust that takes into account the trustor's normative expectations about the blockchains' performances. Philosophical discussion on this aspect can help people take a step out of merely focusing on the judgment of the systems' reliability and begin to think about questions of moral and normative significance and the relevant risks when talking about blockchain trust. The analysis provided here can further be utilized to steer the design and policymaking associated with blockchain implementations, with the aim of indicating directions for developing more trustworthy blockchains and reducing the risk and misplacement of trust.

With these considerations, this chapter provides a comprehensive analysis of the role and risk of trust in blockchain-enabled interactions, proposing a user-centred,

multilayer-structured framework for understanding blockchain trust in a normatively loaded manner. The rest of the chapter is organized as follows. Section 4.2 defines normal users as the main trustor group of blockchain systems, clarifying the reasons why trust is needed in user-blockchain interactions. Section 4.3 structures blockchain trust by integrating the current ways of characterizing users' reliance on blockchains into a blockchain engineering framework, providing a holistic view of how trust is established in accordance with the pivotal elements underlying blockchain-based platforms. Section 4.4 conceptualizes blockchain trust in line with the distinctive feature of the trust phenomenon, which not only clarifies the normatively loaded nature of blockchain trust but also offers a perspective from which the appropriateness one grants to the normative values built into blockchain applications can be scrutinized. Following this idea, section 4.5 examines two of the most promising values put forward by developers of the original type of blockchain that have the potential to ground rich trust decisions. It argues that, contrary to the widespread beliefs, trust decisions built on these values are risky and unjustifiable due to the technical and ethical limits faced by its current implementations.

4.2. Why blockchain trust is needed, for whom?

Much of the research into trust would agree that trust is risky (Luhmann 1979; Baier 1986; Becker 1996; McLeod 2015). By trusting, trustors have to take the risk of being let down and they may lose whatever entrusted to trustees. Yet, why do not people stay away from this vulnerable position? The reason may lie in the basic fact that every social being has limited cognitive and practical power so that nobody is capable of doing everything by oneself (Jones 2012). In everyday life, not only do we need to rely on others to satisfy fundamental human needs (e.g., food, water, and shelter), but we also need to rely on others in the acquisition of basic facts, scientific knowledge, and practical techniques (Hardwig 1991; Simon 2010). Trust provides a way of coping with our essential finitude by relying on others to help, learn, and cooperate, bringing both pragmatic and epistemic value to people in need.

In relationships underpinned by trust, while trustors are optimistic about trustees' commitment and competence in doing certain things, they understand that their trust might be frustrated and they are willing to give discretionary power to the trustee (Baier 1986; Jones 2004). Implicit in this statement is the complex relationship between trust and knowledge. Although we use knowledge to place and withdraw trust (Simon 2010), "it is an important fact about trust that it cannot be given except by those who have only limited knowledge, and usually even less control, over those to whom it is given" (Baier 1992). From this perspective, the need for trust might be considered as a sufficient condition for knowing that one lacks knowledge and control about the trustee. In other words, the lack of knowledge and control could be viewed as a threshold for creating the need for trust. By contrast, if someone were fully aware of or able to control another's action, the discretion and uncertainty involved in the relationship would cease to exist, so would the need for trust.

This entangled relationship between trust and knowledge is particularly distinct when comparing laypeople's perception of controversial technologies with that of scientists. It is evident that public perceptions towards controversial technologies tend to follow the "cognitive miser model" in which value predispositions (e.g., trust) and other heuristic reasons play key roles rather than scientific knowledge (Fiske and Taylor 2013; Kahneman 2011). The situation is the same in cyberspace where consumers have increasingly relied on heuristics for trust instead of being more rational (Pesch and Ishmaev 2019). On the one hand, this is because most people are incapable of rationally assessing the relevant information due to a limit of time, resources, and technical expertise (Nickel 2013). On the other hand, trust provides an easier approach – a cognitive shortcut – for people to make decisions on whether to take the risk of doing something, which, as argued by Luhmann (1979, 8), functions as an effective form of reducing the complexity of living in society. By contrast, for scientists and experts who possess an in-depth understanding of the related technology, attitudes toward the technology are typically formed on the basis of domain-specific and general knowledge rather than trust (Ho et al. 2018). As a result, it is likely that resources endorsed by scientist communities can engender only limited epistemic confidence of laypeople, leading laypeople's judgment to false negatives; or that resources encouraging laypeople's trust cannot reach the same effect in scientist communities, leading laypeople's judgment to false positives. Both situations ask for efficient communication between the two groups. Public communication of scientific knowledge is needed on one side. Predictive resources used by "cognitive misers" should be critically examined and carefully incorporated into the design and communication process on the other.

In this regard, whether there is a need for technology trust depends not only on the practical interest of using a particular technology but also on the extent to which the trustor knows about the technology. When the trustor owns limited knowledge of and control over the trustee's action, either explicitly or implicitly, trust is usually a need for technology adoption, and the acquisition of knowledge can enhance the epistemic reason for trust.

In the case of the Bitcoin blockchain, studies have shown that blockchain knowledge plays an epistemic role in enhancing users' trust, and the lack of knowledge and understandability hinders users' initial trust formation in blockchains (Sas and Khairuddin 2017; Ostern 2018). However, while the word 'user' is adopted almost everywhere with the meaning of someone who uses the Bitcoin network for different purposes, according to the preceding discussion on the relationship between trust and knowledge, it causes confusion regarding whether a certain user has the need for trust. The existence of different classes of users with different levels of knowledge fundamentally explains why blockchain technology is sometimes considered trust-free while sometimes viewed as a trusted technology. On the one hand, the trustlessness utterance is commonly seen in scientific research into blockchain technology as the authors often make unrecognized and unspoken assumptions that the audience can fully understand how the system flows and how to control its functioning by developing and maintaining the codebase. In this case, trust is not a

necessity as little uncertainty and discretion from the system's side are considered. On the other hand, media reports and academic research from other communities often describe a blockchain as a trusted or trustable system potentially in place of third parties since a complete understanding of every detail is hardly possible for people with no technical background. In this case, the performance of the system is considered not-fully-understandable, whether the black box comes from the complex algorithms, the unknown developers and other network participants, or the novel applications. Therefore, even in the case of the original blockchain where trust is expected to be omitted by the developer(s), it still plays an important role in shaping the opinions of people with limited technical expertise.

Thus, an explicit definition regarding who the users requiring trust for blockchain adoption are is needed. Following the above analysis, it is arguable that *normal users* who are actively or passively associated with the network yet have limited, rudimentary, and fragmentary blockchain expertise can be defined as the main trustor group of the original blockchain. For example, active users can be those nodes who contribute their computation power to the execution of the system for economic purposes (i.e., miners) yet do not possess a thorough blockchain knowledge; passive users can be those who merely use the system as an instrument for transactions and investments. By restricting trustors of blockchain-based systems to these specific groups, such a definition partially addresses the conceptual confusion over the understanding of the relationship between trust and blockchain. Equally important, it highlights the inherent risks of trusting complicated systems, which concern not only one's vulnerability with respect to the systems' discretion but also the epistemic impairment position of assessing the actual trustworthiness of the systems (Ishmaev 2018; Nickel 2013). The vulnerable position of the trustor suggests serious moral concern over trust manipulation and mistrust associated with blockchain-based interactions, especially considering blockchains' irreversibility nature that leaves almost no room for redeeming the loss. I will return to the discussion about the risk involved in trusting blockchains later. Before that, we shall take a look at what elements of blockchain-based systems are potential targets of users' trust.

4.3. A framework for understanding the structure of blockchain trust

Based on the above discussion, it can be said that from a user-centered perspective trust is still needed in blockchain-based interactions. To understand users' blockchain trust systematically, this section explores the different elements that potentially invite trust in the original blockchain. These trust-inviting elements are then integrated into the blockchain engineering framework (BEF) outlined by Notheisen, Hawlitschek, and Weinhardt (2017). The resulting framework, named user-centred blockchain trust framework (BTF), explicates how the potential targets of trust are associated with the pivotal elements underlying blockchain-based platforms in multiple layers. In doing so, it provides a holistic view capturing the structure of users'

reliance relations on blockchain applications.

As mentioned, existing research has argued that the original blockchain does not eradicate trust. Instead, it enables a shift of trust from third-party authorities (e.g., banks and governments) to the system's algorithms, the network's stakeholders, and the core values built into the system.

Consider first the trust shifted to the algorithms. An algorithm is a set of rules that give a sequence of operations for solving a specific type of problem. With features of "finiteness, definiteness, input, output, and effectiveness", algorithms generally provide some predictability that allows users to predict the system's outcome (Knuth 1997). As a result, increasing epistemic authority is placed in algorithms to assess and predict the trustworthiness of diverse information sources. The idea of embedding epistemic authority in non-human agents such as algorithms has been argued as a new form of trust that requires us to remain particularly vigilant about the algorithms' transparency (Simon (2010). It seems that blockchain technology has fostered such transparency to a great extent. In the Bitcoin network, while no node is delegated with a privileged position to control the database, participants validate transactions and maintain the database collectively by using consensus algorithms and rules, which ultimately result in a single-valid, tamper-proof, and publicly accessible database that can identify any double-spend attempt without the need for any third party (Nakamoto 2008). Thus, interactions processed by the network are based on algorithmic trust that allows users to predict the system's future behaviour and act accordingly rather than trust between human agents (Swan and De Filippi 2017). In this regard, the underlying algorithms are elements that directly invite users' trust towards the system. Trust in Bitcoin's algorithms, as Lustig and Nardi (2015) state, can be viewed as the trust placed in the legitimate power of the open-source codebase to verify information and direct human action, which is considered more predictable than opaque, large institutions' actions.

Consider second the trust shifted to the network contributors. Although what users directly rely upon is the correct functioning of the blockchain system, the performances of the system are enabled by a chain of network contributors, mainly including developers and miners. At the protocol layer, when users adopt Bitcoin, they put faith in the developer community and regard them as collective trustees who are responsible for maintaining the codebase (Mallard, Méadel, and Musiani 2014). As a result of lacking knowledge and time, normal users have to depend on coders who are able and willing to take the responsibility to write and verify the blockchain code. Considering Bitcoin's open-source nature, while this coder community can be as large as whoever contributes to patch proposals, peer-review, and testing, the trustworthiness of "core developers" (known as maintainers) is of importance since they exert the decision-making power over judging the appropriateness of all pull requests. Trust developed at this layer can be personal or impersonal, depending on whether the trustor places trust in a particular, known developer or the developer community as a professional group. At the application layer, when users adopt Bitcoin, trust is distributed to a network of miners that contribute to validating and

securing transactions collectively (Kasireddy 2018; Werbach 2018).

Consider third the trust shifted to the core values potentially brought about by the system's performances. As embodiments of certain economic, moral, and political assumptions over others, the full connotations of algorithms indicate something beyond the mathematical and symbolic properties (Uspensky and Semenov 1981; Hill 2016). As a distributed database technology initially designed for decentralized and continuous interactions between heterogeneous participants (Glaser 2017), blockchain technology came to prominence with explicit core values – such as decentralization, transparency, verifiability, and accountability – that may contribute to addressing the apprehension about information aggregation and power centralization caused by hierarchical structures. In this regard, it is arguable that one important kind of users' motivations for using blockchain-based systems is that they share similar attitudes with the set of consequences underpinned by the systems' affordance. For example, as argued by De Filippi and Loveluck (2016), the adoption of cryptography tools enables the Bitcoin project to be advocated by cypherpunk groups and libertarians as a means of resistance to traditional authorities and human rights abuses. In trusting the system, people expect that these values could be brought about by the system's performances enabled together by blockchain algorithms, a network of miners, the developer community, and other relevant entities that might impact the actual usage of the system.

Before entering into the discussion on the conception of blockchain trust developed on the basis of these elements, the BEF is introduced as a tool to systematically structure how these trust-inviting elements are associated with the pivotal elements underlying blockchain-based platforms. Fig. 4.1 illustrates the extended framework (i.e., the BTF) resulting from a combination of the BEF and the content newly integrated by this chapter (i.e., the content in the dotted square).

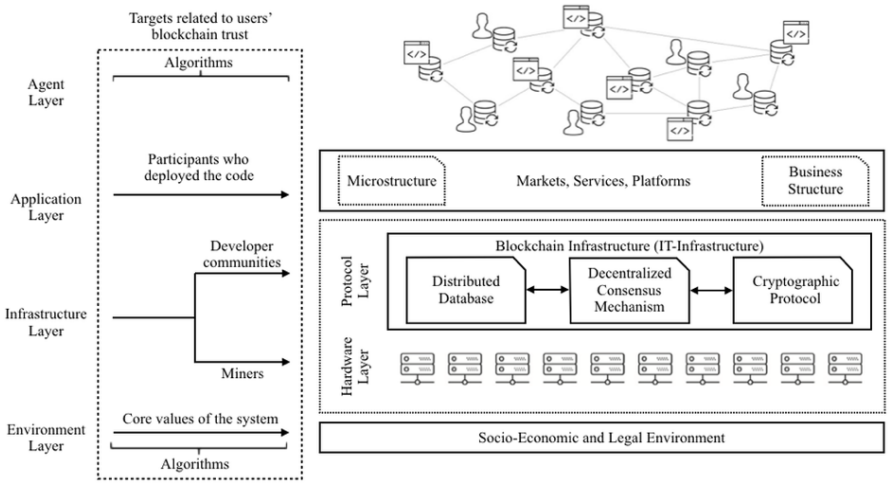

Figure 4.1: User-centred BTF, based on the BEF proposed by Notheisen et al., 2017

The BEF creates a common approach for structuring the layers and pivotal elements of blockchain-based platforms in general. It consists of four layers. (1) The environment layer constructs economic, social, and legal foundations of the actions in other layers, which could be corresponded to the trust shifted to the core values built into blockchains' performances. (2) The infrastructure layer comprises the protocol layer that defines the basic elements of blockchain infrastructure and the hardwire layer that connects a heterogeneous crowd of devices running the virtual machine. On the one side, this layer introduces trust into the developers' collective ability to write and verify the code. On the other side, in the case where mining is necessarily involved, users need to make trust judgment about whether miners will use their computation power to maintain but not to manipulate the integrity of the network. (3) By integrating a full-fledged programming language known as the smart contract functionality, microeconomic designs such as autonomous market mechanisms and services can be built in the application layer. As the code in this layer is controlled by participants who deployed the code (Glaser 2017), the realized blockchain-based services also introduce trust into the application designers. (4) Govern by the applications' rules and characteristics, human and artificial agents can interact in the agent layer and process services available within the self-sufficient, closed ecosystem. By contrast, if the system is not self-sufficient and needs to be bound to external services and interfaces to realize certain functions, it pushes the trust issue back to those third-party authorities, such as exchanges and online markets. In addition, algorithmic trust pervades throughout the design and execution of the ecosystem, providing epistemic authority for users to reasonably expect the system's output and act accordingly.

The user-centred BTF proposed shows that trust shifted from traditional third-party authorities to blockchain-based platforms is multilayer-structured. Intimately linked with different elements constituting blockchain systems, the targets that potentially encourage users' trust include the blockchain algorithms, the relevant stakeholders (i.e., protocol developers, miners, and application designers), and the core values potentially brought about by the system's performances. The resulting framework contributes to first systematically structuring the elements of blockchain-based systems that potentially invite users' trust, providing a way for developers and users to reflect on the actual trustworthiness of these elements when interacting with a blockchain system. Moreover, the trust-inviting elements elucidated in the framework provide the sources on which a distinctive conception of blockchain trust can be built, as I will discuss in the next section. For less demanding accounts that use trust as reliance, trust can probably reside in any of the trust-inviting elements identified above. In this way, trusting a blockchain-based system simply means that people expect that the system will perform reliably for achieving specific goals. Such an expectation can be considered as a result of any or a combination of their judgment on the reliability of the trust-inviting elements of that system, with no additional requirements for building the trust relation. However, if we delve into the reliance relationship grounded in the third category of the trust-inviting elements, a more demanding account of trust seems to be applicable to the systems, which

enables us to take a step out of rational-choice accounts and begin to think about the moral and normative issues involved in blockchain trust.

4.4. Understanding the distinctive feature of trust in the context of blockchain technology

As mentioned, more demanding accounts of trust hold that trust is a distinctive concept that contains some morally, normatively, or affectively loaded elements, going beyond mere expectations about the trustee's reliability. For these accounts, such elements essentially explain why one would like to take a "leap of faith" and form interactions under uncertainty. For these accounts, thus, a blockchain is a viable target of trust if and only if one's attitudes toward the trust-inviting elements of blockchain-based systems can bear a family resemblance to the distinctive feature of trust. Such attitudes show the unique explanatory power of trust that allows one to take the risk of trust. In an effort to explore this question, this section conceptualizes blockchain trust in line with a philosophical account of trust, with the aim of providing a theoretical foundation on which the core values built into blockchains can be reflected.

To understand whether a rich conception of trust can be applied to blockchain systems, it is imperative to first take a look at the nature of the trust notion and trust in technologies as an extension of trust. In philosophical studies into the trust concept, while human-to-human trust is regarded as the original and dominant paradigm of trust relations, the possibility of trust arising between humans and technologies is often overlooked or considered implausible. This situation can be primarily attributed to the widely shared assumption that a trustee should act on the basis of a relatively complete mental state, which can show the motive of the trustee that is considered morally decent. The most influential foundation of this assumption is Baier's (1986) classic explanation about moral trust relationships. For this account, to trust is to rely on another's goodwill and competence to "pursue, promote, preserve, and protect" one's certain goods and vulnerabilities (Alfano 2016). The goodwill of the trustee here shows that one cares about, or at least will not use the discretionary power to harm, the things that are valued by the trustor.

Grew out from this statement, much of the literature challenges the viability of applying the trust notion to technologies by arguing either that technologies can only be paired with strategic reliance, or that this attempt is merely an extension of interpersonal trust (Nickel 2013). Regarding the former aspect, Pettit (2004), for example, argues that machines and technological systems cannot be objects of trust since they lack consciousness and agency manifesting goodwill towards the trustor. Regarding the latter aspect, Pitt (2010) and Cook (2010), for example, argue that trust in a specific technology is eventually an issue of trusting a certain person to do certain things. However, these two challenges are fraught with the problem that the goodwill-based account is not omnipotent. Although accounts built on the trustee's specific mental state might be rich in justifying certain forms of trust (e.g.,

trust in friends), O'Neill (2002) has pointed out that the trustee's goodwill is neither sufficient nor necessary for understanding a wider variety of trust forms – such as trust in professionals and strangers. Affective accounts that highlight emotions of the trustor (e.g., children-to-parents trust) appear to face similar problems of the goodwill-based account, i.e., it is almost impossible to extend these psychological states to explain more diffuse and complex trust relations.

Differently, the normative accounts argue that the distinctive feature of trust lies in one's normative expectations of another's responsible behaviour, which are grounded in our moral standards presumably shared with others, such as promises should be honoured and duties should be performed (Hollis 1998; Walker 2006 ; Jones 2004). As Simpson (2012) puts forward, this sort of account is more suitable for characterizing more distant relations in a communal sense, such as trust in obligation people (e.g., firefighters and doctors). For the normative accounts, trust is more like a stance that the trustor holds toward the trustee, expecting that the trustee will do what they should while leaving their motivation open to different contexts – be it the desire of good repute, goodwill, a good character, the fear of sanctions, the pressure of social constraints, etc. (Walker 2006). For these accounts, it is the normatively loaded expectations that distinguish the trustor's trust from reliance and the trustee's trustworthiness from reliability.

As summarized by Simon (2013), the way of depicting trust is strongly dependent on the particular trust relation in question, since different trustees and situations to which the trust concept applies vigorously shape the actual meaning of trust. In this regard, rather than seeking a single, fixed definition that is amenable to all counterexamples, it seems more plausible to focus on where the value of trust comes from in context (Simpson 2012; van den Hoven 2008). Acknowledging that different forms of trust might emphasize distinct facets of the trust notion, researchers have explored various frameworks in support of the arguments that trust can be invited by and placed in not only physical persons but also roles and professionals (Becker 1996; Pellegrino 1991), institutions (Townley and Garfield 2013), and technologies (Nickel 2013; Coeckelbergh 2012; Nguyen 2019). Although built on different considerations, these accounts shed light on exploring the explanatory power of trust in the context of abstractly-characterized entities.

A particularly interesting and robust trend is the normative accounts mentioned, which provides a relatively consistent way to understand trust when applied to multifarious trustees. For example, Walker (2006) argues that, apart from obligation people, people also hold normative expectations toward institutions and businesses, which are more like a default stance we habitually stand with respect to the good state of their services. When applied to technological systems, such expectations make the trustor believe that they are entitled to what the systems are supposed to do (Nickel 2013). Nickel (2020) further argues that here the systems are the direct targets of trust while the trust developed with engineers and designers behind the systems are considered indirect, impersonal, and abstract. In this sense, similar to roles and professionals, it seems that institutions and technological systems are

also embodiments of certain moral and normative standards that can invite people's shared beliefs about their performances. From this perspective, it seems natural to say that trust can be placed in these abstractly-characterized entities even though we are not aware of those strangers who fill in the roles of a doctor, a banker, or a software engineer. What we generally rely upon is instead the standard performances that we normatively expect of that profession, institution, or technological system rather than any unique tie or personal concern a particular human agent may give to us. This broader conception explains the generation of trust without any complete agential state of the trustee being assumed.

In this regard, it might be said that the essential distinction between the normative expectation we hold toward a specific person and that we hold toward an abstractly-characterized entity lies in the different sources that ground such expectations. While the normativity of the former comes mainly from the moral understandings we presumably shared with others (Walker 2006), that of the latter is suggested by a wide range of normative structures built into these entities' performances – such as different values and norms, laws and regulations, and codes of conduct. This claim resonates with the broader philosophical view that technologies have moral and normative significance and can inform ethical decisions and practices (Verbeek 2011). Coupled together, these views make it possible for technological systems to be designed in a way that encourage users to form their expectations of the systems. For example, if a technological system claims to be privacy-preserving or a product of GDPR-compliance, it is reasonable for one to normatively expect that the system's patterns of action will display explicit cues about privacy protection. Relying on such a system thus carries significant ethical implications about the moral acceptability and social desirability of the system.[2]

The above analysis of trust in technologies provides a feasible way to reach a rich conception of blockchain trust, highlighting the importance of the core values embedded in and potentially embodied by blockchains' performances. As such, the attitude of relying on these values is not just used in a *predictive* sense that one believes that such values will be brought about by the systems' performances, but also in an *evaluative* sense that one thinks that these values are the desirable things that should be folded into the systems. In a word, combined with the trust shifted to the algorithms and network contributors, if a person X trusts a blockchain system Y in a normatively loaded way, not only does X rely on Y's functionality, but also X considers certain moral, political, or social values built into Y's performances to be appropriate.

Betrayal, in this case, is not about how our trust has been frustrated by others' commitments, but about how we feel alienated toward the *appropriateness* we grant to the normative consequences potentially brought about by the system's performances. Just like promises are not always honoured, trust-inviting cues are not always reliable, and they do not lead to the fact that the related entity is indeed

[2]GDPR is short for General Data Protection Regulation.

trustworthy. Recalling the trustor's epistemic vulnerability discussed earlier, when lacking sufficient knowledge, time, and resources to understand a complex system, users might hastily and carelessly bear an unquestioning attitude towards the system and thus trust more than the system's trustworthiness warrants (Nguyen 2020). Thus, we need to assess such appropriateness, in order to approach more warranted trust decisions and avoid the risk of misplacing trust. By presenting the challenges faced by realizing two core values of the original blockchain, in what follows, the focus of this chapter is shifted from the blockchain's trust-inviting elements to the creation of trust-deserving or trustworthy applications.

4.5. Examining the core values related to blockchain technology

As a first step in assessing the appropriateness of what people normatively expect from the blockchain's performances, two questions seem to be germane: (1) whether the embedded values are realized in the systems' socio-technical context, and (2) whether they are realized without conflicting with each other. In this section, these two questions are discussed via an examination of two of the most promising values put forward by the original blockchain, namely, decentralization and transparency. These values are inherent in this technology's technical infrastructure and they are the seeming sources of the blockchain's moral desirability from which justified blockchain trust decisions might be achieved. Nevertheless, this section shows that there is a tension between the pressing values that are intended to be achieved by developers and the predicament situations caused by current implementations. It argues that, unlike the widespread beliefs, trust decisions built on these promised values are risky and unjustifiable due to the ethical limits and practical difficulties involved.

4.5.1. Decentralized network vs. power centralization

While a top-down, centralized authority and hierarchical structures provide useful means of facilitating valid social interactions in modern societies, they are often fraught with a crisis of trust due to data aggregation, undue censorship, surveillance, and the consequent moral apprehension such as the erosion of individual freedom, privacy, and autonomy (Chaum 1985; Al-Saqaf and Seidler 2017). By contrast, data validity of the original blockchain is fuelled by a series of consensus rules and cryptographic tools and executed on a large network of computing devices, peers, and developer communities. By replacing the role and functions of third-party authorities with technical settings, the decentralized database technology proposed by Nakomoto has the potential to mitigate the moral issues engendered by traditional authorities, showing an explicit normative message that the decentralized network is considered more desirable than the traditional mechanism. Such a message can invite the trust of people who favour the moral desirability and other effects potentially brought about by the system's decentralization promise. Based on the rich conception of blockchain trust discussed above, to see whether trust relations grounded in this striking value are well-grounded, we need to take

a look at the real-world performances of blockchain applications for realizing this promise.

As Reijers et al.(2018) point out, the governance of blockchain-based systems can be divided into two categories: on-chain governance where interactions are solely determined by the rule of code and off-chain governance where the reference community might be affected by self-regulation and exogenous rules such as laws and regulations. Yet, it has been argued that there is an inherent degree of centralization existing in both the enforcement of rules and the collective governance of blockchains (Azouvi, Maller, and Meiklejohn 2018). Firstly, at the application layer, it is uncertain to what extent the network's decentralization promise can be realized in the fact that several mining conglomerates control a considerable amount of computing power. The monopoly of mining makes it possible for the conglomerates to collude with each other and manipulate the system. A decentralized network, in this sense, does not guarantee decentralized power (Brekke 2019).

Secondly, the governance structure of the protocol layer is also quite centralized. Research has shown that the same developer has created around 7% of all Bitcoin documents, and half of all the comments in its GitHub repository were written by only 8 contributors (Azouvi, Maller, and Meiklejohn 2018). While the codebase of the project is maintained by only a few developers, vital decisions within the community are reached by the exchange of opinions among members on mailing lists without any transparent decision-making process being known by the multitude of users (Gervais et al. 2014). Such a situation causes concern over the appropriateness that users grant to the blockchain's decentralized setup as it simply uses a few developers to replace the complex social roles previously filled by a wide range of people and institutions. Compared to the well-established special legislation on traditional third parties (e.g., corporations and banks), external rules that can be imposed on individual developers are scarce and limited to general laws (e.g., anti-money laundering). Such a situation is not compatible with the significant role played by developers in determining the functionality of blockchain-based systems that are considered as potential alternatives to traditional third parties.

The increasing power centralization of the two layers, together with some centralized services surrounding the Bitcoin system (e.g., web wallets and exchange platforms), make it unclear whether the system's decentralized infrastructure will lead to power decentralization. This also makes the question of who controls the system of crucial ethical and political importance (Reijers and Coeckelbergh 2018). In the lack of proper rules and regulations, a market- and technocrat-driven governance structure of the system not only indicates the risk of an undemocratic decision-making process, but also makes the wide range of social effects of the system subject to a small group of people (De Filippi and Loveluck 2016). These all make the appropriateness of blockchain's decentralization promise questionable. Therefore, as a joint result of how Bitcoin is implemented today by all actors and processes that are part of the system's socio-technical context, trust decisions invited by one of the system's core values known as decentralization seem risky and

unjustifiable. Explicit regulatory frameworks that can be applied to the developer community as well as the peer-to-peer network are urgently required if we want to make decentralization a trust-deserving property rather than a trust-inviting fiction of blockchain applications.

4.5.2. Data transparency vs. privacy concern

Transparency is arguably another core value that often invites users to trust blockchains in a normatively loaded way. The normative implications of transparency can be understood from the crucial role of information transparency in promoting people's trust in traditional institutions. As characteristics of information, transparency dimensions, including information disclosure, clarity, and accuracy, are positively related to an institution's trustworthiness that can encourage trust (Schnackenberg, Andrew, and Tomlinson 2016). But given the privileged position of centralized authorities played in data processing (e.g., the recording, collection, storage, and using of data), users are almost always placed in a passive and vulnerable position caused by information asymmetry and its knock-on effects. In this regard, an open-source, public blockchain appears to be an ideal medium to facilitate transparency by permitting users fair access right to the database and source code. Rather than relying on the goodwill and a sense of responsibility of centralized authorities and relevant individuals, transparency in blockchain-based systems is guaranteed by network protocols that directly mitigate the situation of information asymmetry.

One flipside of such blockchains' transparent and immutable nature is the challenge posed to private data protection (De Filippi 2016). As all Bitcoin transactions are publicly available and traceable yet not fully anonymous, they can be linked to other information or datasets to reveal the identity of coin owners. Given the risk of re-identification and privacy loss, it can be argued that trust judgment on the public accessibility of a blockchain should be evaluated together with the system's capacity for coping with privacy-related issues. Despite the benefits enabled by blockchains' peculiarities, it is thus important for users to understand the privacy issue involved, particularly considering the fact that in the context of blockchain systems no one is legally responsible for users' privacy loss.

At the same time, blockchain applications are trying to solve this dilemma in different ways. Take the case of the Enigma project that is seen as one of the most promising solutions for preserving privacy in the blockchain context. The way that Enigma addresses the privacy concern is to use a cryptography tool called secure Multi-Party Computation that allows data to be split, encrypted, and computed by nodes at a second layer off the ledger (Zyskind, Nathan, and Pentland 2015). This means that nodes of the network could verify smart contract computations without seeing any decrypted data. Although some metadata is still required to be stored in the ledger to keep track of data ownership and the distribution of data, this solution is much more privacy-friendly than the original blockchain. Furthermore, with the establishment of blockchain-based data markets, projects like Enigma purport to not merely protect privacy but also unlock new value by allowing data owners to

share, trade, and get rewards from their private data. Nevertheless, the moral limits of this solution should not be ignored. As Ishmaev (2019) argues, a market-centric solution for private data can exacerbate the situation of data secondary usage, as well as reduces the multifaced nature of privacy as a moral right to a property that can be measured – if at all – by money. The moral reasons for personal data protection – such as the prevention of informational-based harm, inequality, injustice, and encroachment on moral autonomy – are simply corroded or crowded out in the propertization processes of private data (van den Hoven 2008).

To conclude, based on the argument that the normative values inserted into blockchains' infrastructure are essential building blocks of users' blockchain trust, this section has examined two core values that possibly invite justified trust decisions, in order to scrutinize the appropriateness granted to these trust-inviting cues. In sum, it is argued that the promise of decentralization is restricted by how Bitcoin is implemented today in its socio-technical context, and the promise of transparency should be evaluated together with the application's capacity for preserving privacy and the moral limits involved. Both aspects imply that trust decisions invited by these promised values should be carefully reflected before bestowing appropriateness.

4.6. Conclusions

Whom and what can and should we trust? This is a fundamental philosophical question, as it forms the background of nearly all social cooperation – from dyadic interactions in situations well-modelled as prisoner's dilemmas and stag hunts to large-scale, longitudinal interactions between anonymous groups. Nevertheless, arriving at a well-grounded trust decision is a non-trivial task, especially when it comes to complex and novel systems such as blockchains. On the trustor's side, users' epistemic vulnerability impedes them from collecting and extracting accurate information from a vast amount of resources online, creating barriers to capture a relatively complete picture of the situation. On the trustee's side, while blockchain infrastructure has the potential to revolutionize the way we interact, the entire blockchain industry is still in its infancy. A number of internal and external uncertainties regarding the moral concerns, legal constraints, and technical limitations of blockchain implementations add unforeseen dynamics to our trust decisions made at the moment (Swan 2015).

To explicate the role and risk of trust related to blockchain-based interactions, this chapter has critically engaged with the concept referred to as blockchain trust. It provides a philosophical analysis of the trust notion in the context of blockchain technology, encompassing four aspects: (1) a clarification of the trustor group of blockchain technology; (2) a systematic analysis of the elements potentially inviting users' blockchain trust; (3) an investigation into how the distinctive feature of the trust notion can be understood in blockchain context; and (4) a reflection on the appropriateness one may give to the core values built into blockchains' potential. The reflection provided by this chapter only starts the inquiry about justified

blockchain trust. Future research could build on the conceptual analysis provided and systematically explore regulatory solutions to approach more warranted trust in the context of blockchain technology.

References

Al-Saqaf, Walid, and Nicolas Seidler. 2017. "Blockchain Technology for Social Impact: Opportunities and Challenges Ahead." Journal of Cyber Policy 2.3: 338-354.

Alfano, Mark. 2016. "The Topology of Communities of Trust." Russian Sociological Review 15(4): 30-56.

Azouvi, Sarah, Mary Maller, and Sarah Meiklejohn. 2018. "Egalitarian Society or Benevolent Dictatorship: The State of Cryptocurrency Governance." In International Conference on Financial Cryptography and Data Security, volume: 10958. Springer, Berlin, Heidelberg.

Baier, Annette C. 1986. "Trust and Antitrust." Ethics 96.2: 231-260.

Baier, Annette C. 1992. "Trusting People." Philosophical Perspectives 6: 137-153.

Becker, Lawrence C. 1996. "Trust as Noncognitive Security about Motives." Ethics 107.1: 43-61.

van Lier, Ben. 2017. "Can Cyber-Physical Systems Reliably Collaborate within a Blockchain?" Metaphilosophy 48.5: 698-711.

Brekke, Jaya K. 2019. Disassembling the Trust Machine, Three Cuts on the Political Matter of Blockchain. Accessed 22 February 2020. http://etheses.dur.ac.uk/13174/.

Chaum, David. 1985. "Security Without Identification: Transaction Systems to Make Big Brother Obsolete." Communications of the ACM 28.10: 1030-1044.

Coeckelbergh, Mark. 2012. "Can We Trust Robots?" Ethics and Information Technology 14.1: 53-60.

Coleman, James S. 1990. Foundations of Social Theory. Cambridge: Belknap Press of Harvard University.

Cook, SD Noam. 2010. "Making the Technological Trustworthy." Knowledge, Technology & Policy 23.3-4: 455-459.

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

De Filippi, Primavera. 2016. "The Interplay Between Decentralization and Privacy: The Case of Blockchain Technologies." Journal of Peer Production, Issue 7.

Fiske, Susan T., and Shelley E. Taylor. 2013. Social Cognition: From Brains to Culture. Sage.

Gambetta, Diego. 1988. "Can We Trust Trust?" In Trust: Making and Breaking Cooperative Relations, edited by D Gambetta, 213-237. Oxford: Basil Blackwell.

Gervais, Arthur, Ghassan O. Karame, Vedran Capkun, and Srdjan Capkun. 2014. "Is Bitcoin a Decentralized Currency?" IEEE Security & Privacy 12.3: 54-60.

Glaser, Florian. 2017. "Pervasive Decentralisation of Digital Infrastructures: A Framework for Blockchain Enabled System and Use Case Analysis." In Proceedings of the 50th Hawaii International Conference on System Sciences, 1543-1552. Hawaii, United States.

Hardwig, John. 1991. "The Role of Trust in Knowledge." The Journal of Philosophy 88.12: 693-708.

Hill, Robin K. 2016. "What an Algorithm Is." Philosophy & Technology 29.1: 35-59.

Ho, Shirley S., Alisius D. Leong, Jiemin Looi, Liang Chen, Natalie Pang, and Edson Tandoc Jr. 2019. "Science Literacy or Value Predisposition? A Meta-Analysis of Factors Predicting Public Perceptions of Benefits, Risks, and Acceptance of Nuclear Energy." Environmental Communication 13.4: 457-471.

Hollis, Martin. 1998. Trust Within Reason. Cambridge: Cambridge University Press.

Holton, Richard. 1994. "Deciding to Trust, Coming to Believe." Australasian Journal of Philosophy 72.1: 63-76.

Ishmaev, Georgy. 2018 "Rethinking Trust in the Internet of Things." In Data Protection and Privacy: The Internet of Bodies, edited by De Hert P, Gutwirth S, van Brakel R, and Leenes R, 203-230. Oxford: Hart Publishing.

Ishmaev, Georgy. 2019. "The Ethical Limits of Blockchain-Enabled Markets for Private IoT Data." Philosophy and Technology 1-22.

Jacobs, Mattis. 2020. "How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology." Philosophy & Technology: 1-15.

Jones, Karen. 2004. "Trust and Terror." In Moral Psychology: Feminist Ethics and Social Theory, edited by DesAutels P and Walker MU, 3-18. Maryland: Rowman and Littlefield.

Jones, Karen. 2012. "Trustworthiness." Ethics 123.1: 61-85.

Kahneman, Daniel. 2011. Thinking, Fast and Slow. Toronto: Doubleday Canada.

Kasireddy, Preethi. 2018. "What Do We Mean By "Blockchains Are Trustless?". Accessed 10 October 2019. https://medium.com/@preethikasireddy/eli5-what-do-we-mean-by-blockchains-are-trustless-aa420635d5f6.

Knuth, Donald Ervin. 1997. The Art of Computer Programming. Vol. 1. Massachusetts: Addison-Wesley.

Luhmann, Niklas. 1979. Trust and Power. Chichester: John Wiley.

Lustig, Caitlin, and Bonnie Nardi. 2015. "Algorithmic Authority: The Case of Bitcoin." In 48th Hawaii International Conference on System Sciences, 743-752. Hawaii, United States.

Mallard, Alexandre, Cécile Méadel, and Francesca Musiani. 2014. "The Paradoxes of Distributed Trust: Peer-To-Peer Architecture and User Confidence in Bit-

coin." Journal of Peer Production 1-10.

Mcleod Carolyn. 2015. "Trust." Accessed 1 November 2018. http://Plato.Stanford.Edu/Archives/Fall2015/ Entries/Trust/.

Nakamoto, Satoshi 2008. "Bitcoin: A Peer-To-Peer Electronic Cash System. Bitcoin." Accessed 1 July 2016. https://Bitcoin.Org/Bitcoin.Pdf.

Nguyen, C.Thi. 2020. "Trust as an Unquestioning Attitude." In Oxford Studies in Epistemology.

Nickel, Philip J. 2013. "Trust in Technological Systems." In Norms in Technology, edited by De Vries MJ, Hansson SO, Meijers AWM, 223-237. Dordrecht: Springer.

Nickel, Philip J. 2020. "Trust in Engineering." In Routledge Companion to Philosophy of Engineering, edited by Michelfelder DP and Doorn N.

Notheisen, Benedikt, Florian Hawlitschek, and Christof Weinhardt. 2017. "Breaking Down the Blockchain Hype–Towards A Blockchain Market Engineering Approach." In Proceedings of the 25th European Conference on Information Systems,1062-1080, Guimarães, Portugal.

O'Neill, Onora. 2002. Autonomy and Trust in Bioethics. Cambridge: Cambridge University Press.

Ostern, Nadine. 2018. "Do You Trust a Trust-Free Transaction? Toward a Trust Framework Model for Blockchain Technology." In Thirty Ninth International Conference on Information Systems, San Francisco, United States.

Pellegrino, Edmund. 1991. Trust and Distrust in Professional Ethics. In Ethics, Trust, and the Professions, edited by Pellegrino ED, Veatch RM and Langan JP, 69-85. Washington, D. C.: Georgetown University Press.

Pesch, Udo, and Georgy Ishmaev. 2019. "Fictions and Frictions: Promises, Transaction Costs and the Innovation of Network Technologies." Social Studies of Science 49.2: 264-277.

Pettit, Philip. 2004. "Trust, Reliance and the Internet." Analyse & Kritik 26.1: 108-121.

Pitt, Joseph C. 2010. "It's Not about Technology." Knowledge, Technology & Policy 23.3-4: 445-454.

Reijers, Wessel, and Mark Coeckelbergh. 2018. "The Blockchain As a Narrative Technology: Investigating the Social Ontology and Normative Configurations of Cryptocurrencies." Philosophy & Technology 31.1: 103-130.

Reijers, Wessel, Iris Wuisman, Morshed Mannan, Primavera De Filippi, Christopher Wray, Vienna Rae-Looi, Angela Cubillos Vélez, and Liav Orgad. 2018. "Now the Code Runs Itself: On-Chain and off-Chain Governance of Blockchain Technologies." Topoi, 1-11.

Sas, Corina, and Irni Eliana Khairuddin. 2017. "Design for Trust: An Exploration of the Challenges and Opportunities of Bitcoin Users." In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 6699-6510.

Schnackenberg, Andrew K., and Edward C. Tomlinson. 2016. "Organizational Transparency: A New Perspective on Managing Trust in Organization-Stakeholder Relationships." Journal of Management 42.7: 1784-1810.

Simon, Judith. 2010. "The Entanglement of Trust and Knowledge on the Web." Ethics and Information Technology 12(4): 343-355.

Simon, Judith. 2013. "Trust." In Oxford Bibliographies in Philosophy, edited by Pritchard D. New York: Oxford University Press.

Simpson, Thomas W. 2012. "What Is Trust?" Pacific Philosophical Quarterly 93.4: 550-569.

Simser, Jeffrey. 2015. "Bitcoin and Modern Alchemy: In Code We Trust." Journal of Financial Crime 22(2): 156-169.

Swan, Melanie, and Primavera De Filippi. 2017. "Toward a Philosophy of Blockchain: A Symposium: Introduction." Metaphilosophy 48.5: 603-619.

Swan, Melanie. 2015. Blockchain: Blueprint for a New Economy. California: O'Reilly Media.

Townley, Cynthia, and Jay L. Garfield. 2013. "Public Trust." In Trust: Analytic and Applied Perspectives, edited by Mäkelä P and Townley C, 95-108. Amsterdam: Rodopi Press.

Uspensky, Vladimir A., and Alexei L. Semenov. 1981. "What Are the Gains of the Theory of Algorithms." In Algorithms in Modern Mathematics and Computer Science, edited by Ershov AP and Knuth DE, 100-234. Springer, Berlin.

van den Hoven, Jeroen, Johan Pouwelse, Dirk Helbing, and Stefan Klauser. 2019. "The Blockchain Age: Awareness, Empowerment and Coordination." In Towards Digital Enlightenment, edited by Dirk Helbing, 163-166. Springer, Cham.

van den Hoven, Jeroen. 2008. "Information Technology, Privacy, and the Protection of Personal Data." Information Technology and Moral Philosophy, 301-322.

Velasco, Pablo R. 2017. "Computing Ledgers and the Political Ontology of the Blockchain." Metaphilosophy 48.5: 712-726.

Verbeek, P. P. (2011). Moralizing Technology: Understanding and Designing the Morality of Things. University of Chicago Press.

Walker, Margaret Urban. 2006. Moral Repair: Reconstructing Moral Relations After Wrongdoing. Cambridge University Press.

Weckert, John. 2005. "Trust in Cyberspace." The Impact of the Internet on Our Moral Lives, 95-120.

Werbach, Kevin. 2018. "Trust, But Verify: Why the Blockchain Needs the Law." Berkeley Tech. LJ 33: 487.

Zyskind, Guy, Oz Nathan, and Alex Pentland. 2015. "Enigma: Decentralized Computation Platform with Guaranteed Privacy." Accessed 18 March 2020. https://arxiv.org/abs/1506.03471.

5

Towards trustworthy blockchains: Reflections on blockchain-enabled virtual institutions

This chapter proposes a novel way to understand trust in blockchain technology by analogy with trust placed in institutions. In support of the analysis, a detailed investigation of institutional trust is provided, which is then used as the basis for understanding the nature and ethical limits of blockchain trust. Two interrelated arguments are presented. First, given blockchains' capacity for being institution-like entities by inviting expectations similar to those invited by traditional institutions, blockchain trust is argued to be best conceptualized as a specialized form of trust in institutions. Keeping only the core functionality and certain normative ideas of institutions, this technology broadens our understanding of trust by removing the need for third parties while retaining the value of trust for the trustor. Second, the chapter argues that blockchains' decentralized nature and the implications and effects of this decentralization on trust issues are double-edged. With the erasure of central points, the systems simultaneously crowd out the pivotal role played by traditional institutions and a cadre of representatives in meeting their assigned obligations and securing the functional systems' trustworthy performances. As such, blockchain is positioned as a technology containing both disruptive features that can be embedded with meaningful normative

This chapter is originally published as:

Teng, Y. (2021). Towards trustworthy blockchains: normative reflections on blockchain-enabled virtual institutions. Ethics and Information Technology, 1-13. https://doi.org/10.1007/s10676-021-09581-3

values and inherent ethical limits that pose a direct challenge to the actual trustworthiness of blockchain implementations. Such limits are proposed to be ameliorated by facilitating a shift of responsibility to the groups of people directly associated with the engendering of trust in the blockchain context.

5.1. Introduction

The question of trust is of essential importance to the prominence achieved by blockchain technology. In the whitepaper of Bitcoin, the pseudonymous creator, Satoshi Nakamoto, makes it clear that the primary purpose of creating a decentralized electronic payment system is to remove the need for trusting third-party institutions (e.g., banks) that are often considered necessary for facilitating online transactions between heterogeneous groups of participants (Nakamoto 2008). However, in recent years, increasing research has pointed out that, rather than evaporation of trust, it might be more accurate and less ambiguous to interpret blockchain-enabled "trustlessness" as a shift of trust from centralized authorities to blockchain technology and the associated people, such as developers and miners (Werbach 2018; Sas and Khairuddin 2017; Al-Saqaf and Seidler 2017; Ostern 2018). The trust shifted to blockchain technology is sometimes framed as trust in code (Maurer et al. 2013; Velasco 2017), trust in quasi-entities (Reijers and Coeckelbergh 2018), or trust in algorithmic authority (Lustig and Nardi 2015).[1]

While the above efforts suggest the viability of blockchain trust, little research has explicated the nature and ethical limits of this trust form through the lens of philosophical theories of trust. Conceptualizing blockchain trust in one way or another largely impacts how we understand the role played by this technology in our lives, and more importantly, it can carry different implications shaping what values we want trustworthy blockchains to embody. In philosophical studies of trust, several important accounts have been proposed to understand trust in technologies (Taddeo 2010; Coeckelbergh 2012; Nickel 2013). Yet what is intriguing and unique about blockchain trust is that it seems not just a matter of trust in technologies. The blockchain's potential for providing a self-sufficient way to reach consensus facts without third-party authorities indicates the system's mixed role transgressing between a technological system for achieving functional services and an institution-like entity that can organize relatively stable patterns of social practices. Such a combined position explains why this technology has been referred to as both "an institutional technology" and "a technological institution" (Davidson et al. 2018; Reijers et al. 2016). The institutional aspect of blockchain technology emphasizes the importance of understanding and evaluating blockchain trust based on what we know about trust in institutions. Neglecting the richness and moral significance of institutional trust may conceal what we expect from trustworthy institutions, and thus reduce the tasks that should be addressed by blockchain systems as alternatives to traditional institutions to merely technical aspects.

Unlike trusting a particular person, our trust placed in institutions and those who fill institutional roles is often considered abstract, diffuse, and impersonal (Govier 1997; Luhmann 1979; Coeckelbergh 2015). According to Luhmann (1979, 48), the

[1] While there are different setups of blockchain technology such as public/private/consortium blockchains, for analysis reason, the blockchain systems discussed in this chapter refer to public blockchains such as Bitcoin and Ethereum.

aim of this form of trust (or "system trust" in his term) is to reduce the complexity of interacting with different functional systems (e.g., a financial system) usually seen as necessary for individuals to live in a complex modern society. Although Luhmann's account does not delve into the normative aspect of institutional trust, it takes complex social processes, norms, and the functionality of social systems as important sources of trust that govern our shared expectations about the right ordering and stability of the systems. This view seems to provide a good starting point for understanding blockchain trust given the similar striking capacity of this technology for delivering predefined normative values. As many scholars have argued, the original blockchain is not value-neutral; it is the manifestation and reinforcement of particular norms and values over others (De Filippi and Loveluck 2016; Golumbia 2015; De Filippi and Hassan 2018; Ishmaëv 2019). Besides, applications of this technology may further transform social relations in a way that follows the systems' rigid and non-negotiable features (Reijers and Coeckelbergh 2018). The shared capacity between institutions and blockchains for being normative entities indicates the possibility of understanding blockchain trust in terms of the features of institutional trust.

With these considerations, this chapter presents a novel and meaningful way of conceiving of trust in blockchain technology by analogy with what we understand of trust in institutions. In support of the analysis, two core issues revolving around blockchain trust are examined. First, by discussing how blockchain trust resembles our predictive and normative expectations towards institutions, the nature of blockchain trust is argued to be best understood as a special type of trust in institutions with trust-inviting elements built into, rather than outside, its technical infrastructure. Second, what we know about institutional trust is further utilized as an analytic tool on which the ethical limits of blockchains' trustworthiness can be reflected. As such, a constructive reflection on blockchain trust as a special form of trust in institutions is provided, with the aim of providing perspectives from which the trustworthiness of blockchain applications could be responsibly improved.

It should be emphasized that such an analysis of blockchain trust touches on two core questions of trust as a relational structure:

(1). What constitutes the trustor's trust? This question primarily concerns the trust-establishment phase.

(2). What constitutes the trustee's trustworthiness? This question focuses more on the trust-evaluation phase.

By elucidating these two questions in the blockchain context, this chapter not only contributes to clarifying what people may expect from specific blockchains and how such institution-like systems should be assessed, but more importantly, it builds a normative conception of blockchain trust that could help proactively shape blockchain applications and their effects. The analysis provided in this chapter, thus, provides a way of doing blockchain ethics via a constructive reflection on the most crucial value (i.e., trust) associated with this disruptive technology.

This chapter will proceed as follows. It begins by discussing the trust revolution brought about by the technical potential of blockchains for creating various virtual institutions that could replace third-party authorities in promoting trusted interactions. Given the importance and possibility of exploring blockchain trust in terms of trust in institutions, the chapter then embarks on a detailed investigation of institutional trust. Next, the institutional trust account proposed is applied to analyse the normative aspects of blockchain trust, allowing blockchain trust to be understood as a plausible and meaningful form of trust resembling institutional trust. Here the ethical limits of blockchains' trustworthiness are discussed as a result of removing central authorities. Finally, the limits articulated are used as perspectives from which blockchain implementations' trustworthiness can be properly improved by facilitating a shift of responsibility to the developers and networks of users.[2]

5.2. The trust revolution: Blockchain systems as virtual institutions

The following section discusses how blockchain systems disrupt a traditional way of facilitating trusted interactions. First, it briefly clarifies what is meant by the term "trust" in philosophy and the role played by third parties in promoting interactions between people who have no trust in each other. It then looks at the technical potential of blockchain technology for eliminating the need for third parties and thus revolutionizing the way we trust.

5.2.1. Understanding trust and the role played by third-party authorities

As much research into trust would agree, trust is an elusive concept that has multi-faced nature (Simon 2013; Baier 1994; Ess 2010). In the most general sense, trust can be regarded as a phenomenon that develops within a relation that requires at least two parties: a trustor and a trustee (McLeod 2020; Coeckelbergh 2012; Taddeo 2010). In trust discourse, scholars have proposed several important accounts that can help tease out the complex nature of the trust notion. Gambetta (1988, 217), for example, suggests a rational account by defining trust as a probabilistic assessment of the likely behavior of another. Likewise, Coleman (1990) views trust as a cognitive decision made in line with one's benefit-risk analysis of engaging in some form of cooperation with another. Despite the importance of cognitive reasons for trust, reducing the richness of trust relations to purely cognitive dimension is widely considered narrow and hollow since it does not touch upon the essence of our sense of trust (Hollis 1998; Baier 1986; Hardin 2002).

Unlike reliance, trust is a balance between confidence and vulnerability in that by trusting, one is willing to give up some discretionary power and freedom to the trustee whose behavior one cannot perfectly control or predict (Baier 1986; Wer-

[2]"User" here refers to both miners who contribute to the operation of the network and a wide range of normal users who only use the network as a way to facilitate interactions.

bach 2018). In other words, trust always involves the risk of being letting down that purely rational accounts fail to explain. In most cases, such "giving up" and risk-taking can be explained by an important non-cognitive dimension emphasized by other trust accounts, such as normative accounts that consider trust as reliance on others' responsibility for accomplishing their duties and obligations, e.g., trust in institutional representatives (Hollis 1998; Walker 2006), affective accounts in which emotions and affects play a determinant role for one to develop trust, e.g., children-to-parents trust (Weckert 2005), or motivation-based accounts that highlight the moral significance of the trustee's goodwill towards the trustor, e.g., trust between good friends (Baier 1986). Thus, despite the debate over which dimension is the primary source of trust, human trust is usually thought to be an integrated result of both cognitive and non-cognitive dimensions (Ess 2010; Taddeo 2010), and the question of which particular non-cognitive factor becomes *most* relevant is deeply entwined with the nature of the trust relation in question (Simon 2013).

The cognitive and non-cognitive dimensions of trust are closely engaged with the reasons for trust. As Ferrario et al. (2019) argue, reasons for one to trust another contain two sorts: pragmatic reasons that trusting someone or something can probably improve the trustor's well-being, such as gaining profits, building cooperation, saving time and energy, and preserving moral values, and epistemic reasons that relate to the trustor's belief in the trustee's trustworthiness. This means, on the other side, trust is deeply relational—engaged with a particular person's needs and interests—and highly contextual—impacted by whether there are better alternatives in a specific context. On the other side, for the trustor, the value of trust is achievable insofar as the trustee is in fact trustworthy with respect to the entrusted task (Hardin 2002; Nickel 2015). Both sides show the importance of the trustor's awareness of the trustee's trustworthiness.

Unlike trust, trustworthiness is a quality that indicates to others whether one will act as expected (Taddeo 2010). It allows others to expect the benefit and risk of placing trust reasonably. Yet, arriving at cogent reasons for trust requires the trustor to be familiar with the potential trustee which also explains why trust in tightly-knit groups is widely regarded as the original form of trust (Luhmann 1979). When the two parties are not familiar with-, or do not already trust each other, a credible third-party or middleman who can help them build trusted interactions is often needed. For example, think of Alice and Bob as two teenagers who have no trust in each other but would like to trade stamps, and think of Clark as a credible stamp shop owner in town, who offers the service of facilitating stamp trading for earning a good reputation and small fees. In this case, it is fairly reasonable for Alice and Bob to proceed with the trade through the hand of Clark since, with him, they could trade safely without the need for trusting each other.

This simple way of facilitating trusted interactions between individuals is in fact prevalent in almost all sorts of modern economic activities. For high-stakes decisions and more complex interactions, Clark's role is usually filled by trusted institutions that could provide formal endorsements and indemnity by protecting the

participants' vulnerabilities and interests. By placing trust in third-party authorities rather than one another, participants reduce their risk. Such risk-reducing interactions make it reasonable for participants to engage in an activity. From the direct communicative actors, to a credible third-party like Clark, and then to formal institutions, the shift of trust highlights the fact that, for transactions between strangers, the goods of trust for the trustor are not necessarily linked to a particular trustee but can be achieved by alternatives contingent on social and technological development. As will be discussed below, this also explains why blockchain technology is frequently viewed as an alternative to third-party institutions.

5.2.2. The elimination of third parties: Blockchains as alternatives

In the context of online transactions, institutions for promoting trusted interactions are mainly banks, firms, markets, exchanges, governments, and the relevant financial and legal systems they collectively furnish. While these institutions provide necessary means for economic activity to be processed recurrently and reliably, dependence on centralized entities not only involves extra costs, risks, and uncertainties but also relies heavily on their integrity and credibility (Nakamoto 2008). Along with the financial crisis of 2008, increasing concerns about the drawbacks and insufficiency of trusting centralized authorities have been expressed. With this background, blockchain technology, as the decentralized solution enabling the Bitcoin project, first came to prominence with realizing the ledger function that used to be provided exclusively by centralized institutions.

A blockchain is a distributed transactional database that enables continuous transitions of system states without the intervention of any intermediary (Glaser 2017). The core quality valued by blockchain start-ups, as Dupont and Maurer (2015) state, is blockchains' potential for being record-keeping devices. A record-keeping device (i.e., a ledger) provides a way to create consensus on the factual recording of the state of an economy, which is considered of pivotal importance for coordinating modern commerce (Davidson et al. 2018). Traditionally, such a ledger is issued and kept exclusively by a central authority that monitors all transactions that have ever taken place. By contrast, the Bitcoin blockchain adopts a decentralized and transparent approach with all valid transactions publicly announced to a large network of computers in chronological order, providing an alternative way to ensure the accuracy of transaction records and prevent double-spending attempts.

In cases where no trusted authority is involved, achieving a factual and shared state of the ledger is the main issue faced by any new solution. The Bitcoin blockchain solves this problem by using a consensus algorithm (i.e., proof-of-work) based on cryptographic tools and a series of consensus rules such as a fixed block format, the longest chain rule, and the incentive mechanism. More specifically, new transactions are collected from the memory pool and grouped into a block with other information required by the block format. Nodes competing for the single power of adding a new block to the chain are called miners. They are incentivized to join the

competition by profitable rewards in return for their computation power and electricity. The competition requires them to find a solution to a complex cryptographic puzzle for the issued block as the proof-of-work. After a miner solves the puzzle, the result will be broadcasted to the network and verified by other nodes. And if it is valid, the block will be added to the chain, and that miner can get some coins as rewards. For the blockchain, modifying data in a past block is extremely hard and costly since a malicious user has to assemble a majority of the hash power to redo the proof-of-work of the target block and all blocks after it. Regarding this, the peer-to-peer network is considered robust enough to maintain a single history of order in which all blocks and transactions recorded are valid and immutable (Nakamoto 2008).

In short, the blockchain is designed to facilitate a reliable ledger that could replace those issued by commercial banks, and it has been proven to be secure since no permanent damage has been done to the network since its inception. As transactions processed by the blockchain are validated and verified within the system, the network is able to provide a new basis of trust without relying on a third party (van den Hoven et al. 2019). Considering this third-party-free setting and the fact that Bitcoin meets all criteria of existing legal institutions of digital property, Ishmaëv (2017) further argues that the blockchain can function as a self-sufficient alternative institution of property alongside the traditional structure.

Furthermore, by integrating a fully-fledged, built-in programming language, the Ethereum blockchain introduces another main functionality known as smart contracts to the blockchain industry (Buterin 2013). Essentially, smart contracts are the pieces of code that can be built in a way that only the code determines what will happen once it is triggered (Glaser 2017). Such programmable contracts enable interactive services and market mechanisms to be built on distributed autonomous organizations (DAO) made of software and governed by a network of participants, further releasing this technology's potential for being an institutional technology. As Davidson et al. (2018) put forward, the fact that blockchain technology possesses many elements of market capitalism—such as exchange mechanisms, property rights, code-based law, and financial investments—makes it eligible to create a new mechanism for coordinating market economy. Such a new mechanism has the potential to complement or replace the current mechanism operated collectively by governments, firms, markets, etc. Considering the essential roles played by ledgers and contracts for constituting modernity (Reijers et al. 2016; Dupont and Maurer 2015), it is not surprising to see that ambitious blockchain-based initiatives aiming to create state-like, cloud communities (e.g., the Bitnation project) are also proposed (Tempelhof et al. 2017).

To sum up, the above discussion on blockchains' potential for record-keeping and contract-enforcement provides an analysis of how blockchains can function as virtual institutions and facilitate trusted interactions between participants. This means users of the networks can reliably interact with each other without, apparently, the need for trusting any external authority or anybody in particular. For transactions

enabled by blockchain systems, instead, everything needed seems to be users' trust in these institution-like entities. However, based on the analysis of the complex nature of trust presented above, it can be argued that more is needed to understand the relationship between "blockchain trust" and intuitional trust in addition to the similar functions provided by the two sorts of entities. In other words, a plausible notion of blockchain trust resembling the essence of trust in institutions should also explicitly refer to the normatively loaded expectations one may hold towards the systems. This requires us to first understand the rich meaning of institutional trust, including an understanding of the normative expectations towards institutions, in order to properly grasp and assess blockchain trust as a special form of trust in institutions.

5.3. A conceptual investigation of institutional trust

The functional aspects of blockchains discussed above show the technical potential of the systems for being institution-like entities. The following section works to elaborate on the nature and moral significance of institutional trust, setting the stage for further exploration of blockchain trust.

5.3.1. Beyond prediction: Normative expectations of institutions

As mentioned, trust in institutions is diffuse and does not necessarily depend on personal contact. As Alfano and Huijts (2020) put forward, trust in large-scale institutions can be non-partner-relative, meaning that trustingness and trustworthiness can be valid without a predefined partner. Also, institutional trust could be non-thing-specific. In many cases, "we did not rely on X to do A and Y to do B… but rather that we expected reliable, courteous, and orderly service" of that institution (Walker 2006). Based on the non-partner-relative and non-thing-specific structure of institutional trust, when individuals state that they trust an institution, what they are referring to and relying upon is closer to an acceptable and stable service state of that institution; i.e., they trust that it will do, in a general and abstract sense, what it is institutionalized to do. In this chapter, this institutional trust account is named *the normative account of institutional trust*. Through the lens of this account, the establishment of institutional trust is not exclusively grounded in our predictive expectations about the functions that an institution will provide, but more importantly, it relies on our normative expectations of that institution and individuals who fill the institutional roles to do what they are supposed to do. Such normative attitude links trust to the relevant trustees' responsibility for complying with their duties and obligations assigned by their institutional roles, capturing the non-cognitive dimension of institutional trust.

These two sorts of interrelated expectations echo the two dimensions of trust discussed earlier and are closely related to the trustor's reasons for trust. More specifically, the predictive expectations towards the relevant trustees are commonly

grounded in the trustor's epistemic and pragmatic reasons for trust, i.e., whether one believes that the trustees are trustworthy enough to provide specific functions that can satisfy a particular end of the trustor. Alternatively, the normative expectations towards an institution seem to engage with the trustor's pragmatic reasons for trust, depending on whether particular values and norms one favours are inherent in a given institution and can be delivered by its representatives and overall performances. For the sake of clarity, a sketch of the conceptual structure of institutional trust proposed is shown in Table 5.1.

Table 5.1: The conceptual structure of institutional trust

Dimensions of institutional trust	Institutions' qualities	Reasons for institutional trust
Predictive expectations	Functional aspects	Epistemic/pragmatic reasons
Normative expectations	Normative aspects	Pragmatic reasons

For understanding the essence of the normative account of institutional trust, it is important to discuss how people's normative expectations towards others or institutions are generated and why such expectations are essential for building trust. As Hollis (1998, 34) argues, normative expectations, under either moral or social headings, are not congruent with merely predictive expectations we hold towards functions of objects. Instead, they are grounded in the shared moral understanding that people will act as they should, and the anticipation of others' responsibility for complying with standards and behave responsively (Walker 2006). In other words, these two bases allow us to expect of others that they will act as what the standards require while holding them responsible for meeting those standards (Jones 2004). For instance, when we trust a taxi driver, a dentist, and a delivery person whom we do not know well, we expect of them that they will do their jobs correctly and meet their obligations, promises, and professional standards in a responsible way without assuming their particular concern or regard for us.

When it comes to an institution trustee, normative expectations are often grounded in our shared belief about the normative values stably tied to an institution. According to Turner (1997), institutions are a complex of norms, values, roles, and positions embedded in specific kinds of social structures that can organize fairly enduring patterns of social practices. In other words, institutions can be understood as entities carrying predefined normative qualities, such as moral, social, and legal norms. Such norms, as Lewis (2002) argues, can be interpreted as promises and commitments, providing signals for people to form their beliefs about the actions that should be followed in order to fulfil those promises. Likewise, Bicchieri (2006) depicts norms as "collectively shared scripts" that can guide common anticipation of the corresponding actions that are considered consistent and appropriate under such norms. In this regard, it can be said that normative values inherent in institutions play a significant role in shaping and guiding what one expects from institutions and their representatives.

Considering this, institutions could be viewed as viable entities of trust in the sense

that people can rely on and evaluate them in a normatively loaded way. Such expectations might be thinner than those relevant to interpersonal trust, but they are still natural and comprehensible given our everyday experience with some institutions. For example, in trusting the value of money, one presumes that the economic system and the relevant people will perform in the right way that is considered normatively desirable and established as practically trustworthy (Jalava 2006). Such trust is developed via continual, affirmative experience in using money. It supports one to believe that the system can facilitate the desirable characteristics embedded in fiat money (e.g., acceptability, durability, portability, etc.) stably and recurrently while requiring no specific guarantees. During constant interactions with different sorts of institutions and social systems, such expectations often become a default that is not necessarily assessed every time before interaction (Luhmann 1979, 50). As such, a positive feedback loop of trust between humans and the monetary system can be built via the normative qualities inherent in the system and our daily experience that help confirm the usefulness of the relevant expectations.

5.3.2. Responsible actors as the way to secure institutions' qualities

Similar to interpersonal trust, our shared understanding of the relevant human actors' responsibility for complying with their assigned obligations provides us a way to believe that our expectations invited by institutions' built-in qualities are secured, as we see from Walker (2006, 84),

"I give the bank teller my deposit, but I rely on the institution's competence and fiduciary responsibility, and the system of regulation that ensures and enforces its compliance, and the responsibility of whoever ultimately oversees that system to see to it that my money goes and stays where it is supposed to. "

In this example, responsibility is ascribed to the bank, the associated legal systems, and the human agents who fill the relevant institutional roles. When people interact with the bank teller, on the one hand, they tend to place a default trust in the whole functional system, supposing that the system will work effectively. On the other hand, they presume that the bank teller and other representatives have some sort of legal and moral responsibility for complying with obligations assigned by their institutional roles and are to be held accountable if trust is violated after the fact. As such, people take themselves to be entitled to the right order of particular services of the system and the generally responsive and trustworthy behavior of those representatives.

Although the trustor's premise that individual representatives of institutions will and should be responsible for their obligations generally remains tacit, unreflective, and nonspecific, this premise seems to be crucial for our understanding of institutional roles. As Demolombe and Louis (2006) clarify, an institutional role refers to a set of implicit and explicit rights and obligations in relation to some individuals' position or legal status in an institution. People who fill such a role, accordingly,

can be understood as individuals to whom the predefined set of norms and status functions apply (Searle and Willis 1995). According to Demolombe and Louis, an institutional role contains two sorts of properties—i.e., descriptive and normative properties—that both give specifications of the role and direct our expectations of their performances. For example, the role of bank teller is characterized by descriptive properties: to have specific professional skills and experience, and by normative properties: to have obligations to assist customers with all relevant bank services. In this case, if anyone is in fact a bank teller, it is reasonable for a customer to presume that she is competent in handling particular tasks and has responsibility for doing whatever obligations assigned by the role of bank teller.

In particular, knowing that someone will and should be responsible for doing what they ought to do provides the trustor extra confidence in institutions' trustworthiness in three types of situations. Firstly, since any trust contains the risk of being violated, such a premise makes the trustor reasonably expect that, were things to go wrong, they would ultimately identify someone to be held accountable for the wrong things and get them changed to the right way. Secondly, in ambiguous and flexible situations, the premise that some human actors can finally be found allows trustors to hope that there is some space for negotiation that could benefit themselves. Thirdly, such a premise also drives one to believe that, apart from what is required by the representatives' institutional roles, these individuals are prone to perform in a trust-responsive way since people are inherently reputation-seeking and have the desire to be well regarded (Pettit 1995).

To sum up, the analysis above proposes to understand trust in institutions as predictive and normative expectations towards institutions' performances, with a particular consideration of how the responsibility of institutions and their representatives shapes our expectations of institutions. Accordingly, institutions' trustworthiness is mainly influenced by the functional and normative aspects of their performances, as well as the responsibility of the relevant individuals for securing the realization of institutions' built-in qualities. This is in a nutshell the conceptual structure of institutional trust proposed by this chapter. This structure, on the one hand, gives form and direction to examine the extent to which blockchain trust can be regarded as a type of trust in institutions. On the other hand, and relatedly, it paves the way for a broader reflection on blockchain applications' actual trustworthiness.

5.4. Applying the above trust account to blockchain

The above institutional trust account shows the importance of the normative values built into institutions for the generation of trust and the importance of responsible actors for the realization of institutions' built-in qualities. Applying this account to trust issues related to blockchains, thus, requires an understanding of: (1) whether blockchains contain the capacity for delivering norms that could invite the corresponding expectations similar to those invited by counterpart institutions, and (2) whether the systems can provide a way to secure the realization and maintaining

of predefined functional and normative requirements. While the former question ultimately determines the plausibility of conceptualizing blockchain trust as a meaningful form of trust in institutions, the latter question directly leads us to reflect on blockchain applications' actual trustworthiness.

5.4.1. The normative relevance of blockchain trust

It is often thought that blockchain technology will be eliminating the need for trust. One important claim of this chapter is that the removal of third parties does not eliminate the need for trust, or more specifically, its non-cognitive dimension. It rather shifts the trust to blockchains. First, many empirical studies on technology trust have shown that trust as a value predisposition or a mental shortcut significantly impacts public perceptions and adoption of sophisticated technologies (Ho et al. 2010; Mah et al. 2014). This means that, in a descriptive sense, trust that takes into account non-calculative factors, such as normative and affective sources, could be to some extent seen as a prerequisite for those who lack systematic knowledge and expertise of a technology application to take a "leap of faith" and use the application. This is particularly the case when the application in question is so complicated that reaching a rational assessment of the entire system's trustworthiness is extraordinarily difficult (Corley and Scheufele 2010; Ishmaëv 2018), or when the trust is about innovative practices that are inherently uncertain (van den Berg and Keymolen 2017).

Second, this function of trust, as related to technology adoption and complexity reduction, can be supported by the essence of trust discussed earlier: trust is a way to allow people to accept the fact that dependence on another person or entity will expose them to the possibility of being harmed (Möllering 2006). Thus, on the one hand, relying on a particular technology implicitly or explicitly requires the need for trust to suspend vulnerabilities and risks involved in the use of that technology. Removing the intervention of third parties, hence, tends to shift trust from a system's human masters to the system itself and the network behind (Werbach 2018). On the other hand, given the different criteria people employ to develop and assess trust, the heterogeneity involved in humans' trust in technologies should be specified on a case-by-case basis (Taddeo 2010). Nevertheless, there are approaches that can interpret some common characteristics of trust in technologies. Coeckelbergh (2012), for example, proposes a phenomenological-social approach that captures trust as an emergent and/or embedded property of social relations. In this way, he argues that, as technologies are already part of our lives, trusting technologies is less under the control of individuals but more like a default that emerges from social relations.

The idea of conceptualizing trust in technologies in terms of institutional trust—which this chapter endorses—is essentially an effort to interpret more specific inner connections between humans and technologies. Such an idea is not new, but it has not yet been systematically explored. Nickel (2020; 2013) notes that technologies can be direct objects of trust since they are subscripted by some of the evaluation

standards that are used to reach and justify our trust decisions towards institutions. The way we evaluate sophisticated systems, as he argues, is not merely about whether their functions are reliable or not (like a hammer); we also care, in an evaluative sense, whether they are doing things correctly. In this claim, an analogy between institutions and technologies is drawn in virtue of their similar capacity for being entities that can invite normative evaluation of their performances. The view that technologies contain normative aspects can be better explained by Moor's (2006) clarification of the two categories of normative viewpoints. As he argues, technologies are normative entities because they can be evaluated by:

(1). Non-moral normative viewpoints, which assess particular technologies' performances in terms of their intended purposes or design norms. Such norms can be interpreted as the principles and objectives guiding technologies' performances that do not necessarily draw on ethical consequences;

(2). Moral normative viewpoints, which take moral norms to evaluate those technical performances that are of ethical relevance. This could be the case, for example, when the technical performances can generate ethical consequences or contain built-in moral considerations (Tavani 2015).

On the basis of these two ways of understanding the normative aspects of technologies and the institutional trust account articulated, it can be said that technologies resemble institutions in their design capacity for carrying normative values and inviting relevant expectations about what they are supposed to do. In this regard, it seems that technologies could be plausibly viewed as objects of trust in the sense that they could be relied upon and evaluated in a normatively loaded manner.

The analogy discussed above becomes more striking when the technology trustee in question is the original blockchain. Not only does the system share comparable norm-delivering capacity with institutions, but it is explicitly designed to carry out exactly the same core functions of its counterpart institutions and deliver the set of design norms that are considered desirable in the economic context. For this reason, this chapter argues that blockchain trust is not simply a type of trust in technologies (like trust in an autonomous vehicle) that can be framed as institutional trust in a general sense, but blockchain trust is itself a form of trust in institutions. For creating an alternative to the trust model enabled by third-party authorities, as articulated, the most daunting task of the peer-to-peer network is to reach a shared state of the database that can ensure the validity and irreversibility of all transactions, which is also viewed as the core functionality provided by every blockchain system (Glaser 2017). Essentially, the normative purpose of this task is to provide a global source of truth on which the associated values required for empowering the decentralized solution can be reasonably approached. Such values primarily include (a) data integrity, which indicates the completeness and accuracy of the information shared; (b) data transparency that prevents counterfeits and dishonest behavior by improving information symmetry and audit compliance; (c) data authentication, which ensures a reliable process to verify the identity of a person

or a single piece of data; and (d) data security that makes sure that records issued by the network are tamper-resistant and risk-tolerant. Following Moor's clarification of normative entities, these values could be viewed as the design norms built into the blockchain's infrastructure, which can readily inspire users to generate the corresponding expectations.

Thus, if we consider institutions as a complex of norms and values folding into particular social structures for delivering relatively stable services, the blockchain can be seen as a further step that attempts to keep only the core functionality and certain normative ideas of its counterpart institutions while eliminating these entities as well as their bureaucratic processes and power holders. Moreover, certain parts of the blockchain incorporate explicit considerations as a resistance to the power dynamics enabled by centralized authorities, bringing about effects and implications that are not just normatively but also morally relevant. According to Tavani (2015), Moor's two kinds of standards for evaluating the impacts of a specific technology (i.e., design norms and moral norms) can lead to different levels of trust, from low to high. In this regard, if one's expectations towards the blockchain are about its morally relevant features, the level of trust the trustor places in the system can be higher than those who bear no such expectations.

A proper understanding of the moral features tied to the original blockchain's performances requires a brief review of the moral significance of cryptography-enabled data decentralization. In 1985, David Chaum proposed the idea of using decentralized solutions based on cryptographic techniques to solve moral issues—such as mass surveillance, erosion of democratic rights, and opinion manipulation—entailed by centralized computer systems (Chaum 1985). As a crucial component of Bitcoin protocol, cryptographic techniques thus provide a good starting point for establishing a global decentralized infrastructure that could dilute the power of monopolies and contribute to protecting moral values such as freedom, autonomy, and privacy (Ishmaëv 2019; Scott 2014). In this respect, the distributed database technology might be considered more praiseworthy than traditional solutions, especially in cases when these values are already at stake. For similar reasons, the blockchain has been depicted as a neoliberal project or a libertarian dream through which the control of nation-states on the economy can be reduced so that "governing without governments" might be achieved (De Filippi and Loveluck 2016). Systems built on blockchain technology, thus, have the capacity for bringing about significant effects on challenging authorities and shaping people's understanding of the power-relation of the society (Reijers and Coeckelbergh 2018).

In this regard, the specific norms and values presented by blockchain implementations are very likely to attract the participation of those actors who favour such normative ideas. Also, the profound norm-delivering capacity of this technology inevitably attracts those who are interested in using such capacity for their own purposes (Ishmaëv 2019). These normative aspects of blockchains, thus, can be valid and plausible reasons to invite users' trust. Such reasons fall into the category of the pragmatic reasons discussed earlier, which are deeply relational and engaged

with the trustor's specific interests and needs that might be met by a given trustee's performances. With all these considerations, it seems fair to say that blockchain trust is grounded in and goes beyond our trust placed in institutions. By removing the role of internet aggregators while providing an alternative way to help achieve the value of trust for the trustor, blockchain technology brings about a fundamental change in the way we trust and benefit from the goods of trust. Thus, the normative conception of blockchain trust as a special type of trust in institutions is proposed by analysing the nature of human-to-blockchain relation against the background of theories of trust and institutional trust, which should not be confused with any descriptive claim about trust.

5.4.2. The ethical limits of blockchains' trustworthiness

The above analysis shows the appropriateness and plausibility of conceptualizing blockchain trust based on trust placed in institutions. Nevertheless, the goods of trust only accompany well-grounded trust (McLeod 2020). Thus, it is crucial to distinguish between how trust can be invited and how trust should be evaluated. From the perspective of blockchain ethics, while the former considers the importance of addressing the conceptual vacuum of blockchain trust by understanding its nature, the latter focuses on assessing the implications of blockchain trust with the aim that more well-grounded trust can be achieved.

Despite the advantages produced by blockchains' disruptive features, along with the erasure of central points, blockchain applications' trustworthiness also raises ethical concerns. This chapter argues that the decentralized novelty of blockchain technology has dual effects on trust. It eliminates the risk, cost, and complexity related to third parties while simultaneously crowding out the pivotal role of institutions and a cadre of representatives in meeting their assigned obligations and securing the functional systems' reliable performances.

This means that blockchains' decentralized nature carries significant implications and consequences for issues impacting trust that are of ethical relevance. First, individual representatives do play an important role, especially in unexpected situations. Although there is a risk that, after the fact, human actors of institutions are shown to be incompetent and not responsive to their duties and obligations, these people can be held accountable for their misconducts and even facing punitive measures. In comparison, the lack of control over a blockchain's performance and the lack of clear attribution of responsibility in blockchain communities imply that, were things to go wrong (e.g., loopholes and attacks), nobody would be held accountable for the incidents, and the irreversible nature of the system leaves almost no room for recourse. As Reijers and Coeckelbergh (2018) point out, the high level of blockchains' rigidity is achieved at the cost of a reduction in the dynamic understanding of the freedom and responsibility of the actors involved. In this respect, a market economy built on blockchains may put its trustors in a more vulnerable position than the trust model involving centralized authorities, particularly considering those small networks where attacks are easier to occur.

Second, there are risks deriving from unreasonable normative expectations. Although many expectations related to blockchains seem plausible, such as those related to the design norms and moral norms of the original blockchain discussed earlier, it is not at all surprising that some expectations are not evidence-based. Unrealistic normative expectations, as Buechner and Tavani (2011) mention, also exist in human-to-institution trust. Yet, the fundamental difference between those invited by institutions and blockchains is that the relevant qualities of blockchains are often hidden and less guaranteed (Ishmaëv 2019). Think of the Bitnation project that purports to create blockchain-enabled democratic communities online. A fundamental concern of this idea is that democratic communities in civil society are created by negotiation and compromise between members with diverse backgrounds, conflicting interests, and different conceptions of the common good, but not by a homogenous group of participants who can voluntarily join and leave (De Filippi and Hassan 2018). Thus, the intention of transforming territorial associations into blockchain-based communities would be fatal to democratic values as it tends to eclipse other types of moral and political reasoning. A more profound ethical concern is the non-neutrality of blockchain technology itself. As Golumbia (2015) argues, the basic setting of blockchain technology is considered deeply political, with "right-wing, libertarian, and anti-government" ideology embedded. Organizing democratic communities via blockchains, thus, makes democracy vulnerable to the ideological biases inherent in this technology (Dumbrava 2018). In this regard, if advocates normatively expect that the project can safeguard and promote democracy adequately, their expectations will be frustrated due to the deeply flawed assumptions built into the system.

At the very least, institutional norms are usually under constant scrutiny of democratic debates and examined by long-lasting practices (De Filippi and Hassan 2018). In contrast, we could say that, in addition to trustor's lack of investigation, the generation of unrealistic expectations towards blockchain implementations may also be caused by the absence of actors who are formally responsible for explicating, scrutinizing, and updating the set of assumptions inscribed into the systems and monitoring the actual performance of the systems' norm-delivering capacity. As Jones (2012) argues, trustworthiness requires that trustees are willing and able reliably to signal to others the domains that they are competent and will be responsive to others' dependency. Therefore, compared to institutions where the implementation of the relevant normative ideas is secured by a number of human actors and well-established procedurals, blockchains are designed to float merely in the rules of algorithms. This raises ethical concerns over the reliability of the systems' normative qualities.

The above analysis clarifies that blockchains' decentralized nature and the implications and effects of this decentralization on trust are double-edged. Without the backing of credible parties, the systems put more burden and risk on users themselves without proper measures to redeem unexpected situations and guarantee the systems' actual norm-delivering performances. All these claims seem to point to the ever-pressing need of trustors for being vigilant and reflective knowers. To reach

well-grounded trust in a digital context, resonated with Simon's (2010) view, not only do trustors have a duty to check the integrity and competence of the trusted entity, but they must also scrutinize their standards for evaluating others' trustworthiness. Simply put, users need to be more responsible for their trust decisions as a result of distrusting third parties.

5.5. Towards trustworthy blockchains: A shift of responsibility

Seeking to make trust more well-grounded, nevertheless, is just one side of the coin and restricted by subjectivity-specific differences with respect to users' knowledge, time, and resource. As Keymolen (2019) points out, our ability to establish trust is affected also by the social context in which we are positioned, such as social roles that make each other's actions and expertise more predictable. In the blockchain context, to effectively respond to the challenges faced by the technology, this chapter argues that, apart from a wide range of users, more responsibility should be shifted to developers and active network peers (i.e., miners) who are associated with the actual performance of blockchain applications. Clarifying the specific roles played by these groups and reframing their responsibilities accordingly provide a way to improve our abilities to develop trust by addressing a focus on understanding what is at stake for the development of trustworthy blockchains. Such an effort can be used to inform the design and decision-making related to blockchains-based systems, building affordances that foster warranted trust and foreclose affordances that would undermine warranted trust. In this way, the ethical limits of blockchains' infrastructure discussed earlier are used as perspectives from which the trustworthiness of blockchains might be gradually improved.

While blockchain technology is designed to eliminate the need for centralized authorities, it is not designed to remove the reliance on developers who maintain the actual codebase through the workflow and determine the functionality and the main values of the system (Glaser 2017). As Nickel (2013) clarifies, developers are presupposed to have two trust-related tasks: the first is to make the system as reliable as possible, and the second is to identify the system's trustworthiness to people in a position to trust that system. The core issue here, coupled with the two dimensions of blockchain trust discussed above, is to sufficiently show that the disruptive functions of blockchain technology together with the meaningful normative values imparted can be realized in practice.

For the functional aspects, compared to the big promises made by the original blockchain's whitepaper, its current reference implementation, referred to as Bitcoin Core, is facing many intractable technical issues such as low throughput, high latency, and a tremendous waste of electricity, which are especially apparent in comparison with the efficiency of the incumbent payment gateways they tend to replace, e.g., Visa and PayPal (Swan 2015). While it is clear that the development and practical applicability of blockchain implementations are still in their infancy,

solving the above issues is the shared responsibility of the developer community inherent to their role in the whole ecosystem.

Moreover, given the risk of unexpected situations harming the basic functions of the network, explicit strategies for self-governance and crisis response within the developer community and the peer-to-peer network are hardly optional tasks. A valuable lesson learned from the most infamous incident that occurs in the Ethereum blockchain (i.e., the DAO hack) is that decentralization should not be either-or.[3] The accident shows that in order to protect the network's overall interests, certain sacrifice of the blockchain's immutability and decentralization is in fact considered appropriate and acceptable for the majority of the community. In this sense, effective self-governance adopted to ensure the proper performance of a blockchain might be as useful as the safeguard provided by centralized institutions. However, the current governance structure of blockchain communities is quite technocratic, and the responses provided are relatively arbitrary, two facts which cause concerns about the fragility of the community's decision-making processes and its capacity for dealing with incidents. (De Filippi and Loveluck 2016). Thus, what is lacking is a generic, well-established governance mechanism ready to be applied to interpret and respond to possible contingencies. A way in which laws and regulators can here truly help, as Werbach (2018) notes, is not to offer specific governance rules for the community but to provide the community with jurisprudential insights into how rules should be formulated and enforced in a formal way.

As discussed, the normative ideas inscribed into blockchains are also crucial sources of trust and important criteria for evaluating trust. However, for plenty of blockchain implementations, these ideas are not transparent and well-scrutinized, which makes them easy to be flawed and generate undesirable effects on users and society at large. Some of the assumptions simply fall into naive technological determinism, just like the case of Bitnation. As professionals who have the direct ability to use technical means to express human values, developers and designers can play a vital role in advancing responsible technological innovations by helping realize these values properly (van den Hoven et al. 2015). Indeed, many current proposals seek to embed particular desirable normative goals into blockchain design, such as Enigma and Zcash that aim to create privacy-preserving blockchains and Datawallet that is designed to facilitate data ownership. What is lacking, based on the discussion provided in the above subsection, is a satisfactory explanation and justification of how these norms are embedded in and embodied by the technical design. In this regard, making the normative goals transparent to the public is just the first step. Constantly scrutinizing and updating the systems' built-in assumptions on a case-by-case basis is of central importance for improving the normative qualities of blockchains (Nickel 2013; Ishmaëv 2019), and these are the aspects that developers, network peers, philosophers, policy-makers can all take part in and contribute to approaching more trustworthy blockchains.

[3]For more information about the hack, see https://medium.com/@ogucluturk/the-dao-hack-explained-unfortunate-take-off-of-smart-contracts-2bd8c8db3562.

5.6. Conclusion

This chapter has critically discussed blockchain trust by analogy with trust placed in institutions. Doing so provides a close philosophical reflection on the nature and ethical limits of this trust form. As a result of blockchain's double-edged peculiarities, blockchain trust is characterized, on the one side, as a form of trust grounded in- and going beyond institutional trust. By coding the normative values and technical properties into its basic infrastructure, the original design of blockchain technology touches the most intriguing aspect of trust, i.e., we want our trust to be warranted, more than ever, to dispel our anxieties and worries about the discretionary power possessed by third parties with the hope that the vulnerabilities and risks engendered by placing trust can be minimized to the greatest extent possible. On the other side, blockchain-based systems are confronted with challenges to their actual trustworthiness for functioning as an institution-like entity. Reframing the responsibility shifted to the relevant groups of people in the blockchain context is an essential component of a strategy to address the ethical and societal challenges posed by this disruptive technology. As such, the institutional trust concept is used as an analytical tool to disentangle the double-edged effects of blockchain on trust, and informing ways in which the trustworthiness of blockchain applications could be gradually improved.

References

Al-Saqaf, W., & Seidler, N. (2017). Blockchain technology for social impact: Opportunities and challenges ahead. Journal of Cyber Policy, 2(3), 338-354.

Alfano, M., & Huijts, N. M. A. (2020). Trust and distrust in institutions and governance. In J. Simon (Ed.), Handbook of trust and philosophy. Routledge Taylor & Francis Group.

Baier, A. (1986). Trust and antitrust. Ethics, 96(2): 231-260.

Baier, A. (1994). Trust and its vulnerabilities. Moral prejudices, 130-151.

Bicchieri, C. (2006). The grammar of society: The nature and dynamics of social norms. New York: Cambridge University Press.

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

Buterin, V. (2013). Ethereum white paper. GitHub Repository. https://ethereum.org/en/whitepaper/. Accessed January 5, 2018.

Chaum, D. (1985). Security without identification: Transaction systems to make big brother obsolete. Communications of the ACM, 28(10), 1030-1044.

Coeckelbergh, M. (2012). Can we trust robots? Ethics and information technology, 14(1), 53-60.

Coeckelbergh, M. (2015). Money machines: Electronic financial technologies,

distancing, and responsibility in global finance. Farnham: Ashgate.

Coleman, J. S. (1994). Foundations of social theory. Cambridge: Harvard University Press.

Corley, E. A., & Scheufele, D. A. (2010). Outreach gone wrong? When we talk nano to the public, we are leaving behind key audiences. Scientist, 24(1), 22.

Davidson, S., De Filippi, P., & Potts, J. (2018). Blockchains and the economic institutions of capitalism. Journal of Institutional Economics, 1-20.

De Filippi, P., & Hassan, S. (2018). Blockchain technology as a regulatory technology: From code is law to law is code. arXiv preprint arXiv:1801.02507.

De Filippi, P., & Loveluck, B. (2016). The invisible politics of bitcoin: Governance crisis of a decentralized infrastructure. Internet Policy Review, 5(4).

Demolombe, R., & Louis, V. (2006). Norms, institutional power and roles: Towards a logical framework. In International Symposium on Methodologies for Intelligent Systems (pp. 514-523). Springer, Berlin, Heidelberg.

Dumbrava, C. (2018). Citizenship forecast: Partly cloudy with chances of algorithms. In R. Bauböck (Ed.), Debating transformations of national citizenship (pp. 299-303). Cham: Springer.

DuPont, Q., & Maurer, B. (2015). Ledgers and law in the blockchain. Kings Review. http://kingsreview.co.uk/artic les/ledgers-and-law-in-the-blo
ckchain/. Accessed November 28, 2018.

Ess, C. M. (2010). Trust and new communication technologies: Vicious circles, virtuous circles, possible futures. Knowledge, Technology & Policy, 23(3-4), 287-305.

Ferrario, A., Loi, M., & Viganò, E. (2019). In AI we trust incrementally: A multilayer model of trust to analyze human-artificial intelligence interactions. Philosophy & Technology, 1-17.

Gambetta, D. (1988). Can we trust trust? In Gambetta, Diego (ed.), Trust: Making and breaking cooperative relations, 213, 214. Oxford: Basil Blackwell.

Glaser, F. (2017). Pervasive decentralization of digital infrastructures: A framework for blockchain enabled system and use case analysis. In Proceedings of the 50th Hawaii International Conference on System Sciences, 1543-1552. Hawaii, United States.

Golumbia, D. (2015). Bitcoin as politics: Distributed right-wing extremism. In G. Lovink, N. Tkacz, and P. de vries (Eds), MoneyLab reader: An intervention in digital economy. Amsterdam: Institute of Network Cultures.

Govier, T. (1997). Social trust and human communities. Montreal: McGill-Queen's University Press.

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

Ho, S. S., Scheufele, D. A., & Corley, E. A. (2010). Making sense of policy choices: Understanding the roles of value predispositions, mass media, and cognitive processing in public attitudes toward nanotechnology. Journal of Nanoparticle Research, 12(8), 2703-2715.

Hollis, M. (1998). Trust within reason. Cambridge: Cambridge University Press.

Ishmaëv, G. (2017). Blockchain technology as an institution of property. Metaphilosophy, 48(5), 666-686.

Ishmaëv, G. (2018). Rethinking Trust in the Internet of Things. In R. Leenes, R. van Brakel, S. Gutwirth,
P. de Hert (Eds.), Data protection and privacy: The internet of bodies (pp. 203–230). Oxford: Hart Publishing.

Ishmaëv, G. (2019). Open sourcing normative assumptions on privacy and other moral values in blockchain applications (doctoral dissertation). Delft University of Technology, the Netherlands.

Jalava, J. M. (2006). Trust as a decision: The problems and functions of trust in Luhmannian systems theory (Niklas Luhmann). https://helda.helsinki.
fi/bitstream/handle/10138/23348/trustasa.pdf?sequen. Accessed June 6, 2020.

Jones, K. (2004). Trust and terror. In P. DesAutels
M. U.

Walker (Eds.), Moral psychology: Feminist ethics and social theory (pp. 3-18). Maryland: Rowman & Littlefield.

Jones, K. (2012). Trustworthiness. Ethics, 123(1), 61-85.

Keymolen, E. (2019). When cities become smart, is there still place for trust. European Data Protection Law Review., 5, 156.

Lewis, D. (2002). Convention: A philosophical study. Oxford: Blackwell Publishers Ltd.

Luhmann, Niklas (1979) Trust and power. Chichester: John Wiley.

Lustig, C., & Nardi, B. (2015). Algorithmic authority: The case of Bitcoin. In 48th Hawaii International Conference on System Sciences (pp. 743-752). Hawaii, United States.

Mah, D. N. Y., Hills, P., & Tao, J. (2014). Risk perception, trust and public engagement in nuclear decision-making in Hong Kong. Energy Policy, 73, 368-390.

Maurer, B., Nelms, T. C., & Swartz, L. (2013). "When perhaps the real problem is money itself!": The practical materiality of Bitcoin. Social Semiotics, 23(2), 261-277.

McLeod, C. (2020). Trust. The Stanford Encyclopedia of Philosophy. https://pla to.stanford.edu/archives/fall 2020/entries/trust/. Accessed September 6, 2020.

Möllering, G. (2006). Trust: Reason, routine, reflexivity. Amsterdam: Elsevier.

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

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Bitcoin. https://bitcoin.org/bitcoin.pdf Accessed July 1, 2016.

Nickel, P. J. (2013). Trust in technological systems. In M. J. de Vries, S. O. Hansson & A. W. M. Meijers (Eds.), Norms in technology, philosophy of engineering and technology (pp. 223-237). Dordrecht: Springer.

Nickel, P. J. (2015). Design for the value of trust. In J. van den Hoven, PE. Vermaas, & I. van de Poel (Eds.), Handbook of ethics, values, and technological design: Sources, theory, values and application domains, 551-567. Dordrecht: Springer Netherlands.

Nickel, P. J. (2020). Trust in engineering. In D.P. Michelfelder & N. Doorn, (Eds.), Routledge companion to philosophy of engineering.

Ostern, N. (2018). Do you trust a trust-free transaction? Toward a trust framework model for blockchain technology. In Thirty Ninth International Conference on Information Systems, San Francisco, United States.

Pettit, P. (1995). The cunning of trust. Philosophy & Public Affairs, 24(3), 202-225.

Reijers, W., & Coeckelbergh, M. (2018). The blockchain as a narrative technology: Investigating the social ontology and normative configurations of cryptocurrencies. Philosophy & Technology, 31(1), 103-130.

Reijers, W., O'Brolcháin, F., & Haynes, P. (2016). Governance in blockchain technologies & social contract theories. Ledger, 1, 134-151.

Sas, C., & Khairuddin, I. E. (2017). Design for trust: An exploration of the challenges and opportunities of bitcoin users. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 6499-6510). ACM. Denver, United States.

Scott, B. (2014). Visions of a techno-leviathan: The politics of the Bitcoin blockchain.https://www.e-ir.info/2014/06/01/visions-of-a-techno-leviat
han-the-politics-of-the-bitcoin-blockchain/. Accessed September 15, 2020.

Searle, J. R., & Willis, S. (1995). The construction of social reality. New York: The Free Press.

Simon, J. (2010). The entanglement of trust and knowledge on the Web. Ethics and Information Technology, 12(4), 343-355. https://doi.org/10.
1007/s10676-010-9243-5.

Simon, J. (2013). Trust. In Pritchard, D. (Ed.): Oxford bibliographies in philosophy. New York: Oxford University Press

Swan, M. (2015). Blockchain: Blueprint for a new economy. Sebastopol: O'Reilly Media, Inc.

Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-trust. Minds and machines, 20(2), 243-257.

Tavani, H. T. (2015). Levels of trust in the context of machine ethics. Philosophy & Technology, 28(1), 75-90.

Tempelhof, S. T., Teissonniere, E., Tempelhof, J. F., & Edwards, D. (2017). Bitnation white paper. GitHub repository. https://github.com/Bit-Nation/Pan
gea-Docs. Accessed January 20, 2019.

Turner, J. H. (1997). The institutional order: Economy, kinship, religion, polity, law, and education in evolutionary and comparative perspective. New York: Longman Publishing Group.

van den Berg, B., & Keymolen, E. (2017). Regulating security on the Internet: Control versus trust. International Review of Law, Computers & Technology, 31(2), 188-205.

van den Hoven, J., Pouwelse, J., Helbing, D., & Klauser, S. (2019). The blockchain age: Awareness, empowerment and coordination. In D. Helbing (Ed.), Towards digital enlightenment (pp. 163-166). Springer, Cham.

van den Hoven, J., Vermaas, P. E., & Van de Poel, I. (2015). Design for Values: An Introduction. In J. van den Hoven, PE. Vermaas, & I. van de Poel (Eds.), Handbook of ethics, values, and technological design: Sources, theory, values and application domains. Dordrecht: Springer Netherlands.

Velasco, P. R. (2017). Computing ledgers and the political ontology of the blockchain. Metaphilosophy, 48(5), 712-726.

Walker, M. U. (2006). Moral repair. New York: Cambridge University Press.

Weckert, J. (2005). Trust in cyberspace. In R. J. Cavalier (Ed.), The impact of the internet on our moral lives (pp. 95-117). Albany: SUNY Press.

Werbach, K. (2018). The blockchain and the new architecture of trust. Cambridge: MIT Press.

Summary

The widespread movement of facilitating trusted interactions by using digital systems to formalize procedures and practices is arguably one of the most revolutionary developments that are transforming our world of collaboration and cooperation. From closely tight-knit communities to open society, from familiarity-based to reputation- and institution-based then to technology-based interactions, the proliferation of trust-inviting systems continually shapes the way we connect to others, emancipating efficacy and productivity from region, time, and energy limitations. Nevertheless, this is just half of the story. As this study shows, innovations in this field open up new avenues that can manipulate and exploit our trust, bringing undesirable social and moral consequences to the present and future. Normative reflections on the question – how can trust-inviting systems foster trust appropriately? – provided by this thesis is essentially a humble endeavor to examine and improve the appropriateness of the assumptions and embodiments of the trust concepts adopted by these systems.

Taking several prevalent digital systems that contain explicit design goals related to trust as studying cases, the collection of papers composing this thesis takes a detailed look at three typical forms of trust – including individual trust, institutional trust, and technology trust – engaged with these systems. On the positive side, China's Social Credit System, similar to other reputation-based platforms and credit reporting agencies, can function as an intermediary that provides information in certain areas for participants to counteract unacquaintance and facilitate (or impede) trusted interactions between distantly connected individuals. Digital contact tracing technologies that are endorsed by a complex of social mechanisms, such as laws and regulations, provide a less intrusive way for citizens to cooperate with public policies and make a collective effort to fight against the pandemic. Taking a step further, blockchain applications foster trusted interactions without the intervention of third-party authorities, reducing the risk and cost engaged with traditional institutions while retaining the value of trust for the trustor.

The thesis argues that trust-inviting systems essentially attempt to interpret, translate, and ultimately institutionalize the idea of trustworthiness as a desirable property required for reliable interactions in different contexts. However, the way that trust-inviting systems are currently using to institutionalize the characteristics of trustworthy persons, institutions, and technologies should not be accepted without scrutiny. For each case studied here, a discrepancy is shown between the intention to improve trust and trustworthiness and the means that are adopted to facilitate them. Such cleavage is argued to be primarily caused by flawed understanding of the trust concepts and the resulting ill-suited design applied to given contexts, as

well as problems that emerge from the implementation process. In an effort to ameliorate these issues, it is proposed that a recalibration of the trust concepts can contribute to remedying shortcomings of the current design and development of the systems and providing forward-looking strategies that help shape trust in a more socially and morally desirable way. In a word, it is argued that trust-inviting digital systems should be designed, developed, and deployed in ways that are aligned with the essence of the trust relation in context, in order to achieve proper trust and trustworthy systems. By doing so, the pitfalls identified in each case are used as perspectives from which affordances that foster warranted trust could be built and affordances that would undermine secure trust could be foreclosed accordingly.

The first section has introduced the relevant research questions, motivations, and the analytic approach adopted for analysis throughout this dissertation. The context-sensitive approach proposed argues that digital systems should be designed, developed, and deployed in ways that are aligned with the essence of the trust relation in context. To achieve the research goal of making trust towards individuals, institutions, and technologies in the digital age more justified and well-grounded, this dissertation seeks to clarify the conceptual and practical muddles pivoting around trust and trustworthiness in specific cases and close the gap between the current understanding of the concepts and the conceptualizations needed for remedying current flaws and achieving proper trust.

In line with the interest in investigating how interpersonal trust shapes, and is shaped by digital systems, chapter 2 takes China's SCS as the case, critically examining whether this project can achieve its overarching goal of fostering moral trust between citizens via its current implementations, as well as some logic behind reputation systems in general. To this end, this chapter provides a close ethical reflection on the normative assumptions of trust and trustworthiness made implicitly by the initiatives of the SCS, together with a comparison study of three pilot cities' scoring systems. It argues that the underlying conceptions of trust and trustworthiness assumed by current initiatives can foster trust relations primarily in an instrumental and prudential sense. As a result, a discrepancy is shown between the moral objective of the overall project and the current ways of approaching it. To help address the conceptual and practical issues involved and promote trustworthiness appropriately, well-designed and -audited systems should be seen as a precondition, and thus, a coherent framework for guiding how the trust concepts should be understood and implemented at both the national and local level is urgently required. This makes clear the need for institutions to design trustworthy systems before talking about promoting citizens' trustworthiness in terms of the systems' rules.

At the first sight, the role played by institutions behind the SCS is an intermediary that curates information of participants to direct interactions, just like other reputation-based platforms. However, as a governance approach that contains great comprehensiveness and invasiveness, the SCS is unique and the role of agencies behind it is argued to be not just an intermediary but an indirect and ultimate

trustor to whom citizens should be or appear trustworthy. Thus, although the SCS is mainly used as a representative case to explore how technological systems impact trust between individuals, the complexity related to trust in institutions and the institutions' role as an implicit trustor should not be overlooked.

Chapter 3 takes a closer look at technology-mediated institutional trust with the case of digital contact tracing technologies developed for mitigating the pandemic. It argues that the deficit of institutional trust with respect to personal data protection should be understood as part of the privacy issues over contact tracing apps, and that proper implementation of this privacy-sensitive digital solution should be underpinned not just by legal and technological measures for preserving privacy but also by the trustworthiness of institutions and citizens' proper trust towards institutions. This is because the legal and technological measures adopted may unexpectedly economize on trust but are not in themselves sufficient to encourage the adoption of this digital solution. Considering the relatively complementary features of the trust-based approach and alternatives such as data-protection laws and privacy-sensitive design, a combined strategy is proposed to be closer to what we expect from responsible design and development of digital contact tracing technologies. Additionally, it should be noted that this statement does not indicate that legal and technological approaches are present as solely distrusting strategies, nor does it indicate that these measures cannot contribute to impacting trust and the uptake of the apps. Rather, this statement is meant to emphasize the shortcomings and insufficiency of these approaches, as well as the crucial role played by institutional trust in supporting any policy responses to the public health crisis.

The observations of the relationship between institutional trust and digital contact tracing technologies show the value of trust together with people's concern over bureaucratic structures and the goodwill of power holders. Technology trust in the context of blockchain studied in chapter 4 is essentially an effort to explore technology-enabled trusted interactions that do not depend upon authoritative, traditional institutions such as governments and corporations. The comprehensive analysis of blockchain trust provided contributes to debates on whether blockchain technology is trust-free or trusted by clarifying who the trustor group is, where they place their trust in, and why it is plausible to talk about blockchain trust at all. Based on a reflection on two of the most promising values that can invite users' trust in blockchain technology – namely, decentralization and transparency, this chapter argues that users' trust built on these values is risky and unjustifiable due to the moral and technical limits engaged with how blockchain technology is currently implemented.

Following the rough idea (proposed in chapter 4) that blockchain trust can be understood in a normatively laden way similar to trust towards institutions and business, chapter 5 provides a constructive reflection on what people may expect from specific blockchains and how such institution-like entitles should be assessed. Due to blockchains' double-edged properties, blockchain trust is characterized, on the one hand, as a form of trust grounded in- and going beyond institutional trust.

By building trust-inviting elements into, rather than outside, this technology's basic infrastructure, the original design of blockchain touches the most intriguing aspect of trust as well as the central question of trust discussed at the beginning of this dissertation. Namely, we want our trust to be warranted, more than ever, to dispel our anxieties and worries about the discretionary power possessed by others (including both the trustee and third parties) with the hope that the vulnerabilities and risks engendered by placing trust can be minimized to the greatest extent possible. On the other hand, it is undeniable that blockchain applications that crowd out the pivotal role played by traditional institutions and a whole array of responsible representatives are facing challenges to their actual trustworthiness for functioning as an institution-like entity. This chapter ends by proposing that such limits could be ameliorated by shifting the responsibility to a network of peers, developers, and normal users that are directly associated with blockchain applications.

Acknowledgements

Like many other students, my PhD journey started by a long-haul flight, with feelings of excitement and passion, as well as nervousness, fear, and anxiety. Something perhaps a little bit different from others was that I also brought an electric rice cooker in my backpack, a typically foodie and nostalgist. But fortunately, the trajectory of my PhD is never lonely, as I have met many amazing people who have inspired me when I get stuck in the path of being an independent researcher and people who, like families, have brought great happiness to my life and provided support after I get a health issue later this doctoral journey. I treasure this opportunity to express my sincerest gratitude to these very important people who have raised me up along this fantastic journey.

My first thank goes to my supervisors. Mark, thank you so much for your countless feedback, support, patience, and encouragement. During our meetings in your office throughout the years, with the fine coffee you provided, you have shown me how to be a critical researcher and a warm and generous person at the same time. I appreciate the topic that you helped me draft during our first meeting, the trust you placed in me to do my research independently, and the timely assistance you provided whenever it's needed. I feel lucky to have you as the thick piece of rope alongside my climbing, and I wish you all the best in your new life in Australia. I also would like to express my great gratitude to Filippo, who became my supervisor after Mark's resignation from our university. I am very grateful for your most kind guidance and all the invaluable feedback and edits you gave whenever I sought your help. Sincerely, I am very happy to have you as my last year's supervisor. Hope we could have a chance to collaborate on joint work in the future, and also, I have to say that I really love your humor. Lastly, I want to thank my promotor Jeroen, who made possible my position in this university and gave me great freedom to stick to my interests. The opportunity you kindly provided has indeed changed my life, opening the door of this amazing academic adventure and my career. I appreciate all the fruitful discussions we had and the comments and criticism you provided.

My appreciation also goes to members of my defense committee – WANG Guoyu, Philip Nickel, Ibo van de Poel, Johan Pouwelse, and Nitesh Bharosa. I am honored and grateful that you are interested in my research and would like to participate in this defense. Then, a special thank you to Zachary Pirtle and Ronald Oosting. Each of them helped one article of mine with detailed comments and edits. Thank you, and I will always keep your encouragement and generous support in mind.

I would also like to thank the wonderful people I met in the Philosophy Section

here. Anna, my dear friend, I was fortunate to know you and pleasant to have a lot of ramen and fun with you. Though we are persons with quite distinct hobbies and characters, there is something that can always resonate between us. I wish you could realize your dream one day and we will, of course, have many road trips together. Giulio and Thijs, who become equal parts officemates and friends that brought plenty of happiness and pleasure to my time in our room. Giulio, I have to say that I learned a lot of technical knowledge from you, including cellphones, smartwatches, headphones, coffee machines, monitors, etc. Shuhong, I appreciate all the great talks and fun we had together, as well as Lavinia who had a nice bike trip with me. Also, I hope to thank many colleagues who have provided feedback to my research on different occasions, including Sabine, Casper, Frances, Pieter, Emily, Juan, Nicole, Birna, Taylor, Zhijie, Ximeng, and more. Next, I would like to extend my gratitude to Nathalie, who immediately welcomed me as a new member of our section. Together with Diana, you were always on call and supported me in handling all the complicated procedures involved in doing a PhD here. Thank you so much. Finally, I would also like to thank the peers I met in our faculty, including Bing, Arie, Amir, Ali, Ahmad, and Kartika. We had shared many interesting meetings throughout the four years, and I cherish all the feedback you provided to my presentations.

And of course, I am thankful for all my friends outside my academic path, who made my life incredible and this small, lovely town a second home for me. Raissa, Zhenwu, Kaiyi, Yan, Anton, and Sitong, thanks for coming to my life and always being there by my side, and most importantly, thanks for your tolerance of my careless cooking and the resulting dark cuisines. A very special thank goes to Jing, who shared many important moments with me and helped me countless times with my writing. Thank you for your trust and always support, and the excellent food in particular. A sincere thank also goes to my extraordinary neighbors: Hongjuan, Dadi, and Ling, and friends who often enjoyed great drinks with me: Yuxin, Pan, Guowei. Hereby I would like to thank you all for the nice and relaxing experiences we shared together.

Last but not least, I hope to thank my parents for their unconditional love and considerations. And my grandpa, sister, aunt, and uncle's continued support for me to pursue what I think it's the right to pursue. You taught me how to be an optimistic, upright, and dedicated person with peace and love in mind. Without this lovely family, I will never have a chance to be the person I want.

Yan Teng
May 2021

About the author

Yan Teng (1991) was born in Dalian, Liaoning, China. She got a Bachelor of Philosophy (with honors) from Dalian University of Technology, 2012, and at the same place, she received a Master in Ethics (with honors) in 2016. Between September 2016 and December 2021, she completed her PhD in Ethics of Technology at Delft University of Technology with the funding from China Scholarship Council. Her PhD research focused on philosophy and ethics of trust in the context of technological systems where issues of trust are particularly relevant. In particular, she contributed to providing detailed normative reflections on trust relations engaged with blockchain technology, reputation-based systems, and COVID-19 tracing technologies.

List of Publications

6. **Teng. Y**. (2021) Towards trustworthy blockchains: normative reflections on blockchain-enabled virtual institutions. *Ethics and Information Technology*.

5. **Teng, Y.**, & and Song, Y. (under review). Beyond legislation and technological design: the importance and implications of institutional trust for privacy issues of digital contact tracing. *Science and Engineering Ethics.*

4. **Teng, Y.**, & Alfano, M. (under review). Can Social Credit System promote social trust and trustworthiness? *Philosophy & Technology.*

3. **Teng, Y.** (under review). What does it mean to trust in blockchain technology? *Metaphilosophy.*

2. **Teng, Y**. (under review). Warranted trust in the context of trust-inviting systems. *Ethics and Information Technology*.

1. **Teng, Y.,** Zhang, J., Wang, G, Y. (2015) Public Assessment of Benefits and Risks of Nanotechnology. *Deliberate safety and ethical issues of nanotechnology* (ISBN 978-7-03-044258-1), p242-251. Beijing: Science Press.