# Mediated trust: A theoretical framework to address the trustworthiness of technological trust mediators

Bodó, B.

DOI
10.1177/1461444820939922

Publication date
2021

Document Version
Final published version

Published in
New Media & Society

License
CC BY

*Article*

new media & society
1–23

DOI: 10.1177/1461444820939922
journals.sagepub.com/home/nms


# Mediated trust: A theoretical framework to address the trustworthiness of technological trust mediators

**Balázs Bodó**
University of Amsterdam, The Netherlands

## Abstract

This article considers the impact of digital technologies on the interpersonal and institutional logics of trust production. It introduces the new theoretical concept of technology-mediated trust to analyze the role of complex techno-social assemblages in trust production and distrust management. The first part of the article argues that globalization and digitalization have unleashed a crisis of trust, as traditional institutional and interpersonal logics are not attuned to deal with the risks introduced by the prevalence of digital technologies. In the second part, the article describes how digital intermediation has transformed the traditional logics of interpersonal and institutional trust formation and created new trust-mediating services. Finally, the article asks as follows: why should we trust these technological trust mediators? The conclusion is that at best, it is impossible to establish the trustworthiness of trust mediators, and that at worst, we have no reason to trust them.

## Keywords

Institutional trust, interpersonal trust, online intermediation, regulation, trust

## Introduction

"Trust and technology" promises to be one of the major themes of technology policy discussions in the coming years. While early on the question of trust emerged in the

**Corresponding author:**
Balázs Bodó, Blockchain & Society Policy Research Lab, Institute for Information Law (IViR), University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands.
Email: bodo@uva.nl

contexts of online anonymity, reputation, or e-commerce, at the beginning of 2020s new issues are at stake. The prevalence of online misinformation, for example, has raised questions about the trustworthiness of digital news distributors. The weaponization of Internet services is forcing us to consider what the prerequisites of a trustworthy digital environment might be. In policy debates on platform regulation, some are questioning whether traditional regulatory frameworks can be trusted to address the challenges we face (Suzor, 2019; Van Dijck et al., 2019). Service providers routinely breach users' trust by exposing their personal information when these providers get hacked, or by selling their data to third parties contrary to users' expectations. Automated decision-making systems are being used to make increasingly consequential choices, raising the question of whether they can be trusted to act in fair, just, and transparent ways, or more generally in the interests of their users (Pasquale, 2015). Partly in response, blockchain technologies promise to replace seemingly untrustworthy intermediaries with a technological system designed to minimize the need for trust.

All of these issues highlight the fact that digital technologies[1] shape *how humans trust each other*, and that in order to fulfill this task, they need to be trustworthy. The ultimate question that we need to address is thus as follows: what can we say about the trustworthiness of the technological tools that are used to produce trust?

Trust on an interpersonal level can be defined as the willingness to cooperate with another in the face of uncertainty, contingency, risk,[2] and potential harm. Three major transformations, however, are changing the nature of trust in the information society.

First, globalization and digitization have created new ways in which human societies produce and distribute risk, and have introduced new forms of incalculable uncertainties into everyday life (Beck, 1992). These new risks necessitate new approaches to producing trust and managing distrust. We need forms of trust that operate at the scale of the planetary networks we are embedded in, and that match the global challenges we face. Traditional, systemic, and institutional guarantors of trustworthiness, such as governments and expert systems, still tend to function within the limits of the nation state. Their apparent inability to deal with issues beyond their scale and the falling levels of confidence they face are mutually reinforcing.

Second, our interpersonal relations are increasingly being mediated by digital technologies. Existing institutions are also becoming more reliant on new technologies to fulfill established roles. This mediation is inevitably transforming the nature of the trust that emerges in these contexts. As we use digital technologies to mediate interactions that require or produce trust, *these technologies turn into trust mediators.* Third, partly in response to these developments, new technological forms of trust production have rapidly gained prominence. Consequently, we need to recognize the trust-mediating capacities of the digital technologies that we use, and to assess their trustworthiness from this perspective.

At present, there is no overarching theoretical framework that incorporates all of the relevant dimensions of trust and digital technologies. All of the major works in sociology on interpersonal and institutional trust and risk, such as Beck (1992), Giddens (1990), Fukuyama (1995), Misztal (1996), Sztompka (1999), and Hardin (2002), predated the mass rise of the Internet and failed to anticipate the nature and prominence of today's digital technology. There is a need to revisit the assumptions underlying the discussion of trust

and technology in these accounts. In 1997, for example, Fukuyama was still discussing trust in the context of the nation state. Beck (1992) addressed risk in the context of industrial production, and focused on physical threats such as radiation or chemical poisoning (p. 21). Giddens (1990) did not consider the role of digital technology in the organization of abstract systems of trust production. By contrast, this article starts from the assumption that new, technology-specific challenges have appeared in addition to those discussed at the turn of the millennium, such as globalization and environmental degradation. The rapid proliferation of information technologies is forcing us to consider whether our current theoretical frameworks capture the process in which the trust-producing institutional arrangements and power relations of the industrial era are under the constant assault of digital disruption?

This article is the first step in a long-term project to rethink the social theory of technological trust mediation. It addresses the following questions: (1) How do digital technologies establish new forms of interpersonal and institutional trust? (2) How do digital technologies transform the existing logics of interpersonal and institutional trust? (3) How can we establish the trustworthiness of trust-mediating technologies?

In the canonical models of trust formation, such (Giddens, 1990; Mayer et al., 1995; McKnight et al., 2011; Misztal, 1996) trust is described to have a number of different components. First, trust depends on the personal characteristics of the trustor, such as their propensity to trust, something which I do not discuss in this article. Second, trust is a factor of the perceived or actual characteristics of the trustee, such as its trustworthiness. In turn, the trustworthiness of trustees is a combination of internal factors, such as their ability to fulfill the expectations they face, and external conditions, such as the institutional structures, which envelop both trustor and trustee, and which, as I lay out in the rest of this article, provide structural assurances and situational normality, produce common knowledge and shared expectations, and provide various forms of safeguards and guarantees against uncertainties. This article focuses in this latter dimension: the institutional frameworks, which produce trust and mitigate distrust by signaling, shaping, and defining the trustworthiness of the unknown other whom one needs to trust. In interpersonal relations, this institutional framework is increasingly technological. If, on the other hand, the trustee is the trust-producing technology itself, I argue that the institutional framework is incomplete, which may have serious consequences on the interpersonal trust that emerges in technological settings.

## Institutions of trust production and distrust management

Human societies have developed a number of ways to develop trust and manage distrust. Trust in interpersonal relations emerges through collective social practices such as habits, rituals, memories, and reputation (Misztal, 1996). In closely knit social groups, these practices take place through physical co-presence, and structured by repeated, often ritualized interactions and communal activities. Under the conditions of modernity, however, trust also needs to operate across larger social, temporal, and geographical distances, whereby sophisticated institutional arrangements emerge to produce trust and mitigate distrust among strangers (Luhmann, 2017). In the following, I will focus on institutional forms of trust production, and I will only discuss interpersonal trust if it has an institutional component.

There are multiple logics of institutional trust production. For instance, trust is produced through the creation of familiarity and shared sets of knowledge. Abstract systems and institutions, such as churches, newspapers, civic organizations, firms, professional associations, and other forms of bureaucratic organizations, create shared spaces of common knowledge, interpretative frames, signal sets, and coding rules to re-create the familiarity of interpersonal relations on a larger scale (Zucker, 1985). To paraphrase Sztompka (1999), such contexts create trust by providing normative certainty, transparent social organization, and a stable social order. In addition to normative clarity, they provide procedural safeguards and enforcement mechanisms. They maintain systems of accountability, enact rights and obligations, enforce duties, and safeguard dignity and integrity. The internal rules, norms, and governance mechanisms of such organizations also establish their trustworthiness for outsiders.

The gaps between these different local contexts are bridged by the sophistication and extension of institutions of governmentality (Foucault, 1991) and the social overhead sector (Zucker, 1985). The main role of the latter is strategically to manage distrust of institutional trust producers. As Shapiro (1987) and Sztompka (1999) observed, generalized trust in Western democratic and economic systems is produced through the strategic management and the institutionalization of distrust of the individual institutional constituents. Distrust in this context, which I also follow in this article, is not simply the lack of trust (Cofta, 2006; Hardin, 2004; Marsh and Dibben, 2005) which would result in non-cooperation. There are many situations, in which one cannot avoid relying on an actor with unknown or questionable trustworthiness. Distrust in such cases becomes a strategy to minimize the risks that come with the engagement with an untrustworthy other. Achieving trust through distrust entails establishing systems of accountability, checks and balances, oversight and supervision, backup systems, and insurance, which disincentivize, detect, punish, and remedy the breach of trust by these institutions. Markets, for instance, aggregate and reveal information, price risks, and provide insurance. Legal services create contractual frameworks of control to produce confidence (Shapiro, 1987). Laws and other regulatory mechanisms define norms, sanctions, and institutions of oversight. States establish frameworks of control, monitoring, dispute resolution, sanctions, and enforcement through various arms of government, while the latter also oversee and limit each other's power.[3] Banking provides the infrastructure for the circulation of credit, the universal symbolic token (Giddens, 1990). Ultimately, these systems form a complex, mutually interdependent institutional network of checks and balances, whereby trust is produced through internal governance mechanisms, external control, and the division of power.

These different logics produce a spectrum of trust. At one end of this spectrum, risks, contingencies, or even fate (or fortune) is managed through *faith*: a non-cognitive, non-verifiable belief in some positive outcome (Simmel and Frisby, 2004: 175). At the other end, *confidence* is rooted in knowledge in the face of uncertainty. Confidence relies on the rational, calculative assessment of risk, and the ability and agency of the parties involved to reduce vulnerability. While in theory, trust may lie somewhere in between the two (Seligman, 2000), in practice the different components of trust continuously recombine, substitute for, and complement each other.

## Challenges of trust at scale: globalization and digitization

Different logics of trust operate on different scales: interpersonal trust works on a kinship scale; abstract systems enable cooperation across larger social, economic, and cultural distances. Late modern Western systems of trust production and distrust management mostly operate on the scale of nation states. At present, however, an increasing number of trust-dependent activities are taking place beyond the action radius of current institutional trust frameworks. These changes are the result of two forces: globalization and the rapid proliferation of digital technologies.

*Globalization.* The ancient trade networks that carried tin, silk, spices, and slaves had their own logics of trust production, such as risk-sharing arrangements or informal reputation networks. It was only from the 19th century that global trade systems coalesced into a truly international system, in which a densely woven network of transnational corporations, nation states, and intergovernmental and nongovernmental organizations facilitated trust-requiring economic, political, and social relationships (Buzan and Lawson, 2015). Post-WWII globalization intensified the interdependence of highly different local domains connected by planetary-scale frameworks of finance, production, commerce, telecommunications, media, and governance. These transformations forced a change in the logics of trust production, mostly through the transformation in the scale and nature of the risks produced in and by these frameworks.

Beck (1992) argues that modernization, industrialization, and technological development have introduced new hazards, risks, and insecurities to late modern society. They cannot be understood using existing shared knowledge; they resist established logics of assurance; they fall outside of existing interpretative frames; and they create an unfamiliar world for individuals and institutional actors alike. Global mobility and the global infosphere, for example, increase local cultural heterogeneity; local knowledge and underlying expectations are contrasted against and destabilized by new, unfamiliar knowledge and expectations from across the globe. Economic interdependence exposes local economies to the iron rules and competition of global supply chains and financial and labor markets, resulting in what is often the radical transformation of local economic and social conditions. Local societies also face new, global challenges, such as pandemics, ecological degradation, or mass human migration, which do not respect the boundaries of the nation state. With each such transformation, new institutional logics are needed to address the new forms of risks and the corresponding crises of trust (Beck 1992; Shapiro, 1987; Zucker, 1985).

From this perspective, the majority of post-WWII institutional developments aimed to establish supranational frameworks to produce trust on a global scale. The development of international standards (from weights and measures to telecommunications) and global monitoring and response networks (such as the World Health Organization's Global Outbreak Alert and Response Network); the rapid development of supranational political coordination and jurisdiction (such as the creation of the United Nations and its institutions, and the birth, enlargement, and slow federalization of the European Union); the rapidly growing network of trade agreements (including the World Trade Organization and a number of regional agreements); and the international, mostly private governance

of global production networks and value chains (Gereffi and Korzeniewicz, 1994) offer hard institutional frameworks that produce trust and manage distrust. They create new background and contextual expectations, standards, rules, enforcement mechanisms, organizational structures, procedures, and safeguards at the supranational level. In parallel to the growth of supranational institutional frameworks, empirical studies documented an erosion of trust in various national institutions, suggesting that trust producers at the national level may not be perceived as able to manage risk in a globalized world (Catterberg and Moreno, 2005; Crozier et al., 1975; Inglehart, 1999; Nye et al., 1997; Putnam, 2001; Sztompka, 1999). In response to this perceived crisis of trust, populist and extremist political actors, fringe ideologies, and belief systems stepped into the vacuum by providing a (usually false) sense of ontological security (Giddens, 1990) in a complex and unpredictable world, by identifying scapegoats, giving overly simplistic explanations, and a sense of agency.

*Digitization.* The rise of the Internet intensified and magnified the aforementioned challenges. First, digital technologies permeated every possible locus of interpersonal and institutional trust production. Technology is now used for functions that range from negotiating sex with strangers to predictively policing populations. These technologies introduce completely new and unknown forms of risk to interpersonal and institutional trust relationships.

Second, while the process of institutional trust production is traditionally embedded in interdependent legal, political, economic, social, and cultural milieus; institutional frameworks; and regulatory structures, trust-producing digital technologies do not form part of these carefully designed distrust management frameworks. The reason for this is twofold: first, many technologies consciously resist existing institutions of regulation and control (Yeung, 2019), and second, it is often unclear whether existing institutional trust-production logics are prepared to incorporate planetary-scale technology networks.

Ultimately, we are facing new contexts in which trust and distrust need to be addressed, and we are also able to produce new, technology-based forms of trust. Digital technologies are thus part of both the problem and the possible solution. To describe these parallel developments in the demand for and production of trust, in the following section, I introduce the concept of *mediated trust*.

## Defining mediated trust

The concept of *technology-mediated trust*, or simply *mediated trust*, is currently lacking in the literature on trust. Recent studies of the intersection of trust and technology, such as Botsman (2017), Keymolen (2016), and Werbach (2018), tend to focus on the ways in which digital technologies produce interpersonal trust. The notion of trust mediation, however, looks beyond the perspective of the individual user who relies on digital intermediaries to trust strangers, and incorporates two additional dimensions. First, the concept of mediated trust is used to address cases where digital technology indirectly transforms the established logics of trust production by (re)mediating those interactions where trust traditionally emerges. Second, mediated trust also considers digital

technologies' trust-mediating role in the light of their political economy, power, and the institutional context. Before going further, however, let us step back and consider the definitions of "trust" and "technology" in more detail.

I consider trust to be a basic fact of social life that enables humans to cooperate with each other, despite the inherent uncertainties and risks such cooperation entails (Fukuyama, 1995; Keymolen, 2016; Misztal, 1996: 26–27). Different disciplines, such as psychology, economics, philosophy, or computer science, focus on different aspects of this definition. For psychologists, trust is a mental state that prompts a trustor to accept vulnerability vis-à-vis a trustee (Walker and Ostrom, 2003). In philosophy, scholars have tried to differentiate different forms of trust, from faith to confidence (Keymolen, 2016; Seligman, 2000). In sociology, substantial theoretical work has been done on the conditions of interpersonal trust (Misztal, 1996); the non-personal, institutional aspects of trust (Giddens, 1990; Sztompka, 1999); and the various social and economic consequences of more or less trust in society (Fukuyama, 1995; Putnam, 2001). In addition, much empirical work has been done on trying to quantify and measure trust in society, mainly focusing on democratic institutions, but also on other institutional actors (such as businesses, scientists, and the media) that play a crucial role in democratic societies (see citations at the discussion of the empirical studies on the erosion trust in the context of globalization). For economists, trust manifests itself as a microeconomic problem that is suited to game-theoretical modeling, or as a phenomenon in organizational theory, with macroeconomic and institutional relevance (Shapiro, 1987; Zucker, 1985). Computer scientists have also discussed trust in various theoretical contexts, such as the dependability of software systems (Clarke, 2006), the control and safety of critical systems, and security (Nissenbaum, 2001; Schneier, 2012), while the rise of e-commerce and e-government services in the early 2000s prompted intense empirical research into user trust vis-à-vis technical systems (Corritore et al., 2003; McKnight et al., 2002, 2011; Nixon and Terzis, 2003; Söllner et al., 2016; Tang and Liu, 2015).

The concept of *mediated trust* incorporates elements of these different disciplinary approaches to focus on how digital technologies establish new logics of trust production and change pre-existing ones. For the purposes of this article, I define digital technologies as including data, software, networks, machines, protocols, and standards; economic and political structures; the social and cultural practices that emerge around them; and the institutional and organizational forms that they assume. Due to the deliberately vague contours of this definition, it covers not only the directly visible aspects of technology, in the form of products, services, objects, and interfaces, but also the hidden, invisible, human, and organizational components of these heterogeneous technological assemblages, such as complex institutional frameworks and the processes of research, development, production, finance, and logistics that produce and maintain them (Bijker et al., 1987; Crawford and Joler, 2018; Gillespie, 2010; Gürses and van Hoboken, 2018; Plantin et al., 2018).[4]

The analysis is based on an archelogy of the contemporary (Graves-Brown et al., 2013): a close reading of actual trust-mediating services, their use, the controversies that surround them, and the firms that control them. These elements have been omitted from this article, however, as the specific controversies, configurations, and practices are ultimately ephemeral. I focus instead on the structural elements that can be identified by studying these ephemeral cases.

The concept of mediated trust covers multiple dimensions in which trust and technology interact. First, I discuss *trust produced by technology*, to consider the fact that complex technological systems permeate and affect both the most intimate interpersonal and the most robust institutional precursors of trust. Second, I examine the issue of why we have *trust in technology*.

## *Trust* by *technology*

Trust can be produced in small-scale interpersonal settings, and by larger-scale abstract institutions. Digital technologies have an impact in both domains. In the first case, interpersonal trust is established through physical, temporal, social, cultural, economic, and familial proximity. Shared habits, rituals, reputations, and shared memories create a form of social stability based on familiarity, predictability, reliability, and legibility in social relations.

On the other hand, modern societies need to establish trust beyond the boundaries of interpersonal relationships, and thereby need impersonal institutional arrangements that span wider social, economic, cultural, temporal, and physical distances. Giddens suggests that trust in such disembedded social relations is mediated through symbolic tokens or expert systems. The former refers to "media of interchange which can be 'passed around' without regard to the specific characteristics of individuals or groups that handle them at any particular juncture" (Giddens, 1990: 22), such as reputation, political legitimacy, or money. Expert systems are complex institutional arrangements that create reasonable expectations about the operation of such systems and the results they produce, through sets of rules, supervisory and enforcement mechanisms. The legal system, the medical profession, journalism, science, and government are examples of expert systems that enable the emergence of trust among strangers.

Digital technologies enter these spaces of trust production both directly and indirectly. Some technologies, such as platforms, marketplaces, and resource-sharing services emerge as institutional trust-producers. They also have an indirect impact on the established logics of trust production, as they remediate the loci in which interpersonal trust emerges.

The use of digital technologies also has an impact on the trust produced by traditional expert systems. When healthcare institutions use machine learning-based diagnostic tools, or when law enforcement relies on predictive policing software, their core operations are transformed. The use of digital technologies by private and public entities creates new uncertainties, conflicts of interest, and modes of operation; it restructures values and ethics. All of these affect these abstract systems' trustworthiness and their ability to produce trust. In the following section, I take a deeper look at how we directly and indirectly trust *by* technology, first at the interpersonal level of trust production, and then at the institutional level.

### *Indirect and direct interpersonal trust production by digital technologies*

*Indirect effects of digital technologies on interpersonal trust.* Digital technologies mediate representations of the self: the ways we perceive and make ourselves legible to, and perceive, others. When we rely on digital communication tools to tell others who we are, or learn about the world, we mediate our experiences through a combination of objects,

interfaces, and software, created by concrete, often commercial entities that subject us to their particular rules, priorities, shortcomings, business models, and politics.

We may use messaging apps to keep in touch with our closest family members, but this exposes us to what are often incalculable risks, such as law enforcement agencies or unknown subcontractors also listening in to these exchanges (Ehrenkranz, 2019; Gartenberg, 2019). We may see others demonstrating certain habits online, without being aware of the role of technology in shaping these habits. The compulsive posting of heavily photo-shopped selfies on social media may be a genuinely harmless pastime, or a result of unforeseen pathologies[5] or carefully calculated nudges. If Zuboff's (2019) prediction is correct, such behavior will continue to multiply as capitalism moves beyond surveillance and prediction, toward shaping consumer decisions and forming new habits. As we are also increasingly reliant on digital services to recall the past, we will have to accept that our memories are controlled by intermediaries with their own particular interests, pressures, and priorities. The European Right to be Forgotten, for instance, enables European citizens to delist certain information from search engines, based on data protection and privacy rules (Van Hoboken, 2013). The identity, integrity, and reputation of individuals depend on whether the social memory is mediated digitally, orally, or on paper.

When we base our interpersonal trust on the digital representation of the other, we have to allow for how this other is filtered by the act of mediation. Some of these filters are literal and alter the person's appearance. Other filters are more symbolic, but they are no less consequential. Our traditional interpersonal logics of trust need to take account of the fact that the raw material of interpersonal trust has changed. The trust-related affordances of digital technologies are also arranged into technological assemblages that aim to directly produce trust.

*Directly produced interpersonal trust.* Botsman (2017) and Keymolen (2016) have extensively discussed the emergence of new digital services that facilitate trust-dependent economic interactions among strangers on a global scale. Sharing platforms such as Airbnb or Uber and e-commerce sites such as eBay, social media, and dating apps produce trust between buyers and sellers, information sources and information consumers, service providers, customers, and potential sex-partners who need trust to engage. Most of these services render their users legible through the aggregation of a mix of signals optimized toward trust: records of past actions, profile pictures, social networks and institutional affiliations, reviews, and reviews of reviews. The source of trust is knowledge about the users' reputation: perceived capacity to fulfill a specific task, and certain signals about their personal benevolence and integrity. The technological trust producer collects, aggregates, and makes available data about their users in these dimensions, but does not necessarily verify them in any meaningful or systematic, transparent, and accountable way (Houser, 2020; Kerr, 2019). Other systems, such as blockchains, aim to minimize the need to trust the other by limiting the actions of their users and removing the option of non-compliance through their design (Werbach, 2018). We may not be able to rely on the reputation of an anonymous user, but we may have some confidence in the technology to force the counterparty to respect a set of pre-defined rules.

The trust produced by these technologies is interpersonal, but a contextual one. The trust mediators' aim is to merge invisibly into the background. While technology-mediated trust

cannot be separated from the trustworthiness of trust mediators, the latter have little incentive to take responsibility for any breach of trust in the interactions they structure. Users are expected to conduct their own due diligence based on information provided by the trust mediator. If they make a bad call, they may be left with little or no recourse. While users of e-commerce and sharing platforms are protected from certain, obvious cases of fraud, for example, they have limited recourse against the failure of a trust mediator to verify the trustworthiness of their users and proactively remove "bad apples" across the board.

Technology has substantial agency in both directly and indirectly produced trust. This agency forces us to consider the trustworthiness of trust mediators. Humans ultimately learn to reflect upon and interpret how the media affects the nature of conveyed information (McLuhan, 1964; Ong and Hartley, 2012). People living under oppressive political regimes learn how to gather information from untrustworthy government-controlled newspapers and how to express themselves on phone lines and in homes that are potentially tapped. A similar awareness of the agency of technology and its impact on the nature of interpersonal trust is yet to fully develop. Too often, we fail to interrogate the agency of trust technologies and allow them to sink into the background.

*Institutional trust production and digital technologies.* Digital technologies also impact the institutional logics that manage and disperse distrust across larger cultural, geographic, economic, and social distances. Institutional trust producers strengthen confidence in society by defining the governance of a social or economic domain. They set and enforce the rules, monitor compliance, and resolve conflicts in a transparent, accountable, and predictable manner (Sztompka, 1999). Instruments of control also build confidence: contracts spell out the rules that have been agreed upon by independent parties. Other forms of institutional trust include monitoring and oversight mechanisms to aggregate and disseminate information about the behavior of other actors. Markets, for example, aggregate information in prices. Yet, another form of institutional trust mitigates the negative consequences of a breach of trust. Insurance, for example, builds confidence by compensating harm. Some of these institutional trust production mechanisms are private and/or operate in a decentralized manner, but to some extent, they are all embedded in and subject to the legislative, enforcement, and adjudicative powers of the state.

Digital technologies have an impact on these institutional trust production logics. They also enable the emergence of new, technological forms of institutional trust production.

*Indirect institutional effects of technological trust production.* Traditional trust-producing institutions increasingly rely on digital technologies to fulfill their roles. Some of these technologies support humans by providing more accurate, up-to-date, detailed information, or by suggesting and ranking decision alternatives. Other technologies are intended to replace humans altogether, because they are expected to make fewer mistakes, or be more efficient. Both these developments affect the trustworthiness of traditional institutions of trust production.

Insurance companies may scrape public social media profiles for signs of licentious behavior, for example, and set premiums accordingly. Police departments deploy software to predict crime and allocate resources. Public agencies in the fields of welfare and

healthcare have installed technological systems to counter perceived bias, corruption, simple errors, or low-quality decisions. Sometimes, humans are removed from the decision-making process altogether: news items are recommended by invisible or incomprehensible machine learning models, rather than journalists (Bodó, 2019), and Amazon employment contracts are sometimes terminated automatically.

These transformations are bringing new risks, changing the nature of the trust produced by these institutions, and impacting the latter's trustworthiness. Even if technological systems do no more than formalize and encode internal rules and procedures into technological decision-support systems, this already removes a considerable amount of human agency and discretion from the system. Bureaucratic systems may become blind machines with little or no human capacity to adjust rules to local circumstances, recognize and apply exceptions, or interpret rules when necessary (Eubanks, 2017). By contrast, the machine learning systems that are deployed in predictive policing, for example, or as news recommenders, are black boxes that produce probabilistic outcomes (Ananny, 2019). The lack of transparency of the process that converts input data into decisions carries the risk of potential algorithmic bias, discrimination, and unaccountability (Zuiderveen Borgesius, 2019). The probabilistic nature of the outcome reduces control and introduces uncertainty. There are clear conflicts between both of these elements and notions of trust and trustworthiness. If and when institutions become slot machines where non-transparent automation produces probabilistic outcomes, confidence in them will inevitably fall, and this, in turn, will necessitate new institutional arrangements to deal with the growing distrust.

*New technologies of institutional trust production.* Digital technologies have also created new logics of institutional trust production. The need for trust in digitally mediated interactions produced a plethora of technical solutions ranging from various cybersecurity tools, via online reputation systems, to technological transparency and accountability frameworks, such as open source code, or decentralized architectures. As they matured, these approaches coalesced into a number of complex trust production logics and led to the institutionalization of technological trust production. Due to limited space, I only discuss here three of these: *reputation, control*, and *insight*. A more detailed account is the task of future work.

Probably the most visible and intuitive institutionalized form of technological trust production is the global, interaction-specific catalog of human *reputation*. The US credit scoring industry (Lauer, 2017) was the first to produce reputation-based trust, which was independent from any geographically, temporally, socially, culturally, or politically defined context. With the rise of the creative class (Gandini, 2016) and precarious (digital) labor, a corresponding digital infrastructure emerged to facilitate the economic involvement of a dispersed and atomized workforce through the commodification of individual reputation. Signals of trustworthiness are fine-tuned to the requirements of specific interactions (i.e. hospitality, transportation) and standardized across highly different local contexts. Trust mediators can also mobilize automation to monitor and police interactions, and incorporate external trust-relevant signals in the online reputation. Automation and standardization allow interpersonal trust to be produced on unprecedented scales and speeds. Such centralized, systematic, and automatized accounts of

reputation transform trust and trustworthiness, a form of social capital, into a commodity, an industrially produced asset that can be quantified, traded, enclosed, and sanctioned. The emerging Black markets in trustworthy online profiles (Damaini et al., 2018) and nascent discussions and laws on reputation portability (Hesse and Teubner, 2019) are just two indications of the importance of reputation-based technology-mediated trust. In a closely related manner, the Chinese social scoring system arguably achieves the same effect, with no less social impact (Elgan, 2019).

Blockchain systems follow a different logic, that of *control*. These systems try to minimize the need for trust and produce confidence by hard-coding rules into the system, both at the level of infrastructure and in their application (smart contracts). This ensures that the behavior of the system is predictable. Crypto tokens, the value distribution mechanisms in these systems are similar to the trust generating abstract tokens described by Giddens (1990), which circulate across different localities without changing their meaning.

Machine learning–based systems produce trust from *insight*. More and more exact and comparable data; more quantifiable, verifiable, and objectifiable knowledge; and more transparency has led to more predictable, arguably trustworthy private and public institutions through increasing the calculability of risks, reduction of uncertainties, and limitation of contingencies (Beniger, 1986; Foucault, 1991; Porter, 1996). Technologies that rely on the systematic, automated analysis of large datasets, such as machine learning, and automated recommendation and decision systems in commerce, search, ad-tech, fintech, news, or social media build on the same logic of mechanical objectification to produce trust between information producers and consumers (Tang and Liu, 2015). They transform unprecedented insights into their users' actions into suggestions of arguably trustworthy connections and interactions. Google's PageRank algorithm transforms insight into the Internet's linking structure into measures of relevance. Facebook and Twitter increasingly analyze patterns of information posting, sharing, and consumption to filter untrustworthy information, and to populate our information environments with familiar, trusted sources. There are many struggles, conflicts, and failures around this logic. The nature, working, and reliability of such insight-based trust production often remain opaque or incomprehensible. The supposedly trustworthy suggestions or decisions are often contentious or outright wrong. Yet, despite the concerns, the power of the objectification logic and the ubiquity and convenience of such systems make them hard to avoid.

These various logics of direct technological institutional trust production are not necessarily new in and by themselves. Professional associations have always played an important role in maintaining reputations. Rites, standards, protocols, or contracts are timeless instruments of control. The scientific revolution was launched by the shift toward the current logics of quantification, objectification, and insight. Yet, the technological modes of institutional trust production also differ substantially from the traditional logics of institutional trust, as they come with a number of new, known, and unknown, risks.

Technological trust producers rely on automation, the minimization of human oversight, and the ability continuously to upgrade their technical design in response to the changes in their technical and social environment (Gürses and van Hoboken, 2018).

They tend to be weakly embedded in the institutional distrust management frameworks, including regulation. The scale, speed, modularity, temporality, and adaptability of technological trust-mediators differ starkly from the design of our traditional institutions of trust, which is still deeply rooted in temporal and social predictability, familiarity, and stability.

Digital technologies have clear agency in the interpersonal and institutional processes of trust mediation. The more the technology is focused on producing trust, the more its trustworthiness needs to be externally assessable, verifiable, and comparable to other abstract systems of trust production. In the next section, I will discuss the question of trust *in* trust mediating technology.

## *Trust* in *technology*

It is impossible to discuss the trust produced *by* technology without considering the trustworthiness *of* trust-producing technologies. The agency of technology is crucial, yet, its influence on the trust it produces is largely unknown.

The question of trust in digital technology has a long history in Computer science/Information Science/Computer Mediated Communication research (Cheshire, 2011; Clarke, 2006; Harper, 2014; McKnight et al., 2002, 2011; Söllner et al., 2016). Trust research in these disciplines is very empirically focused, and pays a lot of attention of the trust-related dispositions, intentions, and behavior of the user vis-à-vis a digital technology. In contrast, possibly because it is harder to operationalize, the institutional conditions of the trustworthiness seem to be undertheorized. This article may contribute to the existing models by elaborating the institutional aspects of trustworthiness, which may affect the individual behavior, but operate independent of both the trustor's mental state and the intrinsic qualities of technologies. The issue, as I detail below, is the independent (institutional) assessability and verifiability of trustworthiness, which affects the validity of the technology-mediated trustworthiness of individuals wishing to cooperate with the help of trust mediators.

The main reason for this latter approach is that the strategies by which we engage with and via trustworthy systems can be starkly different from those that we mobilize when we need to rely on systems that are not to be trusted. The agency of technology needs to be recognized so it can be properly addressed. The first step in this process is to inquire whether we can trust technology to produce trust. As I argue in the following section, structural hurdles prevent the establishment of this trustworthiness.

The most straightforward way to approach the trustworthiness of trust mediators is through the ability, benevolence, and integrity (ABI) framework (Mayer et al., 1995). "Ability" or competence relates to the skills and expertise the trustee must possess in order to meet the relevant expectations with regard to the context where trust is needed. "Benevolence" refers to the trustee's willingness to act beyond their own self-interest, to the benefit of the trustor. "Integrity" signals that the trustor and the trustee share some fundamental moral, ethical, or legal principles, which guide the actions of the trustee.

*Competence.* The competence of trust mediators is context-specific. A search engine, for example, is expected to show the most relevant answers to a specific information query.

The electronic marketplace must link supply and demand in a trustworthy way. A smart contract must be bug-free and work as promised. Organizational safeguards should be in place if technology fails. Automated decisions should be explainable and contestable. Conflict resolution and arbitration regimes should be able to review and resolve complaints. Fraud should be detected, prevented, or remedied. If internal procedures fail, judicial review should remain an option.

The problem with building confidence in the technical aspects of system competence is the *instability of the technology and the fragmentation of contexts and experiences*. We can only judge the trustworthiness of technology based on our own past experience, or based on others' experiences. None of this is helpful, however, if the same technology is deployed in radically different contexts, if it produces probabilistic outcomes, if the technology changes quickly, or if the technology produces personalized outcomes that differ for each user. The experience of technology's ability to fulfill expectations will be limited to each single case. What I experienced today may not be what I will experience tomorrow, in a different context or at a different location, or it may not be what someone else experienced at the same time or under similar conditions. Individuals and societies lack appropriate procedures, infrastructures, and institutions to observe, accumulate, and aggregate such unique experiences across greater temporal and socio-geographic distances (Bodó et al., 2017). Such aggregation is the core competence of the technologies themselves, of course, but no mechanisms have emerged to date that have external access to such aggregates or that build comparable knowledge independently. Given their lack of reliable firsthand trust signals, users must rely on secondary signs, Keynesian beauty contests (Keynes, 1964), and speculative bubbles to guess the competence of trust mediators (Csigó, 2016).

On the other hand, the problem with assessing the organizational aspect of competence is *incompleteness.* Most readers are likely to be aware of situations in which social media companies, content platforms, or search engines have censored information. It is usually hard to get a comprehensive explanation of the process or reasons behind such censorship, and it often takes considerable effort to contest such a decision. According to Suzor (2019), this is because technology firms operate in a state of lawlessness. Technology companies' competitive advantage rests upon their ability to scale their operation technologically with minimal or no human intervention. Human oversight does not scale well, meaning that trust mediators have an economic incentive to minimize the costs of the non-scaling human, organizational aspects of their operations by standardizing procedures, limiting their scope, and outsourcing such activity to subcontractors in low-income countries. This standardization and outsourcing disembed human oversight from the local norms, standards, applicable laws, and institutional frameworks that would otherwise address such issues. Such reduced governance mechanisms offer limited, decontextualized, imprecise, and inscrutable procedures for users to handle the risks and harms produced by the technologies.

The unstable nature of the technical components of trust mediators results in competence uncertainty. Organizational safeguards could potentially alleviate this uncertainty, but there are substantial economic disincentives for firms to go beyond the bare minimum. Taken together, these logics make the external assessment of the competence of trust mediators extremely problematic.

*Benevolence.* The benevolence aspect of trust mediators concerns their ability and willingness to act in the interests of their users, potentially at the expense of their own self-interest. Hardin (2002) refers to this as "encapsulated interest," Zucker (1985) calls it "independence from self-interest," and Parsons (1939) describes it as "other-orientation." There are two systemic challenges associated with the assumption that trust mediators are benevolent.

First, there are severe and persistent asymmetries of information and power between trust mediators and users that prevent the latter from forming reasonable expectations about the benevolence of trust mediators. Trust mediators know about their users' most intimate details; preserve long-forgotten memories; and analyze communication, transactions, social interactions, purchases, interests, and psychological profiles. Such extreme exposure forms the raw material for trust mediation. On the other hand, trust mediators remain completely non-transparent. The workings of their technology are impenetrable; their internal rules, procedures, organization, and responsibilities remain opaque; and the details of business relationships with third parties are kept secret. This information asymmetry translates into power asymmetry. Users have very little ability to learn about, assess, or exercise control over the operation of trust-mediating technologies, while trust mediators themselves are in a position to exercise control over certain aspects of their users' lives. In addition, most users have little choice as to whether and how they engage with trust mediators. Some trust mediators occupy monopoly or quasi-monopoly positions, and can get away with offering "take it or leave it" choices, because for the individual, the cost of exclusion is simply too high (Feld, 2019). Leaving often means being locked out of whole domains of trust relationships.

Second, trust-mediating technologies are situated in a complex web of interests in which users are not the only, and often not the most important, stakeholders. The interest of users in relation to privacy, in relation to the goals served by algorithmic decisions, or to how users are represented on and served by trust-mediating technological systems, compete against the interests of other parties. The company wants to make savings on human oversight costs, in order to meet the profit expectations of shareholders and investors. Sellers and advertisers who pay for the services of trust mediators want to see and control user behavior. Governments want trust mediators to help them enforce the law. All of these interests directly compete with the legitimate interests of users.

Trust mediators are expected to encapsulate the interests of those parties that are able to exercise some form of power or control over them. Shareholders control trust mediators on the basis of the stock price and the company board. Advertisers and other businesses exercise control through the fees they pay and the business they bring. Governments exercise control over trust mediators through regulation. Users, meanwhile, have little power to negotiate the terms of service; they do not pay fees in many cases (or at least not directly); and they have little voice in the governance of the technologies. This relative powerlessness, in comparison to all the other interests, as well as vis-à-vis the trust mediators themselves, forms the second systemic challenge to the benevolence of trust mediators.

These extreme power and information asymmetries prevent users from basing their expectations of the benevolence of trust mediators on a rational, calculative form of trust.

What they are left with is faith, the same relationship the faithful have with inscrutable, all-powerful, and life-controlling deities.

*Integrity.* Finally, by "integrity," we mean that that the trustee's actions are congruent with their words, and that there is some shared moral and ethical ground between the trustor and the trustee.

At present, the integrity of trust mediators is under renewed scrutiny. When Google's famous 'Don't be evil' motto is silently removed from the company's code of conduct (Conger, 2018), and when Facebook argues in court that the Cambridge Analytica affair did not constitute a breach of privacy (Biddle, 2019), such companies are setting themselves apart from the value-expectations of their users and trying to establish new norms. When tech CEOs have to step down due to their personal conduct (Isaac, 2017), when tech employees publicly object to their firms doing business with authoritarian regimes (Gallagher, 2019) or contentious domestic law-enforcement practices (Chan, 2019), and when consumers pay attention to firms' questionable environmental or labor records (Merchant, 2017), the integrity of these companies is called into question.

There are systemic reasons behind a sustained value incongruence between users and highly innovative technologies on a global scale. The two most important logics at work here are the disruptive nature of innovation and the difficulty of maintaining integrity across a diverse and conflicting global value landscape.

First, digital innovation is disruptive. Similar eras when innovation led to the radical transformations of social, economic, cultural, and political relations and practices have been associated with rapid declines in trust and the rise of new modes of trust production (Zucker, 1985). Disruption replaces existing institutions and logics of social and economic organization and trust production with new ones, and thereby destroys the predictability and familiarity of the economic and social relationships that these institutions formed and maintained. It is hard to assume natural and frictionless value congruence within the logic of disruptive innovation when it aims to destroy and rebuild pre-existing trust expectations, logics, values, and institutions.

Second, trust mediators gain their power from the standardization and automation of trust production across widely different local contexts. The global visibility of trust mediators' local actions makes it very difficult for them to isolate the effects of their commitment to incompatible value regimes in different local contexts. Many Western technology companies have had to consider the reputational and trust-related costs of doing business in countries that are seen as oppressive from a Western perspective. Similar concerns have been raised about doing business with Western military actors, secret services, or law enforcement bodies. A number of value communities struggle to establish their own norms over trust-mediating services. Attempts to neutralize such value conflicts by removing human discretion and introducing automation only exposes and amplifies them, as Twitter recognized when the artificial intelligence (AI) system intended to remove Nazi hate speech reportedly started to filter Republican politicians (Panetta, 2019).

The integrity of trust mediators ultimately boils down to commercial considerations. Maintaining a level of integrity that reflects fundamental Western values may come at the price of not doing business in some other parts of the world. On the other hand, a

profit-prioritizing approach may make it difficult to maintain integrity in one or more local contexts.

It would seem reasonable to establish the trustworthiness of trust mediators before we use them to produce interpersonal and institutional trust, and allow pre-existing modes of interpersonal and institutional trust production to be transformed. Yet, systemic barriers are preventing both users and societies from building confidence in trust-mediating technologies. In the best-case scenario, it is impossible to establish the trustworthiness of trust mediators. In the worst case, however, they are not to be trusted at all, because they are unable to produce trust in a trustworthy manner, because they do not have their users' best interests at heart, or because they lack the necessary moral integrity to do so. The implications of these different scenarios will be addressed briefly in the conclusion.

## Conclusion: shifting the sources of trust in trust mediators

To date, the source of trust in digital technology has mainly been ideological, rooted in a mix of American 1960s counterculture, neoliberal economic thought, and libertarian political beliefs, which suggested that better technological tools lead to more just, equitable, inclusive, and democratic social, political, and economic structures; that regulation comes at the cost of innovation; that the powers of central government are dangerous, but that disruption based on market innovation is beneficial; and that progress is inevitable (Barbrook and Cameron, 1996; Morozov, 2014). This militant optimism (Rossetto, 2018) was able to create and support trust and dispel distrust of technology when all other procedural sources of establishing trustworthiness were lacking, and when firsthand experiences clashed with expectations.

At the beginning of the 2020s, faith in technology in general, and unregulated innovation in particular is dwindling. Highly contentious social, cultural, political, and legal conflicts called into question the competence of technological trust mediators and our reasons to trust them. A wide range of issues are debated, such as the liability for data breaches (Shackelford et al., 2015); intermediary liability for spreading hate speech or misinformation (Klonick, 2017); the democratic function, control, and oversight of recommender systems (Helberger, 2019), the control of discrimination by AI systems (Zuiderveen Borgesius, 2019); or the privacy guarantees of contract tracing apps, used to control pandemics (PEPP-PT Team, 2020). The debates also illustrate how difficult it is to calculate the costs and benefits, the governance, the vulnerabilities, and the risks of trusting technology with producing trust.

It may seem that the ubiquity and power of digital intermediaries make it almost impossible to address their "lawlessness" (Suzor, 2019). Yet, there are a number of potential approaches to increase their trustworthiness. The detailed description of this institutional approach goes beyond the limits of this article, and will be discussed in future work. Here, I refer only the two most important challenges that the institutionalization of technological trust production faces: the limits of internal trust guarantees and the conditions of external accountability.

Enthusiasts of blockchain technology and the decentralized web argue that certain technology design choices, such as open sourcing software; decentralization; the possibility of open participation in the design and operation of technology; radical

transparency in all aspects of the system, including standards, protocols, algorithms, transaction records, and governance; the standardization and objectification of decisions; and the general reduction of problems to mathematical indicators, are adequate internal logics to address systemic trust issues in trust mediation (Bodó and Giannopoulou, 2019).

I argue that these architectural trust guarantees are necessary but may not be sufficient precursors to trustworthiness. Technological safeguards must be complemented by adequate internal governance structures and clear external accountability. Rules need to be unambiguous, due process should exist to adjudicate conflicts, clear paths are required to redress mistakes, and so on. Such organizational measures cannot be completely incongruent with local contexts, norms, customs, and institutions, and they thus need to be embedded in local institutional interdependencies of distrust management.

This embeddedness is the precondition of external accountability. It includes, but should not be limited to, the regulation of trust mediators, issues of jurisdiction and enforcement, the creation of legal certainty, and establishing the applicable laws. The local embeddedness of technological trust mediators includes issues such as where they store their data or host their infrastructures, where they develop their technologies, how they are present in the local economy, and how they relate to local actors, from communities to professional associations. All of these interfaces, exposures, and frictions provide local insight and control over the operation of global-scale trust technologies and integrate them in the institutional networks of distrust. At present, these different guarantees of the trustworthiness of trust mediators are implemented in a fragmentary and incomplete manner, if at all. In the long run, these isolated approaches will need to develop together.

Digital trust mediation has been a relatively marginal service to date, but it is quickly becoming a core element of the digital infrastructure. This implies that we need better analytical tools to assess the known and unknown risks associated with digital technologies, allowing us to manage distrust, design them to be more trustworthy, and rely on the trust they produce.

## Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The Lab has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 759681). The paper has been written with the support of a Fellowship grant at the Weizenbaum Institute, Berlin. The author is immensely grateful for the ideas, inspiration, critique, helpful comments of the members of the Lab: João Pedro Quintais, Alexandra Giannopoulou, and Valeria Ferrari; the members of the Distributed Trust Consortium: Esther Keymolen, Seda Gürses, Jaap-Henk Hoepman, and Jurgen Goossens; the members of the 'Trust in distributed environments' Research Group at the Weizenbaum Institute; and the editors and anonymous reviewers of NMS.

## ORCID iD

Balázs Bodó https://orcid.org/0000-0001-5623-5448

## Notes

1. For a detailed definition, see Section 'Defining mediated trust'.
2. Despite important differences in the literature, I use the term "risk" to refer to future events, both calculable and unforeseeable, which necessitate a "leap of faith," or trust to manage, to refer to Beck's work on "risk society" and the unforeseen risks digital (trust) technologies may bring forward.
3. There is substantial disagreement in the legal literature on the impact of (contractual or legal) control and trust. Some argue that trust and control are mutually exclusive. Others suggest that a need for control signals mistrust, and thus undermines trust. Still others suggest that trust is a precondition for entering into contractual relations. There is evidence that legal certainty and regulation is an important enabler of trust. Trust and control (through contracts) can also complement and strengthen each other. See Woolthuis et al. (2005) for an overview of the literature. I have no intention to take a position in that debate.
4. This definition regards the technical component as part of a complex techno-social assemblage with institutional, and human elements. It is in line with the mainstream STS literature, but is somewhat in contrast with a more traditional separation of information systems and system providers in the Computer Science and Information Systems literature, such as Söllner et al. (2016) or McKnight et al. (2011). I agree with Gürses and van Hoboken (2018) that the agile turn in software development rendered the strict separation of information systems and system providers obsolete.
5. Technology operators may both cause and design addiction (Eyal, 2014; Kuss and Griffiths, 2011). Plastic surgeons have warned about 'Snapchat dysmorphia': patients requesting surgery to resemble photo-manipulated selfies (Ramphul and Mejias, 2018).

## References

Ananny M (2019) *Probably speech, maybe free: toward a probabilistic understanding of online expression and platform governance*. Report, 21 August. New York: Knight First Amendment Institute. Available at: https://knightcolumbia.org/content/probably-speech-maybe-free-toward-a-probabilistic-understanding-of-online-expression-and-platform-governance

Barbrook R and Cameron A (1996) The Californian ideology. *Science as Culture* 6(1): 44–72.

Beck U (1992) *Risk Society: Towards a New Modernity (Theory, Culture & Society)*. London; Newbury Park, CA: SAGE.

Beniger JR (1986) *The Control Revolution : Technological and Economic Origins of the Information Society*. Cambridge, MA: Harvard University Press.

Biddle S (2019) In court, Facebook blames users for destroying right to privacy. *The Intercept*, 14 June. Available at: https://theintercept.com/2019/06/14/facebook-privacy-policy-court/

Bijker WE, Hughes TP, Pinch TJ, et al. (1987) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, MA: MIT Press.

Bodó B (2019) Selling news to audiences: a qualitative inquiry into the emerging logics of algorithmic news personalization in European quality news media. *Digital Journalism* 7: 1054–1075.

Bodó B and Giannopoulou A (2019) The logics of technology decentralization: the case of distributed ledger technologies. In: Ragnedda M and Destefanis G (eds) *Blockchain and Web 3.0: Social, Economic, and Technological Challenges*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3330590

Bodó B, Helberger N, Irion K, et al. (2017) Tackling the algorithmic control crisis: the technical, legal, and ethical challenges of research into algorithmic agents. *Yale Journal of Law & Technology* 19: 133–180.

Botsman R (2017) *Who Can You Trust? How Technology Brought Us Together and Why It Might Drive Us Apart*. 1st ed. New York: PublicAffairs.

Buzan B and Lawson G (2015) *The Global Transformation: History, Modernity and the Making of International Relations (Cambridge Studies in International Relations 135)*. Cambridge: Cambridge University Press.

Catterberg G and Moreno A (2005) The individual bases of political trust: trends in new and established democracies. *International Journal of Public Opinion Research* 18(1): 31–48.

Chan R (2019) Microsoft workers are asking the company to cancel its $8 million of contracts with ICE, after their colleagues at GitHub take a stand. *Business Insider Nederland*. 10 October. Available at: https://www.businessinsider.nl/microsoft-github-ice-contracts-solidarity-2019-10/

Cheshire C (2011) Online trust, trustworthiness, or assurance? *Daedalus* 140(4): 49–58.

Clarke K (ed.) (2006) *Trust in Technology: A Socio-technical Perspective (Computer Supported Cooperative Work 36)*. Dordrecht: Springer.

Cofta P (2006) Distrust. In: *Proceedings of the 8th international conference on electronic commerce the new E-commerce: innovations for conquering current barriers, obstacles and limitations to conducting successful business on the Internet—ICEC'06*, Fredericton, NB, Canada, 13–16 August: New York: ACM Press.

Conger K (2018) Google removes 'don't be evil' clause from its code of conduct. *Gizmodo*, 18 May. Available at: https://gizmodo.com/google-removes-nearly-all-mentions-of-dont-be-evil-from-1826153393

Corritore CL, Kracher B and Wiedenbeck S (2003) Editorial. *International Journal of Human-computer Studies, Trust and Technology* 58(6): 633–635.

Crawford K and Joler V (2018) Anatomy of an AI system: the Amazon echo as an anatomical map of human labor, data and planetary resources. *AI Now Institute and Share Lab*. Available at: http://www.anatomyof.ai

Crozier M, Huntington SP and Watanuki J (1975) *The Crisis of Democracy: Report on the Governability of Democracies to the Trilateral Commission*. New York: New York University Press.

Csigó P (2016) *The Neopopular Bubble*. Budapest: CEU Press.

Damaini AA, Nugroho GS and Suyoto S (2018) Fraud crime mitigation of mobile application users for online transportation. *International Journal of Interactive Mobile Technologies* 12(3): 153.

Ehrenkranz M (2019) Snapchat employees allegedly misused Internal tools to snoop on users: report. *Gizmodo*, 23 May. Available at: https://gizmodo.com/snapchat-employees-allegedly-misused-internal-tools-to-1834989316

Elgan M (2019) Uh-Oh: Silicon Valley is building a Chinese-style social credit system. *Fast Company*, 26 August. Available at: https://www.fastcompany.com/90394048/uh-oh-silicon-valley-is-building-a-chinese-style-social-credit-system

Eubanks V (2017) *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Eyal N (2014) *Hooked: How to Build Habit-forming Products*. New York: Portfolio; Penguin.

Feld H (2019) *The Case for the Digital Platform Act: Market Structure and Regulation of Digital Platforms*. New York: Roosevelt Institute, Public Knowledge. Available at: https://www.publicknowledge.org/assets/uploads/documents/Case_for_the_Digital_Platform_Act_Harold_Feld_2019.pdf

Foucault M (1991) Governmentality. In: Burchell G, Gordon C and Miller P (eds) *The Foucault Effect: Studies in Governmentality*. Chicago, IL: University of Chicago Press, pp. 87–104.

Fukuyama F (1995) *Trust: The Social Virtues and the Creation of Prosperity*. New York: Free Press.

Gallagher R (2019) Google employees uncover ongoing work on censored China search (Blog). *The Intercept*, 4 March. Available at: https://theintercept.com/2019/03/04/google-ongoing-project-dragonfly/

Gandini A (2016) *Reputation Economy: Understanding Knowledge Work in Digital Society*. London: Palgrave Macmillan.

Gartenberg C (2019) Apple's hired contractors are listening to your recorded Siri conversations, too. *The Verge*, 26 July. Available at: https://www.theverge.com/2019/7/26/8932064/apple-siri-private-conversation-recording-explanation-alexa-google-assistant

Gereffi G and Korzeniewicz M (eds) (1994) *Commodity Chains and Global Capitalism (Contributions in Economics and Economic History 149)*. Westport, CT: Greenwood Press.

Giddens A (1990) *The Consequences of Modernity*. Cambridge: Polity Press.

Gillespie T (2010) The politics of "platforms." *New Media & Society* 12(3): 347–364.

Graves-Brown P, Harrison R, Piccini A, et al. (2013) *The Oxford Handbook of the Archaeology of the Contemporary World (Oxford Handbooks in Archaeology)*. 1st ed. Oxford; New York: Oxford University Press.

Gürses S and van Hoboken J (2018) Privacy after the agile turn. In: Selinger E, Polonetsky J and Tene O (eds) *The Cambridge Handbook of Consumer Privacy*. Cambridge: Cambridge University Press, pp. 579–601.

Hardin R (2002) *Trust and Trustworthiness (The Russell Sage Foundation Series on Trust)*, vol. 4. New York: Russell Sage Foundation.

Hardin R (ed.) (2004) *Distrust (Russell Sage Foundation Series on Trust)* vol. 8. New York: Russell Sage Foundation.

Harper RHR (ed.) (2014) *Trust, Computing, and Society*. New York: Cambridge University Press.

Helberger N (2019) On the democratic role of news recommenders. *Digital Journalism* 7(8): 993–1012.

Hesse M and Teubner T (2019) Reputation portability: quo vadis? *Electronic Markets*. Epub ahead of print 13 September. DOI: 10.1007/s12525-019-00367-6.

Houser K (2020) Airbnb claims its AI can predict whether guests are psychopaths. *Futurism*, 4 January. Available at: https://futurism.com/the-byte/airbnb-ai-predict-psychopaths

Inglehart R (1999) Trust, well-being and democracy. In: Warren ME (ed.) *Democracy and Trust*. Cambridge: Cambridge University Press, pp. 88–120.

Isaac M (2017) Uber founder Travis Kalanick resigns as C.E.O. *The New York Times*, 21 June. Available at: https://www.nytimes.com/2017/06/21/technology/uber-ceo-travis-kalanick.html

Kerr D (2019) Some Uber drivers aren't who you think they are. *CNET*, 26 November. Available at: https://www.cnet.com/news/uber-drivers-using-fake-identities-isnt-just-a-london-problem/

Keymolen ELO (2016) *Trust on the Line: A Philosophical Exploration of Trust in the Networked Era*. Rotterdam: Erasmus University Rotterdam. Available at: hdl.handle.net/1765/93210

Keynes JM (1964) *The General Theory of Employment, Interest, and Money*. San Diego, CA: Harcourt Brace Jovanovich.

Klonick K (2017) The new governors: the people, rules, and processes governing online speech. *Harvard Law Review* 131: 1598.

Kuss DJ and Griffiths MD (2011) Online social networking and addiction: a review of the psychological literature. *International Journal of Environmental Research and Public Health* 8(9): 3528–3552.

Lauer J (2017) *Creditworthy: A History of Consumer Surveillance and Financial Identity in America (Columbia Studies in the History of U.S. Capitalism)*. New York: Columbia University Press.

Luhmann N (2017) *Trust and Power*. Malden, MA: Polity Press.

McKnight DH, Carter M, Thatcher JB, et al. (2011) Trust in a specific technology: an investigation of its components and measures. *ACM Transactions on Management Information Systems* 2(2): 121–125.

McKnight DH, Choudhury V and Kacmar C (2002) Developing and validating trust measures for E-commerce: an integrative typology. *Information Systems Research* 13(3): 334–359.

McLuhan M (1964) *Understanding Media; the Extensions of Man*. 1st ed. New York: McGraw-Hill.

Marsh S and Dibben MR (2005) Trust, untrust, distrust and mistrust: an exploration of the dark(er) side. In: Herrmann P, Issarny V and Shiu S (eds) *Trust Management: Lecture Notes in Computer Science*, vol. 3477. Berlin; Heidelberg: Springer, pp. 17–33.

Mayer RC, Davis JH and David Schoorman F (1995) An integrative model of organizational trust. *The Academy of Management Review* 20(3): 709–734.

Merchant B (2017) Life and death in Apple's forbidden city. *The Observer*, 18 June. Available at: https://www.theguardian.com/technology/2017/jun/18/foxconn-life-death-forbidden-city-longhua-suicide-apple-iphone-brian-merchant-one-device-extract

Misztal BA (1996) *Trust in Modern Societies: The Search for the Bases of Social Order*. Cambridge: Polity Press.

Morozov E (2014) *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: PublicAffairs.

Nissenbaum H (2001) Securing trust online: wisdom or oxymoron? *Boston University Law Review* 81(3): 101–131.

Nixon P and Terzis S (eds) (2003) *Proceeding of the Trust Management: First International Conference, Itrust 2003: Lecture Notes in Computer Science*, vol. 2692. Berlin: Springer.

Nye JS, Zelikow P and King DC (1997) *Why People Don't Trust Government*. Cambridge, MA: Harvard University Press.

Ong WJ and Hartley J (2012) *Orality and Literacy: The Technologizing of the Word (Orality and Literary)*. London; New York: Routledge.

Panetta G (2019) Twitter reportedly won't use an algorithm to crack down on white supremacists because some GOP politicians could end up getting barred too. *Business Insider*. 25 April. Available at: https://www.businessinsider.nl/twitter-algorithm-crackdown-white-supremacy-gop-politicians-report-2019-4/

Parsons T (1939) The professions and social structure. *Social Forces* 17(4): 457–467.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

PEPP-PT Team (2020) Documentation for pan-European privacy-preserving proximity tracing (PEPP-PT). Available at: https://github.com/pepp-pt/pepp-pt-documentation

Plantin J-C, Lagoze C, Edwards PN, et al. (2018) Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society* 20(1): 293–310.

Porter TM (1996) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Putnam RD (2001) *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.

Ramphul K and Mejias SG (2018) Is "Snapchat dysmorphia" a real issue? *Cureus* 10(3): 2263.

Rossetto L (2018) Beyond the digital revolution: the need for militant optimism. *Wired*, 18 September. Available at: https://www.wired.com/story/wired25-louis-rossetto-tech-militant-optimism/

Schneier B (2012) *Liars and Outliers: Enabling the Trust That Society Needs to Thrive*. Indianapolis, IN: Wiley.
Seligman AB (2000) *The Problem of Trust*. Princeton, NJ: Princeton University Press.
Shackelford SJ, Proia AA, Martell B, et al. (2015) Toward a global cybersecurity standard of care: exploring the implications of the 2014 NIST cybersecurity framework on shaping reasonable national and international cybersecurity practices. *Texas International Law Journal* 50: 305.
Shapiro S (1987) The social control of impersonal trust. *American Journal of Sociology* 93(3): 623–658.
Simmel G and Frisby D (2004) *The Philosophy of Money*. London: Routledge.
Söllner M, Hoffmann A and Leimeister JM (2016) Why different trust relationships matter for information systems users. *European Journal of Information Systems* 25(3): 274–287.
Suzor NP (2019) *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge: Cambridge University Press.
Sztompka P (1999) *Trust: A Sociological Theory (Cambridge Cultural Social Studies)*. Cambridge: Cambridge University Press.
Tang J and Liu H (2015) *Trust in Social Media*. San Rafael, CA: Morgan & Claypool.
van Dijck J, Nieborg D and Poell T (2019) Reframing platform power. *Internet Policy Review* 8(2): 1414.
van Hoboken J (2013) *The Proposed Right to Be Forgotten Seen from the Perspective of Our Right to Remember*. Luxembourg: European Commission. Available at: http://dx.publications.europa.eu/10.2788/51998
Walker J and Ostrom E (2003) *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*. New York: Russell Sage Foundation.
Werbach K (2018) *The Blockchain and the New Architecture of Trust*. Cambridge, MA: MIT Press.
Woolthuis RJAK, Hillebrand B and Nooteboom B (2005) Trust, contract and relationship development. *Organization Studies* 26(6): 813–840.
Yeung K (2019) Regulation by blockchain: the emerging battle for supremacy between the code of law and code as law. *The Modern Law Review* 82(2): 207–239.
Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power*. London: Profile Books.
Zucker LG (1985) Production of trust: institutional sources of economic structure, 1840 to 1920. In: Cummings LL and Staw B (eds) *Research in Organizational Behavior*. Greenwich, CT: JAI Press, pp. 53–111.
Zuiderveen Borgesius FJ (2019) Discrimination, artificial intelligence, and algorithmic decision-making. Report, 7 February. Strasbourg: Council of Europe. Available at: https://www.coe.int/en/web/artificial-intelligence/-/news-of-the-european-commission-against-racism-and-intolerance-ecri-

## Author biography

Balázs Bodó is an associate professor, senior research scientist at the Institute for Information Law, University of Amsterdam, and the Principal Investigator of the European Research Council funded Blockhain and Society Policy Research Lab. He is a two-time Fulbright Scholar (2006-7, Stanford University; 2012 Harvard University), and a former Marie Skłodowska-Curie fellow (2013-15). His academic research focuses on domains and processes of knowledge production and exchange: their histories; social, economic and technological organization; and regulation.