Introduction: Trust, Responsibility, and Digital Governance

Sebastian Bücker, Marcus Düwell, Andreas Kaminski, Michael Leyer

Digital tools are increasingly influencing and shaping more aspects of our lives. We communicate via information technology, we develop a shared understanding by reading shared texts, watching pictures and videos, we make decisions based on information and recommendations provided by digital systems. When we are in contact with professionals, like physicians, financial advisers, lecturers, or travel agents, they often communicate with us based on information they obtain from digital tools; often, this communication itself is shaped by digital systems. More and more digital systems are supposed to implement functional roles that were previously performed by humans, and even our social and personal relationships are often initiated, mediated, and shaped by digital tools.

This ubiquity of digital tools comes with specific challenges. The question of whether we can trust the information we receive, the institutions we have to deal with and to what extent we can trust each other, depends to a large extent on the way digital tools are shaped, how they function, and how they are controlled. The physician searching in a computer for medicine for an illness is not in a position to judge how the program is influenced by the pharmaceutical industry. The student who asks ChatGPT about the subject for his homework cannot judge to which extent the given information can be taken as testimony. The teen who shares photos from the last vacation on social media cannot judge who has access to this data and what conclusions are drawn from it. We all are dependent on the governance of these digital aids and our trust in the social world and our trust in each other depends on the quality of this digital governance.

But how can we tell whether that is actually the case? This question is often taken to mean that we should develop a set of criteria and then evaluate digital systems by checking whether they meet those criteria. Criteria that are often proposed in these contexts are, for example, transparency, non-discrimination, non-abusiveness, etc. Next to the question of whether there is consensus about what the criteria should be, first there are some questions that are much more basic and tacitly presupposed in debates on the normative assessment of digital systems, namely questions regarding our ability to understand the digital systems in the first place. Before we can ask

whether a digital system is designed in a non-discriminatory way, we have to ask: Are we even able to understand how digital systems are designed? Do we possess the capacity to understand how they shape our actions, our communications, and our decisions? If we are not able to understand how digital systems work, the ethical and legal assessment of whether they meet normative criteria cannot even be posed. And if this more basic capacity of understanding and evaluating is at stake, we can furthermore ask: How can this capacity to assess systems and their effects be cultivated and developed? How can digital governance support individuals and institutions in building such capacities? Thus, this edition asks those questions that are necessarily required for being able to evaluate digital systems, independently of what the criteria for a normative assessment may be.

1. Two Pathways Towards Trustworthy Systems and Responsible Digital Governance

In line with these considerations, we can distinguish two approaches to evaluate digital systems which are, however, in no way mutually exclusive, as we will see. Most approaches start by analyzing values and evaluating systems against them; meanwhile, our approach focuses on the question of what capacities are required to evaluate systems in the first place. We aim to briefly outline the difference and reflect on their mutual relationship.

1.1. The Value-Approach

Starting from certain values seems a promising pathway, since we seem to agree on common values. Whatever we may expect from well-functioning digital tools, we at least demand some basic normative features. An example: What would we think about a digital tool that systematically refuses access for people of a certain gender or skin color? Of course, we do not want systems to have a fundamental sexist or racist bias. Likewise, if a mortgage-decision tool always ruled against applicants named 'Peter', we would doubt its trustworthiness – no matter how accurately it might function otherwise. This line of investigation assumes that defining criteria for responsible digital governance involves a double task: First, identifying the values that matter; second, determining how to ensure that digital governance aligns with them. This perspective is, for example, central to discussions of 'Value-Sensitive Design' (Friedman et al. 2013; Hillerbrand 2021; Van de Poel 2018). Many approaches follow similar ideas though without explicitly adopting this framework.

1.2. The Capacity-Approach

This volume follows an alternative line of investigation, which we call the capacity approach. This approach begins with the assumption that, prior to any evaluation, the user must be in a position to relate to the digital tools in an autonomous and responsible way. This entails that the ubiquitous presence of digital tools must not undermine the capacity of human beings to act responsibly and to relate to each other in a responsible manner. Such responsibility presupposes that the user has the capacity to understand digital systems and their impacts at least sufficiently to position themselves in relation to them – that is, to engage with them both epistemically and practically. It also requires that the digital systems be shaped so as to be sensitive to the capacity of the user to exercise independent judgment in dealing with them.

A first critical reaction could be the following: It is not up to the digital tool whether the user is competent to use it; rather, whether the user has the relevant capacities depends on the user. This comment is entirely correct. Different people will have different capacities; they will be capable of competently dealing with such tools to varying degrees. But the problem of judging a digital system as trustworthy cannot be remedied even if everybody had a degree in computer science: First, even in a digitalized world we still need physicians, bakers, bus drivers, and philosophers. Second, quite often the most technically competent people, e.g., developers or computer scientists, don't possess the capacities to account for all aspects of a system's behavior, especially in the realm of Artificial Intelligence. The relevant capacities are, however, not only a matter of individual training and education. Rather, we can observe that certain systems are designed based on fundamental assumptions about the capacities of the intended user.

Let us explain this: If the city council sends letters to all citizens to inform them about an upcoming election, it assumes that all citizens are capable of reading. It furthermore assumes that people are capable of informing themselves about the programs of political parties and choosing which party they want to vote for. Finally, it assumes that people are capable of going to the election office at the right moment (thus, people being capable of determining the right day and time, finding the right place, etc.). All of these are implicit but necessary assumptions about the capacities of the addressee of those letters which are embedded in the very practice of sending out those letters. Thus, these assumptions are necessary conditions without which the practice of sending out those letters would not even be intelligible. These 'necessarily presupposed capacities' concerning knowledge and practices that are implicitly presupposed in a certain context, can be distinguished from capacities and knowledge about the function of the election in general. Voters are not supposed to know the details of the organization of the election in general, the mechanics of vote counting, or the legal regulations governing the procedure. The electoral au-

thorities need to have much more knowledge about the entire process, but the voter needs some basic capacities to participate meaningfully in this practice. Similarly, the driver of a car needs some basic practical and judgmental capacities to be able to drive responsibly, while the mechanic needs more detailed technical skills that the driver of the car does not need to have to be a responsible driver. Moreover, cars should be designed in a way that the usual driving skills are sufficient to deal responsibly with a car.

Along these lines, we can observe that the way a digital tool is designed reveals which capacities are implicitly presupposed regarding its user. Some digital tools are created in a way that they require the capacities of a computer scientist to be used in a competent way. Some tools function in a way that forces the user to blindly 'trust' them, without being able to take a competent stance towards the information received from them. If a physician uses certain digital tools, these can be designed in a way that the competence of the physician as a responsible decision-maker remains untouched. But they can also be designed so that the tool effectively takes over the medical decision. If this were the case, the physician could hardly be held accountable for advice given based on those tools. Accordingly, the patient will have reasons not to trust a physician whom they suspect of receiving advice on grounds that are not intelligible to them.

The question investigated in this volume is not so much whether we can trust certain digital tools, but rather whether it will be possible to have trust in each other and in relevant institutions in a world mediated and shaped by digital tools. Whether the development of trust is possible and justified depends on whether those tools are governed in a responsible way. Whether the digital governance is realized responsibly will depend on the way in which the design of the digital tools enables human beings to deal with them, without undermining their capacity to relate responsibly to each other. This, however, depends primarily on how the design of the tools relates to the capacities of the assumed user.

In that sense, the capacity approach and the value approach are not independent of each other. The assessment of whether or not digital systems are designed in a responsible way, in the outlined sense, is a necessary prerequisite for them being morally acceptable. In some sense, this is a normative assessment that comes prior to further normative considerations. That means, independent of what users may expect from those digital systems, this would be a kind of basic requirement, and it likewise represents a basic normative quality of digital systems. In that sense, the capacity approach is not only relevant to a particular liberal framework but has much broader relevance. Of course, one can admit that the capacity approach is particular convincing for normative theories that assume the self-determination and autonomy of agents to be of central value, or that are grounded in human dignity and human rights. These concepts are, however, the starting points of most constitutions

and the international human rights-framework. In that sense, such assumptions are broadly shared.

With this short outline the context in which this book is situated has been sketched. It will not result in a checklist for determining the responsibility and trustworthiness of digital governance. Rather, the book will investigate the field in such a way as to clarify the pathways through which these questions can be answered.

2. Outline of the Book

Our edition is divided into three sections.

The first section **Capacities That Enable Digital Governance** outlines the program of the approach. The contribution by **Andreas Kaminski**, **Marcus Düwell** and **Philipp Richter**, *The Capacity-Oriented Approach*, presents the underlying conceptual structure. The text builds on an idea by Christoph Hubig, namely that the ethics of technology is not simply the application of ethics to a specific domain, but rather an effort to secure the capacities for ethical reflection within the technological domain itself. This perspective reveals how prudential and deontological considerations complementarily build upon this foundational premise of all moral judgments and ethical reflection. The contribution by **Sebastian Bücker** and **Nico Formánek** *Precautions for Medical Decision Support by LLMs* analyzes on which theoretical foundations (medical) propositions made by LLMs are built, showing in which sense these systems cannot partake in the practice of "giving and asking for reasons" (Brandom). This motivates specific precautions for whenever LLMs shall be incorporated as a support for medical decision-making. In their text *Individual and Organizational Capabilities for Assessing the "Trustworthiness" of AI Systems in Healthcare Settings. The Crucial Role of Structural Empowerment*, **Oliver Behn**, **Marc Jungtäubl**, **Michael Leyer,** and **Mascha Will-Zocholl** explore the impact of AI systems in organizations. The text identifies various dimensions that are characteristic features of a structural empowerment of employees that should be guaranteed by the use of AI systems. On this basis they ask what governance structures organizations should establish to enable employees to assess the AI's trustworthiness and to ensure the continuous development and maintenance of AI-related capacities.

The second section **Governance That Enables Understanding** examines the possibilities of epistemically grasping digital systems from a normative perspective. In their article *A Right to Explanations of AI Decisions*, **Elena Dubovitskaya** and **Gregor Bosold** examine whether the right to an explanation – recognized in European frameworks such as the General Data Protection Regulation (GDPR) – may be grounded in more fundamental rights. They also explore the potential role that local explanations might play in empowering individuals affected by automated de-

cisions, such as those used in credit scoring. **Florian Möslein** and **Michael Birkner** in their article *Cross-Chain Governance* explore the challenges in governing cross-chain blockchains, which enable interoperability between otherwise separated and heterogeneous blockchain ecosystems, thereby introducing novel complexities for blockchain governance. Governing such complexities in a responsible and trustworthy manner should, they argue, require not just technical solutions, but also address legal coordination, participant's evaluative capacities, and the design of systems capable of fostering trust. In *Blockchain-Based Governance of Financial Markets. Examining The SEC's Approach to Building Trust in Centralized and Decentralized Crypto Exchanges* **Sebastian Omlor** and **Hans Wilke** investigate how subsuming new digital systems under old legal paradigms can lead to incompatible, undesirable, and unfeasible obligations. They therefore characterize both traditional markets for equity securities and contrast these with recent markets for crypto assets. Building on this analysis, they show that the paradigm of recent markets has shifted to a point that they cannot be governed by regulatory frameworks designed for traditional markets for equity securities, thus finally outlining the necessary conditions for designing a framework that would be compatible with the paradigms of markets for crypto exchanges.

The third section **Enabling Conditions for Trust and Responsibility** explores the relation between capacities, trust, and responsibility. **Sebastian Bartsch**, **Marcus Düwell**, **Jan-Hendrik Schmidt,** and **Alexander Benlian** investigate in *Ethics and Regulation of Al Systems in Medicine. The Example of Cancer Detection* how accountability of healthcare professionals is possible if AI systems form an integrative part of medical practice. The text explores the ethical and regulatory implication of employing AI systems in this context and proposes a model for distributing accountability between the parties involved. **Andreas Kaminski** begins in *Trust in AI* from the compelling arguments that it is a category mistake to speak of trust in technology. However, as he shows, reducing trust to a narrow epistemic concept such as reliability is not the only option. Instead, he develops a conception that opens a path to speaking meaningfully of trust in technology without personalizing technology or deflating trust into an epistemic construct. In *Explainable AI as a Component of Building Trust. The Case of Regulating Credit Scoring*, **Katja Langenbucher** traces how the concepts of trust, trustworthiness, and transparency entered EU policy discourse and legislation. She then examines their relationship to explainable AI and explains how differing approaches to explanation have added a second layer of complexity. Her main contribution, however, lies in showing how various forms of explanation can enable normative responses to model-based decisions, illustrated through the example of credit scoring. In their contribution *"Trust" and "Trustworthiness"* in the AI Act, **Lucia Franke** and **Benjamin Müller** examine how the notions of trust and trustworthiness are employed in the AI Act and the EU's Ethical Guidelines. By reconstructing the meaning of trust from these documents, they conclude that it is primarily under-

stood as reliability. They argue that this reduced understanding of trust neglects an essential dimension of trust relations – namely, freedom.

References

Friedman, Batya et al. (2013): "Value Sensitive Design and Information Systems", in: Neelke Doorn et al. (eds.): Early Engagement and New Technologies. Opening up the Laboratory. Dordrecht: Springer, pp. 55–95.

Hillerbrand, Rafela (2021): "Value Sensitive Design", in: Armin Grunwald and Rafaela Hillerbrand (eds.): Handbuch Technikethik. Stuttgart: Metzler, pp. 466–471.

Van de Poel, Ibo (2018): Design for Value Change. Ethics and Information Technology 23, pp. 27–31.