Electronic Markets
https://doi.org/10.1007/s12525-020-00441-4

INVITED PAPER

Trustworthy artificial intelligence

Scott Thiebes[1] · Sebastian Lins[1] · Ali Sunyaev[1]

Received: 13 May 2020 / Accepted: 9 September 2020

Abstract

Artificial intelligence (AI) brings forth many opportunities to contribute to the wellbeing of individuals and the advancement of economies and societies, but also a variety of novel ethical, legal, social, and technological challenges. Trustworthy AI (TAI) bases on the idea that trust builds the foundation of societies, economies, and sustainable development, and that individuals, organizations, and societies will therefore only ever be able to realize the full potential of AI, if trust can be established in its development, deployment, and use. With this article we aim to introduce the concept of TAI and its five foundational principles (1) beneficence, (2) non-maleficence, (3) autonomy, (4) justice, and (5) explicability. We further draw on these five principles to develop a data-driven research framework for TAI and demonstrate its utility by delineating fruitful avenues for future research, particularly with regard to the distributed ledger technology-based realization of TAI.

Keywords Trustworthy artificial intelligence · Artificial intelligence · Trust · Framework · Distributed ledger technology · Blockchain

JEL classification M15 O30 A13 C80

Introduction

Artificial intelligence (AI) enables computers to execute tasks that are easy for people to perform but difficult to describe formally (Pandl et al. 2020). It is one of the most-discussed technology trends in research and practice today, and estimated to deliver an additional global economic output of around USD 13 trillion by the year 2030 (Bughin et al. 2018). Although AI has been around and researched for decades, it is especially the recent advances in the subfields of machine learning and deep learning that not only result in manifold opportunities to contribute to the wellbeing of individuals as well as the prosperity and advancement of organizations and societies but, also in a variety of novel ethical, legal, and social challenges that may severely impede AI's value contributions, if not handled appropriately (Floridi 2019; Floridi et al. 2018). Examples of issues that are associated with the rapid development and proliferation of AI are manifold. They range from risks of infringing individuals' privacy (e.g., swapping people's faces in images or videos via DeepFakes (Turton and Martin 2020) or involuntarily tracking individuals over the Internet via the Clearview AI (Hill 2020)), or the presence of racial bias in widely used AI-based systems (Obermeyer et al. 2019), to the rapid and uncontrolled creation of economic losses via autonomous trading agents (e.g., the loss of millions of dollars through erroneous algorithms in high-frequency trading (Harford 2012)).

To maximize the benefits of AI while at the same time mitigating or even preventing its risks and dangers, the concept of trustworthy AI (TAI) promotes the idea that individuals, organizations, and societies will only ever be able to achieve the full potential of AI if trust can be established in its development, deployment, and use (Independent High-Level Expert Group on Artificial Intelligence 2019). If, for

Responsible Editor: Rainer Alt

✉ Ali Sunyaev
sunyaev@kit.edu

Scott Thiebes
scott.thiebes@kit.edu

Sebastian Lins
sebastian.lins@kit.edu

[1] Department of Economics and Management, Karlsruhe Institute of Technology, Institute AIFB - Building 05.20, KIT-Campus South, 76128 Karlsruhe, Germany

Published online: 01 October 2020

example, neither physicians nor patients trust an AI-based system's diagnoses or treatment recommendations, it is unlikely that either of them will follow the recommendations, even if the treatments may increase the patients' well-being. Similarly, if neither drivers nor the general public trust autonomous cars, they will never replace common, manually steered cars, even if it is suggested that completely autonomous traffic might reduce congestion or help avoiding accidents (Condliffe 2017). However, the importance of TAI is not limited to areas like health care or autonomous driving but extends to other areas as well. Electronic markets, for example, are increasingly augmented with AI-based systems such as customer service chatbots (Adam et al. 2020). Likewise, several cloud providers recently began offering 'AI as a Service' (AIaaS), referring to web services for organizations and individuals interested in training, building, and deploying AI-based systems (Dakkak et al. 2019; Rai et al. 2019). Although cost- and time-saving opportunities have triggered a widespread implementation of AI-based systems and services in electronic markets, trust persists to play a pivotal role in any buyer-seller relationship (Bauer et al. 2019; Marella et al. 2020). Consequently, TAI is of increasing relevance to electronic markets and its research community.

Prevalent research on achieving TAI not only covers AI-related research domains like ethical computing, AI ethics, or human-computer interaction but also cuts many cognate research areas such as information systems (IS), marketing, and management that have focused on achieving trust in electronic markets and the role of trust in technology adoption for decades. Today, researchers in areas related to TAI have already created a vast body of knowledge on certain aspects of TAI. There are, for example, currently more than 60 high-level guidelines for the development and deployment of ethical AI (Jobin et al. 2019). Similarly, explainable AI is a topic of heightened interest within research, aiming to achieve transparency such that the results of an AI can be better understood by human experts (Adadi and Berrada 2018). Overall, TAI is a highly interdisciplinary and dynamic field of research, with knowledge on technical and non-technical means to realize TAI being scattered across research disciplines, thus making it challenging to grasp the status quo on its realization.

With this article, we aim to contribute to the ongoing debates around the importance of TAI and provide guidance to those who are interested in engaging with this increasingly important concept. To do so, we first highlight the need for TAI, review extant trust conceptualizations in the IS domain, and introduce the TAI concept, including a definition as well as the five TAI principles beneficence, non-maleficence, autonomy, justice, and explicability. Afterward, we draw on an abstract AI co-creation process and the five outlined TAI principles to develop a data-driven research framework for TAI (named *DaRe4TAI*). This framework outlines tensions between the current state of AI and the five TAI principles to inform future research opportunities on technical and non-technical means in support of realizing TAI. We then demonstrate the framework's utility on the example of delineating fruitful avenues for future research. In particular, we examine the realization of TAI based on distributed ledger technology (DLT) because the unique combination of DLT's inherent characteristics (e.g., tamper resistance, transparency, and confidentiality) present it as a promising technical means to address several, albeit not all, of the prevalent tensions inherent in the TAI principles. Finally, we end this article with a brief conclusion.

Toward a definition of trustworthy artificial intelligence

The need for trustworthy artificial intelligence

Since the term 'artificial intelligence' was conceived at a workshop at Dartmouth College in 1956 (John et al. 2006), the field has experienced several waves of rapid progress (Haenlein and Kaplan 2019). Especially the ground-breaking advances in the subfields of machine learning and deep learning that have been made since the early 2010s and the increasing rate at which those advances are made, have fueled people's imagination of a reality interspersed with intelligent agents contributing to the wellbeing and prosperity of individuals, organizations, and societies. However, it is becoming increasingly evident that AI is not the *'magic bullet'* some would like to believe it is and that AI, just like any other technology, will not only bring forth many benefits but will also be accompanied with a variety of novel ethical, legal, and social challenges (Floridi 2019; Floridi et al. 2018). In response to the growing awareness of the challenges that are induced by AI, we have seen multiple calls for *beneficial AI* (Future of Life Institute 2017), *responsible AI* (Chinese National Governance Committee for the New Generation Artificial Intelligence 2019; Université de Montréal 2017; Wiens et al. 2019)*,* or *ethical AI* (Floridi et al. 2018; UK House of Lords 2017) during the last few years. Irrespective of the exact terminology, all of these calls refer to essentially the same objectives, namely, the advancement of AI such that its benefits are maximized while its risks and dangers are mitigated or prevented. Likewise, the independent High-Level Expert Group on Artificial Intelligence of the European Commission published its Ethics Guidelines for Trustworthy AI in early 2019. These guidelines have quickly gained traction in research and practice and have laid the foundation for the adoption of the term *trustworthy AI* in other guidelines and frameworks like the OECD principles on AI (OECD 2019) or the White House AI principles (Vought 2020).

In its essence, TAI is based on the idea that trust builds the foundation of societies, economies, and sustainable development, and that therefore the global society will only ever be able to realize the full potential of AI if trust can be established in it (Independent High-Level Expert Group on Artificial Intelligence 2019). Yet, TAI is a highly interdisciplinary and dynamic field of research, comprising multifarious research discussions and streams that are scattered across disciplines, including psychology, sociology, economics, management, computer science, and IS. Opinions and interpretations about what makes AI trustworthy vary, preconditions and (ethical and regulatory) requirements that have to be fulfilled are unequally prioritized across the globe, and knowledge on technical and non-technical means to realize TAI is ever-increasing. Considering that *'trust'* in general is a complex phenomenon that has sparked many scholarly debates in recent decades, it is not surprising that the conceptualization of trust in AI and what makes AI trustworthy-as of today-remains inconclusive and highly discussed in research and practice. Grasping the status quo on a definition of TAI and its realization thus remains challenging.

Extant trust conceptualizations

Trust is a complex phenomenon that has sparked many scholarly debates from researchers of diverse disciplines, including psychology, sociology, economics, management, computer science, and IS. In its basic notion, trust is commonly defined as an individual's willingness to depend on another party because the individual lacks (total) control over the other party, thereby creating potential for opportunistic behavior of the trusted party (Mayer et al. 1995). In such situations, individuals must willingly put themselves at risk or in vulnerable positions by delegating responsibility for actions to another (trusted) party (J. D. Lee and See 2004). Nevertheless, various perspectives on trust exist in literature, comprising different dimensions and (partially opposing) interpretations (McKnight et al. 2002). Moreover, trust develops over time as trust relationships evolve, starting with initial trust where an individual has no prior experience with the other party, which then further develops to knowledge-based trust, where the individual knows the other party well enough to predict the party's behavior in a situation (Lewicki and Bunker 1996; McKnight et al. 2011; Paul and McDaniel Jr 2004). As a result of the plurality of perspectives on this concept, there is no commonly accepted definition of trust (Lansing and Sunyaev 2016; Söllner et al. 2016) but rather a need for contextualized trust conceptualizations (Jarvenpaa et al. 2004).

Trust plays a particularly important role in almost any IS-enabled situation in which either uncertainty prevails or undesirable outcomes are possible (McKnight et al. 2011). Most IS research nowadays employs a dualistic perspective on trust (see Table 1). First, trust in a specific person or organization (a moral and volitional agent) (Lankton et al. 2015; McKnight et al. 2011), such as trust in an e-vendor (Gefen et al. 2003) or virtual team members (Robert et al. 2009). Second, trust in a specific technology or IT artefact (lacking volition and moral agency) (Lankton et al. 2015; McKnight et al. 2011), such as trusting an online shopping platform (Vance et al. 2008) or a cloud service (Lansing and Sunyaev 2016). Both types of trust are highly relevant in the context of AI. For example, organizations need to trust providers of AI-based systems, to deploy reliable AI-based systems (e.g., in the form of AIaaS), to not exploit contractual loopholes, and to process data confidentially. Likewise, the organization also needs to trust in the underlying technology itself, like trusting

Table 1 Overview of common trusting beliefs related to persons and technologies

Trust in persons (e.g., Mayer et al. 1995; McKnight et al. 2002).	Trust in technology	
	Trust in IT artifacts based on system characteristics (e.g., McKnight et al. 2011; Thatcher et al. 2010)	*Trust in automation technology and autonomous systems* (e.g., J. D. Lee and See 2004)
Competence / Ability: One has the ability to do for the other person what the other person needs to have done (McKnight et al. 2002). Group of skills, competencies, and characteristics that enable a party to have influence within some specific domain (Mayer et al. 1995).	**Functionality:** The belief that the specific technology has the capability, functionality, or features to do for one what one needs to be done.	**Performance:** The competency or expertise as demonstrated by the automation's ability to achieve the operator's goals.
Benevolence: One cares about the welfare of the other person and is therefore motivated to act in the other person's interest, does not act opportunistically toward the other.	**Helpfulness:** The belief that the specific technology provides adequate and responsive help for users.	**Purpose:** The degree to which the automation is being used within the realm of the designer's intent.
Integrity: The extent to which a trustee adheres to a set of principles that the trustor finds acceptable.	**Reliability / Predictability:** The belief that the specific technology will consistently operate properly (McKnight et al. 2011) and its behavior can be forecast (Thatcher et al. 2010).	**Process:** The degree to which the automation's algorithms are appropriate for the situation and able to achieve the operator's goals.

that a delployed AI-based system itself functions as expected, handles failures adequately, and ensures effective recovery.

Specific trust in people and trust in technology not only differ in terms of the nature of the object of dependence but also on important trusting beliefs. Interpersonal trusting beliefs reflect judgments that the other party has suitable attributes and motives for performing as expected in a risky situation (Mayer et al. 1995), whereas technology-related trust necessarily reflects beliefs about a technology's characteristics rather than its motives (McKnight et al. 2011). Extant research has commonly agreed that individuals express expectations about a person's competence (i.e., its ability to do what the individual needs), benevolence (i.e., its care and motivation to act in the individual's interests), and integrity (i.e., its honesty and promise-keeping) (McKnight et al. 2002). In contrast, individuals' trust in technology commonly concerns the technology's functionality (i.e., providing features needed to complete a task), its helpfulness (i.e., help functions will provide necessary advice), and its reliability (i.e., technology will consistently operate properly) (McKnight et al. 2011; Thatcher et al. 2010). Nevertheless, these different trusting beliefs are highly related, as for example, the competence of a person and the functionality of a technology represent individuals' expectations about their capability (McKnight et al. 2011).

Specific trust in persons and technology can be further positioned in a nomological net of trust, comprising general trusting beliefs and institution-based trust as antecedents and trusting intentions as consequence of specific trust in persons or technology (McKnight et al. 2011; McKnight et al. 2002; see Fig. 1). General trusting beliefs typically comprise an individual's propensity to trust people or technology (i.e., the general tendency to be willing to depend on technology across a broad spectrum of situations and technologies), faith in humanity or general technology (i.e., one assumes technologies are usually consistent, reliable, functional, and provide the help needed), and trusting stance toward people or technology (i.e., regardless of what one assumes about technology generally, one presumes that one will achieve better outcomes by assuming the technology can be relied on) (McKnight et al. 2011). Institution-based trust as a structural concept and further antecedent of specific trust refers to the belief that success is likely because of supportive situations and structures tied to a specific context (Gefen et al. 2003; McKnight et al. 2011). Institution-based trust is composed of situational normality (i.e., a belief that when a situation is viewed as normal and well-ordered, one can extend trust to something new in the situation) and structural assurance (i.e., a belief that adequate support exists, such as legal, contractual, or physical, to ensure success). The trust literature suggests a causal ordering among trust constructs, such that an individual's general trusting beliefs in people or technology directly influences institution-based trust and indirectly shapes trust in a specific person or technology (McKnight et al. 2011; McKnight and Chervany 2001). Moreover, trust in a specific person or technology has an impact on an individuals' trusting intentions, referring to an individual's intention to engage in trust-related behavior, such as sharing personal information, making a purchase, using a system or acting on information provided by a website (McKnight et al. 2002).

While all of these different types of trust spanning up the nomological net are applicable and relevant in the context of AI, in this study we ground our work on specific trust in persons and technology and their respective trusting beliefs. In particular, we strive for a thorough contextualization of specific trust in AI-based systems given the unique characteristics of AI, namely, its human-like and autonomous behavior.

AI-based systems' autonomous and intelligence-based capabilities allow them to have a great degree of self-governance, which enables them to respond to situations that were not pre-programmed or explicitly anticipated during their development, and to make independent decisions and action selection with little or no control by their users (Parasuraman et al. 2000). In general, autonomous systems are generative and learn, evolve and permanently change their functional capacities as a result of the input of operational and contextual information (Hancock 2017). AI-based systems' actions necessarily become more indeterminate across time and are thus more challenging to predict (Hancock 2017), making trust interactions between humans and AI-based systems more complex and difficult to understand than trust interactions between humans and non-AI technologies. Related research has shown that trust in a technology which is perceived as human-like by its user or is highly automated and autonomous, differs from classical interpersonal trust and also classical trust in technology (Lansing and Sunyaev 2016). For example, extant research has frequently focused on recommendation agents (e.g., Al-Natour et al. 2011; Benbasat and Wang 2005) or websites (e.g., Ray et al. 2011; Vance et al. 2008) as IT artifacts with a high degree of humanness (i.e., they have the form or characteristics of humans; Lankton et al. 2015). It, thus, seems reasonable for users to associate human-like trusting beliefs with an online recommendation agent that has voice and animation as in Benbasat and Wang (2005).

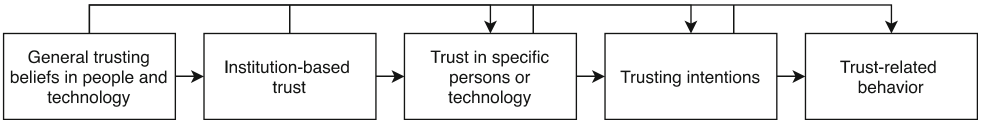

General trusting beliefs in people and technology
Institution-based trust
Trust in specific persons or technology
Trusting intentions
Trust-related behavior

Fig. 1 Simplified nomological net of trust (adapted from McKnight et al. 2011; McKnight et al. 2002)

To understand trust relationships in such scenarios, two research streams emerged that either adapted the concept of interpersonal trust to conceptualize trust in human-like IT artifacts (Lankton et al. 2015), or revised the concept of trust in technology to account for automation technology and autonomous systems (J. D. Lee and See 2004). The adaptation of interpersonal trust is rooted in theories of social responses toward computing (Thatcher et al. 2013). The underpinning assumption of this approach is that intelligent IT artifacts have moral agency (e.g., may behave benevolently) and, as such, have the capacity to act in the best interest of the user, for example, by offering better or worse advice (Thatcher et al. 2013). Contrarily, related research on trust in automated and autonomous systems takes another perspective and has developed three trusting beliefs, namely, performance, process, and purpose (J. D. Lee and See 2004; see Table 1). Performance thereby refers to the current and historical operation of an automated system and includes characteristics such as reliability, predictability, and ability. Process relates to the degree to which an automated system's algorithms are appropriate for the situation and are able to achieve the user's goals. Purpose refers to the degree to which an automated system is being used within the realm of the designer's intent. These trust in automation beliefs therefore take a more technical focus, yet they still relate to prevalent beliefs of trust in human-like IT artifacts. Purpose, for example, corresponds to helpfulness and benevolence and reflects the perception that an automated system has a positive orientation toward the user.

These recent adaptations of trust in specific technology contexts inform our conceptualization of trust in AI-based systems, since such systems are human-like and autonomous. For example, an AI-based system might autonomously learn from available input data, detect certain patterns and make inferences, which then result in the system acting human-like. Such a decision might (in the worst case) treat a person less favorably, evoking feelings of unfairness in this person and reducing their trust into the AI-based system. In the following, we therefore build on extant trust conceptualizations, and particularly, integrate both lenses on specific trust in technology to describe trust in AI-based systems.

Definition and principles of trustworthy artificial intelligence

For this article, we propose that AI is perceived as trustworthy by its users (e.g., consumers, organizations, society) when it is developed, deployed, and used in ways that not only ensure its compliance with all relevant laws and its robustness but especially its adherence to general ethical principles (Independent High-Level Expert Group on Artificial Intelligence 2019).

Several frameworks and guidelines that promote (ethical) principles for TAI have been developed and published by researchers, industry, and policymakers in the recent past. Table 2 summarizes key aspects of a non-exhaustive list of important frameworks and guidelines related to TAI. For a comprehensive comparison, we refer interested readers to Hagendorff (2020). In particular, we adopt the five principles of ethical AI (henceforth TAI principles) beneficence, non-maleficence, autonomy, justice, and explicability of Floridi et al. (2018), which have to be fulfilled by an AI-based system to be perceived as trustworthy. These five principles not only synthesize various pertinent frameworks and guidelines but are also particularly relevant for electronic markets because they reflect a socio-technical view, emphasizing the interaction between people and technology that is needed to realize TAI. In the following, we outline the five principles as well as their relation to TAI in more detail and offer a brief overview of past research efforts related to each principle. Table 3 provides a description of each principle, their relation to existing trusting beliefs, and an overview of which principles are included in the discussed frameworks and guidelines.

Beneficence

Beneficence refers to the development, deployment, and use of AI that is beneficial to humanity and the planet in the sense that it promotes the well-being of humans and the environment, and respects basic human rights (Floridi et al. 2018). Although beneficence is found in all of the frameworks and guidelines discussed here, it is taken into account to varying degrees. While, for example, some of the proposed frameworks and guidelines focus this principle on the well-being of humanity (i.e., Asilomar AI principles, UK AI Code), others extend it to all sentient beings and even the environment (i.e., Montreal Declaration, AI4People, EU TAI Guidelines, OECD Principles on AI). Moreover, the Chinese AI principles further extend this principle to the need for harmony, whereas the White House AI Principles do not directly list beneficence as a key principle but instead state that "AI is expected to have a positive impact across sectors of social and economic life" (Vought 2020) and that US agencies should "[…] carefully consider the full societal costs, benefits, and distributional effects before considering regulations related to the development and deployment of AI applications" (Vought 2020). The beneficence principle aligns with the trusting beliefs benevolence, helpfulness, and purpose since AI-based systems that fulfill this principle should in general act in the users' best interest, try to help or achieve certain benefits while being genuinely concerned, and not acting opportunistically or manipulatively (McKnight et al. 2002).

Research related to the beneficence principle mostly stems from the areas of ethical computing and AI ethics, which focus on discussing foundational ethical themes (e.g., general ethics frameworks) (Floridi 2019; Floridi and Cowls 2019; Floridi et al. 2018; Hagendorff 2020) and how to embed values that

Table 2 Overview of key aspects of pertinent frameworks and guidelines for TAI

Framework/guidelines	Issued by *(in)*	Terminology	Description
Asilomar AI Principles	Future of Life Institute (2017)	Beneficial AI	Describes 23 principles of beneficial AI. The principles are organized into three categories: research issues, ethics and values, and long-term issues.
Montreal Declaration of Responsible AI (Montreal Declaration)	Université de Montréal (2017)	Responsible AI	Provides ten ethical principles that promote the fundamental interests of people and groups and, based on these, eight recommendations for the development of responsible AI.
UK AI Code	UK House of Lords (2017)	Ethical AI	Defines five overarching principles for an ethical AI code, intended to position the UK as a future leader in AI.
AI4People	Floridi et al. (2018)	Ethical AI	A synthesis of six pertinent frameworks and guidelines, which resulted in five foundational principles for ethical AI. Based on the principles, a set of 20 action points in the four categories assessment, development, incentivization, and support is proposed.
Ethics Guidelines for Trustworthy AI (EU TAI Guidelines)	European Commission (Independent High-Level Expert Group on Artificial Intelligence 2019)	Trustworthy AI	Defines four principles of TAI and based on these derives seven key requirements for achieving TAI. Further provides an assessment list for the operationalization of the seven key requirements.
OECD Principles on AI	OECD (2019)	Trustworthy AI	Recommends “five complementary values-based principles for the responsible stewardship of trustworthy AI” (OECD 2019). In addition to the OECD member states, other countries (e.g., Argentina, Brazil, Colombia, Costa Rica, Peru, and Romania) have signed up to follow the OECD principles.
Governance Principles for the New Generation Artificial Intelligence (Chinese AI Principles)	Chinese National Governance Committee for the New Generation Artificial Intelligence (2019)	Responsible AI	Provides a framework and action guidelines for the governance of AI, based on eight principles for the development of responsible AI.
White House AI Principles	White House’s Office of Science and Technology Policy (Vought 2020)	Trustworthy AI	Defines ten principles for stewardship of AI applications and the development of trustworthy AI. These principles are to be considered by US agencies during the development of regulatory and non-regulatory actions on AI.

promote wellbeing into AI at the design and development stages (de Swarte et al. 2019). From an IS perspective, the beneficence principle demands organizations to consider, for example, the environment (e.g., being sustainable and environmentally friendly when using computing resources to deploy AI) as well as the societal impact of AI services and products offered (e.g., embedding AI-based chatbots that truly support consumers instead of only gathering further consumer data).

Non-maleficence

Non-maleficence advocates the development, deployment, and use of AI such that it avoids bringing harm to people (Floridi et al. 2018). Although similar to beneficence, which emphasizes the creation of AI that actively acts towards the wellbeing of humanity, non-maleficence represents a distinct principle that represents a key aspect of all considered frameworks and guidelines. Non-maleficence especially concerns the protection of people’s privacy (expressed by the Asilomar AI Principles, Montreal Declaration, UK AI Code, AI4People, EU TAI Guidelines, Chinese AI Principles) and security (expressed by the AI4People, EU TAI Guidelines, OECD Principles on AI, White House AI Principles), as well as their safety (expressed by the Asilomar AI Principles, UK AI Code, AI4People, EU TAI Guidelines, OECD Principles on AI, Chinese AI Principles, White House AI Principles). An interesting facet of this principle’s safety aspect thereby revolves around artificial general intelligence (i.e., computer programs that can control themselves and solve tasks in a variety of different domains) and how we can ensure that

Table 3 Relation of TAI principles to existing trusting beliefs and the discussed TAI frameworks and guidelines

TAI principle	Description	Relation to existing trusting beliefs	Frameworks / guidelines							
			Asilomar AI Principles	Montreal Declaration	UK AI Code	AI4People	EU TAI Guidelines	OECD Principles on AI	Chinese AI Principles	White House AI Principles
Beneficence	Beneficence refers to the development, deployment, and use of AI that is beneficial to humanity in the sense that it promotes the well-being of humans and respects basic human rights.	• Benevolence (P), • Helpfulness (T), • Purpose (A)	✓	✓	✓	✓	✓	✓	✓	○
Non-maleficence	Non-maleficence advocates the development, deployment, and use of AI in a way that avoids bringing harm to people.	• Integrity (P), • Reliability (T), • Process (A)	✓	✓	✓	✓	✓	✓	✓	✓
Autonomy	Autonomy focuses on the promotion of human autonomy, agency, and oversight, and therefore may include the restriction of AI-based systems' autonomy, where necessary.	–	✓	✓	✓	✓	✓	✓	○	○
Justice	Justice encompasses the utilization of AI to amend past inequities, the creation of shareable and subsequent distribution of benefits through AI, and thwarting the creation of new harms and inequities by AI.	• Integrity (P), • Reliability (T), • Process (A)	✓	✓	✓	✓	✓	✓	✓	✓
Explicability	Explicability entails the development, deployment, and use of explainable AI by producing (more) interpretable AI models whilst maintaining high levels of performance and accuracy. In its ethical sense, explicability further comprises the creation of accountable AI.	• Competence (P), • Functionality (T), • Performance (A)	✓	✓	✓	✓	✓	✓	✓	○

Note: ✓ *= key aspect of the framework/guideline,* ○ *= mentioned in the framework/guideline*

P = relating to interpersonal trusting beliefs, T = relating to trusting beliefs in technology, A = relating to trusting beliefs in automation and autonomous systems artificial general intelligence, once it becomes a reality, behaves in a non-harmful a way (Goertzel 2014). Non-maleficence relates to the trusting beliefs integrity, reliability, and process because it requires AI-based systems to act honestly and consistently, and to sincerely adhere to ethical and other pre-defined principles.

Extant research has proposed several approaches to protect people's privacy during the training and operation of an AI, like adding noise to data and models (Sarwate and Chaudhuri 2013), the use of trusted execution environments (Tramer and Boneh 2019), or federated learning for AI model training (Smith et al. 2017). While past research related to the non-maleficence principle especially investigated means for the development and deployment of safe and secure AI in the areas of autonomous driving (Koopman and Wagner 2017) and medicine (Wiens et al. 2019), the non-maleficence principle is highly important for electronic markets due to the exchange and analysis of highly sensitive consumer and intellectual property data. For example, organizations offering AIaaS must implement adequate data governance and protection mechanisms such that collected as well as AI-generated data about individuals is not used in a way that impedes their privacy and such that users are enabled to better understand the consequences of data disclosure.

Autonomy

Autonomy is the third TAI principle. Given that extant TAI frameworks and guidelines provide slightly different understandings of this principle, it lacks a precise definition. While some mainly focus on the promotion of human autonomy, agency, and oversight (e.g., EU TAI Guidelines), others also consider the restriction of AI-based systems' autonomy, where necessary (e.g., the Montreal Declaration) (Floridi and Cowls 2019). Floridi et al. (2018) refer to this as meta-autonomy and humans retaining the right to decide when to decide at any given time. Only two guidelines do not directly address the need for autonomy, The Chinese AI Principles abstractly refer to the need for 'controllability', stating that "controllability of AI systems should be improved continuously" (Chinese National Governance Committee for the New Generation Artificial Intelligence 2019) but do not further discuss their understanding of this concept. Similarly, the White House AI Principles use autonomy to motivate several other principles, stating that AI may impede or contribute to human autonomy, but do not explicitly refer to autonomy as a key principle in itself. The autonomy principle is not directly related to extant trusting beliefs but reflects a means to mitigate integrity and reliability risks by balancing between human- and machine-led decision-making. In addition, autonomy aligns with openness, a sub dimension of the process belief of automation technologies (J. D. Lee and See 2004), that refers to the willingness to give and receive ideas, which will increase trust into another party (Mishra 1992; Schindler and Thomas 1993).

Research on AI autonomy is diverse and involves, for example, the autonomy of robots (Noorman and Johnson 2014), human-robot interactions (Goodrich and Schultz 2007), or the coordination of several autonomous agents (Yan et al. 2013). Of particular concern in relation to this principle is research on trust in autonomous systems such as autonomous vehicles (Schaefer et al. 2016; Stormont 2008), as well as research on adjustable autonomy, which refers to agents dynamically changing their autonomy and transferring it to other entities (Mostafa et al. 2019). For organizations, this principle implies that they should, for example, consider implementing proper oversight mechanisms (e.g., keeping the human-in-the-loop) to ensure autonomy when embedding AI into their electronic services and products.

Justice

Like non-maleficence, *justice* is as key aspect of all eight frameworks and guidelines discussed in this article, albeit it is also referred to as fairness by some. Justice is not to be understood judicially, as in adhering to laws and regulations, but instead in an ethical way (Floridi and Cowls 2019). As such, all frameworks and guidelines exhibit similar but slightly distinctive views on justice, which can be summarized as (1) the utilization of AI to amend past inequities like discrimination, (2) the creation of shareable and subsequent distribution of benefits through AI, and (3) thwarting the creation of new harms and inequities by AI (Floridi et al. 2018). Regarding the utilization of AI to amend past inequities, for example, the White House AI Principles state that US agencies should consider "[…] whether the AI application at issue may reduce levels of unlawful, unfair, or otherwise unintended discrimination as compared to existing processes" (Vought 2020). The Asilomar AI Principles on the other hand express the need for 'Shared Benefit' and 'Shared Prosperity', thus emphasizing the creation of shareable and subsequent distribution of benefits. An example for avoiding the creation of new harms and inequities can be found in the 'Equity' principle of the Montreal Declaration, which reads as "[t]he development and use of [AI] must contribute to the creation of a just and equitable society" (Université de Montréal 2017). Similar to non-maleficence, justice aligns with the trusting beliefs integrity, reliability, and process, ensuring that ethical principles are fulfilled by an AI-based system.

Justice in its various shapes is an important aspect of contemporary AI research. Central research themes concerning the justice principle are, for instance, identifying the presence of racial and other biases in current AI-based systems (Mehrabi et al. 2019), means for quantifying the fairness or absence thereof in AI-based systems (Bellamy et al. 2019), and approaches for mitigating or even avoiding bias in AI-based systems (Mehrabi et al. 2019). Similar to most of the other TAI principles, much of the current research relating to the justice principle is conducted in medical contexts. Nevertheless, the justice principle is also highly relevant for electronic markets as, for example, AI-based product recommendations may be disturbed by popularity biases, where popular products would be presented more to the public, while such a recommendation may not be a result of good quality (Mehrabi et al. 2019).

Explicability

Explicability is the fifth and last TAI principle. According to Floridi et al. (2018), explicability comprises an epistemological sense as well as an ethical sense. In its epistemological sense, explicability entails the creation of explainable AI by producing (more) interpretable AI models whilst maintaining high levels of performance and accuracy. In its ethical sense, explicability comprises the creation of accountable AI. Within the eight frameworks and guidelines considered in this work, explicability can be found under different terms and to varying degrees. The Asilomar AI Principles and the UK AI Code, for example, convey this principle by formulating the need for transparent AI and intelligibility of AI, respectively. Similarly, the EU TAI Guidelines and the OECD Principles on AI call for transparent and accountable AI, whereas the Chinese AI Principles call for the continuous improvement of the transparency, interpretability, reliability, and controllability of AI. The White House AI Principles, on the other hand, refer to transparency and accountability within several of their ten principles but do not explicitly state both as a requirement for TAI. Explicability relates also to the trusting beliefs competence, functionality, and performance in the sense that explainable and interpretable AI proves that it has the capability, functionality, or features to do what needs to be done. Thus, an individual will tend to trust the AI if its algorithms can be understood and seem capable of achieving the individual's goals in the current situation.

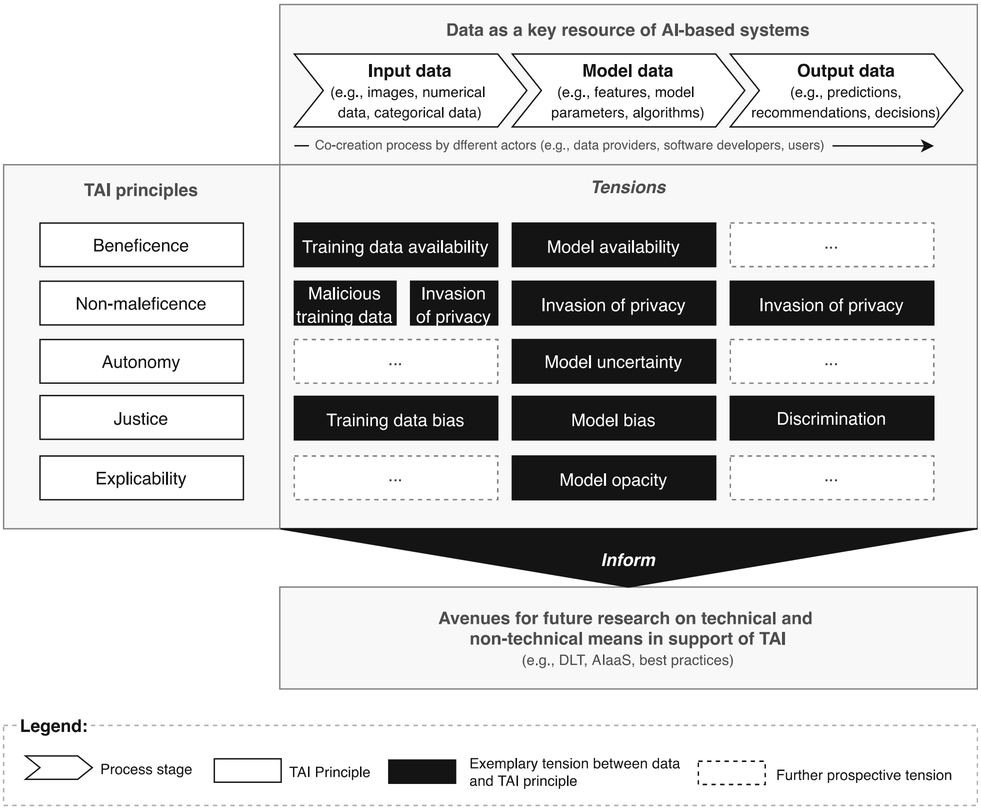
Explicability, in its two meanings, is perhaps the most prevalent theme in contemporary AI research. A central reason for this lies in the fact that today's AI-based systems are complex systems that mostly function as black boxes and therefore suffer from opacity and a lack of accountability. Their sub-symbolic representation of state is often inaccessible and non-transparent to humans, thus limiting individuals in fully understanding and trusting the produced outputs. Floridi et al. (2018) consider explicability an enabling principle for TAI, as it augments the four previously discussed principles. Toward this end, "[one] must be able to understand the good or harm [AI] is actually doing to society, and in which ways" (Floridi and Cowls 2019) for it to be beneficent and non-maleficent. Likewise, we must be able to anticipate an AI's predictions and decisions to make informed decisions about the degree of autonomy we attribute to that AI, and must also ensure accountability to hold someone legally responsible in case of an AI failure, thus supporting the justice principle.

Extant research efforts on explainable AI can be divided into research focusing on the creation of transparent and interpretable models (e.g., via decision trees, rule-based learning, or Bayesian models) and research focusing on establishing post-hoc explainability (e.g., via heat maps, or backpropagation) (Barredo Arrieta et al. 2020). Another prominent stream of research concerned with the explainability of AI encompasses the quantification of uncertainties (Begoli et al. 2019). Furthermore, there are also first research efforts in the direction of auditing AI (e.g., Cremers et al. 2019). In the IS domain, explicability of AI is of major importance since it will not only allow organizations to meet compliance requirements when employing AI (e.g., by means of enabling independent third-party audits) but will also be a key driver for acceptance of AI by managers, the general workforce, and consumers (Hagras 2018; Rai 2020).

Limitations of extant principles, frameworks, and guidelines

Despite their value for a realization of TAI, the outlined principles and the corresponding frameworks and guidelines also exhibit two major limitations. First, as noted in the EU TAI Guidelines, several TAI principles may at times conflict with each other. Take, for example, the beneficence and justice principles. Extant research shows that AI can be employed for purposes of predictive policing (i.e., using mathematical models to forecast what crimes will happen when and where) and therefore benefit society by allowing for a better allocation of police staff and reducing crime rates (Courtland 2018). However, ethnicity and other socio-demographic characteristics are often-used data in the training of AI models for predictive policing. Training AI models on the grounds of such characteristics induces a form of discrimination, essentially violating the justice principle. Depending on the specific application cases, the conflicts between certain TAI principles are inherent to those principles and therefore difficult or even impossible to fully resolve without making trade-offs. We leave a discussion of such trade-offs to ethics and legal experts and instead focus on another limitation for the remainder of this article. The second major limitation of the outlined TAI principles concerns the fact that they are highly general and that extant frameworks and guidelines provide little to no guidance for how they can or should be transferred into practice, nor how they can inform future research on technical and non-technical means in support of a realization of TAI. In this article, we attempt to address this limitation by

Fig. 2 Data-driven research framework for TAI (DaRe4TAI)


Data as a key resource of AI-based systems
Input data
(e.g., images, numerical data, categorical data)
Model data
(e.g., features, model parameters, algorithms)
Output data
(e.g., predictions, recommendations, decisions)
Co-creation process by dfferent actors (e.g., data providers, software developers, users)
TAI principles
Beneficence
Non-maleficence
Autonomy
Justice
Explicability
Tensions
Training data availability
Model availability
...
Malicious training data
Invasion of privacy
Invasion of privacy
Invasion of privacy
...
Model uncertainty
...
Training data bias
Model bias
Discrimination
...
Model opacity
...
Inform
Avenues for future research on technical and non-technical means in support of TAI
(e.g., DLT, AIaaS, best practices)
Legend:
Process stage
TAI Principle
Exemplary tension between data and TAI principle
Further prospective tension

presenting a data-driven research framework for TAI in the following.

DaRe4TAI – A data-driven research framework for trustworthy artificial intelligence

Framework overview

While there exist various approaches to create AI-based systems, most systems that are developed, deployed, and in use today rely on machine learning, or for that matter, deep learning methods. They are thus based on an abstract AI co-creation process, comprising three important stages: input, modeling, and output. Throughout the process, different actors (e.g., data providers, software developers, users) co-create value by transforming large amounts of input data (e.g., images, numerical data, categorical data) into output data (e.g., predictions, recommendations, decisions) via the design, training, and subsequent application of AI models. Besides input and output data, the AI models themselves thereby also constitute an important form of data that is being generated during the systems' design and training.

Drawing on this abstract AI co-creation process, and in line with calls for more research that treats data as a key resource of IS (Selz 2020), the guiding notion for the development of our data-driven research framework (*DaRe4TAI*; Fig. 2) is the idea that data in its various forms (i.e., input data, model data, output data) and functions (i.e., for training or for inference) represents the central, single most important resource for AI-based systems. At the same time, the nascent stream of data ethics tells us that data in itself can be a source of manifold ethical problems (Floridi and Taddeo 2016). By analyzing how the different actors in the co-creation process interact with each other at any of the three process stages (i.e., input, model, output), through the *collection, curation, generation, analysis, and use* of data, we can identify tensions between the current state of AI development, deployment, and use and the five outlined TAI principles beneficence, non-maleficence, autonomy, justice, and explicability. These tensions, which form the backbone of *DaRe4TAI*, should thereby not to be understood from a negative point of view, as in contributing to mistrust in AI, but instead as illustrations of challenges that need to be overcome in certain scenarios to achieve TAI from a user (i.e., consumers, organizations, society) perspective. Thus, they inform future research opportunities on technical and non-technical means in support of realizing TAI. In the following, we briefly outline exemplary tensions that we identified for data at each stage of the AI co-creation process and the five TAI principles.

Input stage

Input data plays an important dual role in AI-based systems. On the one hand, it serves as the main resource for the training of AI models.[1] On the other hand, input data is transferred into output data via a trained AI model, once an AI has been deployed. Within our framework, this dualistic role of input data may lead to the emergence of several tensions between the different forms of input data and the TAI principles.

Next to advances in machine learning and deep learning methods, the growing availability of training data represents a fundamental reason for the recent advancements of AI (Pandl et al. 2020). However, the creation of high-quality training data is costly and time-consuming, especially when expert knowledge is required (e.g., labeling of thousands of medical images). As a result, large, high-quality data sets are either under the control of a few large enterprises or, in the case of freely available high-quality data sets, are limited to a few specific application areas (e.g., a certain disease). We identify this situation as creating a tension between input data and the beneficence principle (*tension: training data availability*). In line with the view of institutions-based trust in technology being an antecedent to trust in a specific technology (McKnight et al. 2011), we argue, that the limited availability of large, high-quality data in certain areas, which constraints our ability to promote human well-being through AI in those areas, could lead society to perceive the entire class of AI-based systems as not beneficent (enough). Nevertheless, this does not necessarily imply that all data should always be freely available, but instead calls for technical and non-technical means to create large, high-quality data sets and enable their availability (proprietary or open access) in areas that are particularly beneficial to society (e.g., medicine).

Feeding low-quality or even malicious input data into an AI model's training process, on the other hand, bears the risk that the AI-based systems themselves might behave unintendedly or even maliciously. Microsoft, for example, released the AI-based chatbot *Tay* on Twitter that caused subsequent controversy when the bot began posting inflammatory and offensive tweets through its Twitter account, causing Microsoft to shut down the service only 16 h after its launch (P. Lee 2016). We discern this as a tension between input data and the non-maleficence TAI principle (*tension: malicious training data*), since unintended or malicious behavior of an AI-based system will most likely exert a negative impact on users' trust in the AI.

AI possesses the inherent ability to infringe people's privacy. Even if an AI-based system's overall purpose is beneficent and of interest for its users, their trust in such an AI-based system might still derogate if their data is involuntarily used for purposes of training or inference. Contemporary smart speakers, for example, have repeatedly been criticized for constantly eavesdropping on their users to train the underlying AI (e.g., Amazon's Alexa, Apple's Siri, or Google's Assistant). Consequently, the potential for privacy infringements concerning input data represents yet another tension between input data and the non-maleficence principle (*tension: invasion of privacy*).

Next to the limited availability of high-quality training data, training data that are already available are routinely biased toward certain groups of people, which in the past has led to the implicit discrimination of those groups of people that are underrepresented in the training data (Bellamy et al. 2019). Obermeyer et al. (2019), for example, found that a widely used AI in US hospitals is negatively biased toward Black patients, since it uses health costs spent on patients as a proxy for health needs and since on average less money is spent on Black patients in the US. Such training data bias creates tension between input data and the TAI principle of justice, which demands fairness and the avoidance of discrimination (*tension: training data bias*).

Model stage

AI models are responsible for translating input data into output data. In line with our guiding notion that data is the single, most important resource for contemporary AI-based systems, we argue that AI models themselves constitute an important form of data and identify several tensions between the model and the five TAI principles.

Similar to input data, the development and training of an AI model is an expensive and time-consuming task. As a form of intellectual property, AI models increasingly represent an important factor in achieving competitive advantages (Haenlein and Kaplan 2019; Makridakis 2017). Attempts to protect competitive advantages can thereby contribute to the fact that particularly promising AI models are not shared and that AI as a specific class of technology are perceived as not beneficent (enough) by the society (i.e., the whole of AI-based systems not acting in societies best interest). We argue that, analog to the limited availability of training input data, this creates a tension between model data and the beneficence principle because the potential for contributing to human well-being is not being fully realized for these AI models (*tension: model availability*). Again, we stress that this tension does not necessarily imply that all AI models have to be freely available to everyone, but that it instead calls for technical (e.g., pre-trained models in AIaaS) and non-technical means (e.g., licensing models) to make promising AI models more widely available where they can be highly beneficial to society.

Extant research has further shown, that under certain circumstances, parameters of AI models can be analyzed to generate insights about the underlying training data (Shokri et al.

[1] Note: we also consider the data labels needed for some training approaches as input data.

2017). In extreme cases, such insights could be used to identify individuals who contributed their data, which in turn represents a privacy infringement that could undermine those very individuals' trust in AI-based systems. We, thus, also see a tension between model data and the non-maleficence principle (*tension: invasion of privacy*).

Inferences made by AI models are associated with some uncertainty. Although there exist first approaches in research and practice to quantify such uncertainties, these approaches are often still in their infancy and are not broadly available for all use cases (Begoli et al. 2019). However, being able to adequately quantify the uncertainties in AI models is a fundamental aspect in deciding how much autonomy should be given to an AI-based system. Users' inability to adequately quantify uncertainties of AI models, therefore, creates a tension between model data and the autonomy principle (*tension: model uncertainty*).

Current AI-based systems routinely contain socially constructed biases (Obermeyer et al. 2019). Next to the bias in training input data, another source of bias is the overemphasis of certain aspects (e.g., skin color or place of residence) by developers of AI models during the design of an AI model. Considering, for instance, the above example of an AI-based system widely used in US hospitals again, the bias cannot only be found in the training data itself (i.e., on average less is money is spent on Black patients) but also in the fact that such obviously biased data was chosen as a major feature for the model, without correcting for it. Similar to the previously described bias in input data, we therefore see this bias in AI models as creating a tension between model data and the justice principle (*tension: model bias*).

Lastly, the opacity of most current AI models is one of the most popular topics of contemporary AI research (Rudin 2019). Despite extensive efforts that are being directed toward tackling this issue and creating so-called explainable AI, we still lack the ability to fully understand the inner functioning of most AI models, especially those constructed using deep learning. Not only does this impede the interpretability of output data but also obstruct establishing accountability. As such, we view model opacity as creating a tension between model data and the explicability principle (*tension: model opacity*).

Output stage

Output data is the last form of data in our framework. It is generated by applying input data to a previously trained model. We identify two exemplary tensions between output data and the introduced TAI principles.

First, similar to input data and model data before, output data that is being generated by AI-based systems can represent an infringement of people's privacy. However, in contrast to the invasion of privacy tension on the input stage and the model stage, where the AI-based system's purpose might have been benevolent but the collection, processing, and analysis of the users' data without their consent stipulated a privacy infringement, the privacy invasion on the output stage occurs simply due to the fact that the AI's purpose is malicious and an invasion of privacy. In early 2020, for example, it was reported that an AI-based system of the New York City-based firm Clearview would be able to construct complete profiles of individuals from data publicly available on the Internet, which quickly raised suspicion and fear of 'the end of privacy' (Hill 2020). Another, perhaps more extreme, example concerns the recent upcoming of so-called DeepFakes. Although the AI behind DeepFakes could potentially be used to do good or at least to do no harm (e.g., replacing actors' faces in movies with faces of their younger selves), it was primarily used for a malicious purpose, namely the creation of adult films with faces of celebrities. In addition to the obvious privacy infringement of using those celebrities' images without their consent (i.e., an invasion of privacy on the input stage), the very nature of the output in itself constitutes a privacy infringement. In the long run, such acts undermine users' trust in those AI-based systems, which is why we identify a tension between output data and the non-maleficence principle (*tension: invasion of privacy*).

Second, sensitive output data could not only be used to invade people's privacy but also to discriminate against them. Toward this end, AI-based systems have, for example, been shown to be able to infer individuals' political views or sexual orientation based on the Facebook pages that they liked (Gibney 2018) or predicted the mental state of Facebook users based on an analysis of their posts (Goggin 2019). Again, despite the fact that such AI-based systems could as well be used to do good, it is their malicious use (here discrimination) or possibly even the inherently malicious purpose with which a system was designed and developed, that could eventually undermine users' trust not only in other users or developers of the AI but also in the AI-based system itself. We therefore also discern a tension between data at the output stage and the justice principle (*tension: discrimination*).

Future research on the distributed ledger technology-based realization of trustworthy artificial intelligence

To demonstrate the utility of *DaRe4TAI*, this section focuses on deriving fruitful avenues for future research on a technical means to realize TAI, namely DLT. Our focus on DLT is thereby grounded in two observations. First, DLT allows for the operation of a highly available, append-only, peer-to-peer database (i.e., a distributed ledger) in situations where uncertainty prevails and undesirable outcomes are possible (Zhang and Jacobsen 2018). It enables the coordination of economic activity through the creation of secure, transparent, and decentralized electronic markets (Berg et al. 2019; Kollmann et al. 2019; Subramanian 2017), and is probably best known under the name of blockchain, which is a specific type of DLT (Kannengiesser et al. 2020; Sunyaev 2020). Second, there is a nascent stream of literature combining DLT with AI to address diverse issues of current AI-based systems (Pandl et al. 2020). The unique combination of DLT's inherent characteristics (e.g., tamper resistance, transparency, and confidentiality) presents it as a promising technical means to address several, albeit not all, of the aforementioned tensions between data at the input and model stages and the outlined TAI principles. Especially DLT-based data markets, DLT-based federated learning, and DLT-based transparency, accountability, and explainability are fruitful avenues for further research to better address these tensions and eventually realize TAI. In the following, we briefly discuss each avenue concerning the tensions that it might help address and derive exemplary research questions (Table 4).

DLT-based data markets

DLT-based data markets are a popular stream of research that focuses on the use of tokens to securely, efficiently, and inexpensively trade valuable data online through distributed ledgers. They are of particular interest for the electronic markets community and could serve as a means to address several tensions at the input and model stages.

For example, DLT-based data markets provide the ability to create economic incentives, which could not only stimulate the democratization of access to extant, high-quality AI training data (i.e., addressing the *training data availability* tension) but as well encourage greater participation by the general public to drive the generation of new, more diverse data sets (i.e., addressing the *training data bias* tension). However, despite first technical solutions being developed by researchers from the IS, computer science, and related disciplines (Ozercan et al. 2018; Özyilmaz et al. 2018; Xiong and Xiong 2019; Zhao et al. 2019), the question of how to effectively design token economies (e.g., to democratize data access or to encourage the generation of more diverse data sets) remains a focal theme of contemporary DLT research. Adding to this, several researchers have raised concerns over the potential consequences of over-emphasizing economic incentives for the sharing of personal data because they could especially motivate those in need to share their data and without making

Table 4 Fruitful avenues of future research on the DLT-based realization of TAI, related tensions, and exemplary research questions

Avenue	Addressable tensions *(stage)*	Potential future research questions
DLT-based data markets	Training data availability *(input stage)*	• How can DLT be used to democratize access to high-quality training data to increase the beneficence of AI?
	Training data bias *(input stage)*	• How to design a token economy such that it is effective in stimulating public participation and the generation of more diverse AI training data? • What are the potential negative consequences of a token economy that could interfere with the realization of TAI and how can they be prevented?
	Model availability *(model stage)*	• How can DLT be used to democratize access to high-quality AI models to increase the beneficence of those models? • How can AI-related assets (e.g., training data, model data, algorithms) be modeled as tokens?
DLT-based federated learning	Invasion of privacy *(input stage)*	• What is the performance overhead of DLT-based federated learning for complex AI models? • How to improve the efficiency of DLT-based federated learning in real-world application scenarios? • How does DLT-based federated learning affect data providers' privacy concerns and trust in data processors? • How to design a token economy such that it is effective in stimulating participation in federated learning networks?
	Invasion of privacy *(model stage)*	• How can DLT be employed to prevent inference attacks in federated learning networks?
DLT-based transparency, accountability, and explainability	Malicious training data *(input stage)*	• How can DLT support the continuous auditing of training data provenance?
	Model uncertainty *(model stage)*	• How can DLT-based continuous auditing aid in the (real-time) quantification of model uncertainties?
	Model opacity *(model stage)*	• How can tamper resistant trails of the data flows within AI-based systems stored on distributed ledgers support the attainment of explainable AI?

informed decisions (Thiebes et al. 2020). The potential negative consequences of a token economy that could conflict with the realization of TAI, therefore, warrant further research. Lastly, the potential of a token economy for TAI extends beyond the creation of data markets for input training data to the trading and licensing of other AI-related assets such as models or algorithms (i.e., addressing the *model availability* tension) (Sarpatwar et al. 2019). Yet, analog to the question of how to design token economies capable of effectively democratizing access to input training data, the design of effective token economies for democratizing access to AI models requires further research. Furthermore, the modeling of assets as tokens is a topic of ongoing research (Kim and Chung 2018; Laskowski et al. 2019) and, thus far, we lack knowledge on what and how AI-related assets (e.g., training data, model data, algorithms) can be modeled and represented adequately as tokens.

DLT-based federated learning

Next to DLT-based data markets, DLT can also serve to organize the federated (i.e., decentralized) training of AI models (Dinh and Thai 2018). In such a federated learning scenario, no input training data is directly shared. Instead, partial AI models are being trained by nodes participating in the federated learning network, while training data provenance and the integrity of the partial AI models are preserved using a distributed ledger (Pandl et al. 2020). Since no training data is directly shared among participants of the network, DLT-based federated learning seems to be particularly auspicious for addressing the *invasion of privacy* tensions at the input and model stages. However, there remain several issues that warrant further research before we will be able to deploy DLT-based federated learning in real-world use cases.

Most research prototypes, for example, employ DLT-based federated learning to train relatively simple AI models (Pandl et al. 2020; Preuveneers et al. 2018), while extant research indicates that DLT-based federated learning induces a performance overhead of 5% to 15% (Preuveneers et al. 2018). Although this might at first not seem like a large overhead, it could ultimately render DLT-based federated learning prohibitively expensive for more complex AI models. Future research should thus seek to explore the application of DLT-based federated learning to more complex AI models and investigate ways to reduce the induced performance overhead in real-world application scenarios. Furthermore, despite increased confidentiality, research has also shown that federated learning is potentially vulnerable to inference attacks, whereby an adversary can aim to extract information about private training data by inferring the AI model multiple times (Melis et al. 2019; Wang et al. 2019). In addition to employing DLT for preserving training data provenance and AI model integrity, future research should therefore also explore how DLT could help with preventing inference attacks on federated learning networks.

Despite technical questions, several non-technical questions require further research. It is, for example, not clear whether the promises of increased privacy due to the application of (DLT-based) federated learning may actually strengthen data providers' trust in an AI-based system's ability to adequately protect their data and ultimately their willingness to contribute their data for purposes of training AI models. Moreover, and similar to the previously described token economy for AI training data, DLT can provide a ledger for incentivizing participation in federated learning networks. However, also similar to DLT-based data markets, we still lack substantive knowledge on how to successfully design such a token economy for DLT-based federated learning.

DLT-based transparency, accountability, and explainability

The last avenue of DLT-related research on the realization of TAI that we discuss in this article concerns achieving accountability and explainability of AI through DLT.

A central facet of establishing AI accountability concerns our ability to independently audit AI (i.e., AI's auditability), especially in terms of data provenance (i.e., addressing the *malicious training data* tension) and the degree of uncertainty with which AI models make their predictions (i.e., addressing the *model uncertainty* tension). Owing to DLTs characteristics (e.g., decentralization, high tamper resistance), research has recently begun exploring the application of DLT for auditing purposes in organizational contexts (Hofman et al. 2019), while first research results also indicate the feasibility of DLT for the auditing of AI (Dillenberger et al. 2019). However, the development and deployment of AI are highly dynamic, with training data and algorithms (and thus model uncertainty) rapidly changing and constantly evolving. Effective auditing of AI, therefore, does not only warrant creating an independent, tamper-resistant audit trail, but also the continuous updating and assessment of this audit trail using continuous auditing procedures (Lins et al. 2019). Toward this end, future research should explore how, on the one hand, DLT can support the continuous auditing of training data provenance, and how, on the other hand, DLT-based continuous auditing can aid in the (real-time) quantification of model uncertainties.

Finally, extant research has proposed the use of DLT to establish explainability of AI (i.e., addressing the *model opacity* tension). DLT is thereby ought to serve as a tamper resistant trail for tracking the flow of data within AI-based systems, which may then be further analyzed to create explainable AI models (Dinh and Thai 2018). However, looking at the recent literature, we see that the concept of using DLT to create explainable AI is at the idea stage at most and that it remains unknown how DLT can support the attainment of explainable AI. Toward this end, future research should seek to move DLT-based explainable AI beyond the idea stage and explore means for how tamper resistant trails of the data flows within AI-based systems, stored on distributed ledgers can actively support the attainment of explainable AI.

Conclusion

In this article, we introduced the concept of TAI as a promising research topic for IS research, delineated its background, positioned it in related trust conceptualizations, and contextualized the five TAI principles beneficence, non-maleficence, autonomy, justice, and explicability to the IS context. Further, we drew on a data-driven perspective toward AI to develop the research framework *DaRe4TAI* that provides guidance to those enticed to study technical and non-technical means in support of TAI, and demonstrated its feasibility on the example of fruitful avenues for future research on the DLT-based realization of TAI. In doing so, we highlight a vast space of TAI research opportunities for the IS and other research communities that is not limited to the recent AI hype topic of explainability. Especially for the field of electronic markets, TAI provides several promising avenues of future research, including and beyond its DLT-based realization.

The tensions between data at the different stages of the AI co-creation process and the five TAI principles that we outlined here represent only a subset of tensions. Nevertheless, we are convinced that *DaRe4TAI* provides a good starting ground for exploring further tensions and, thus, revealing additional avenues for future research on technical and non-technical means in support of TAI.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Adam, M., Wessel, M., & Benlian, A. (2020). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 1–19. https://doi.org/10.1007/s12525-020-00414-7.

Al-Natour, S., Benbasat, I., & Cenfetelli, R. (2011). The adoption of online shopping assistants: Perceived similarity as an antecedent to evaluative beliefs. *Journal of the Association for Information Systems, 12*(5), 347–374. https://doi.org/10.17705/1jais.00267.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bauer, I., Zavolokina, L., & Schwabe, G. (2019). Is there a market for trusted car data? *Electronic Markets*, 1–15. https://doi.org/10.1007/s12525-019-00368-5.

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence, 1*, 20–23. https://doi.org/10.1038/s42256-018-0004-1.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., et al. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development, 63*(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287.

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems, 6*(3), 72–101. https://doi.org/10.17705/1jais.00065.

Berg, C., Davidson, S., & Potts, J. (2019). Blockchain technology as economic infrastructure: Revisiting the electronic markets hypothesis. *Frontiers in Blockchain, 2*(22), 1–6. https://doi.org/10.3389/fbloc.2019.00022.

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute, Brussels, San Francisco, Shanghai, Stockholm. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20frontier%20Modeling%20the%20impact%20of%20AI%20on%20the%20world%20economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-economy-September-2018.ashx

Chinese National Governance Committee for the New Generation Artificial Intelligence. (2019). Governance Principles for the New Generation Artificial Intelligence–Developing Responsible Artificial Intelligence. Retrieved from https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html

Condliffe, J. (2017). A single autonomous Car has a huge impact on alleviating traffic. MIT technology review. Retrieved from https://www.technologyreview.com/s/607841/a-single-autonomous-car-has-a-huge-impact-on-alleviating-traffic/

Courtland, R. (2018). Bias detectives: The researchers striving to make algorithms fair. *Nature, 558*(7710), 357–357. https://doi.org/10.1038/d41586-018-05469-3.

Cremers, A, B., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., … Wrobel, S. (2019). Trustworthy use of artificial intelligence. Priorities From a Philosophical, Ethical, Legal, and Technological Viewpoint as a Basis for Certification of Artificial Intelligence. Retrieved from https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf

Dakkak, A., Li, C., Gonzalo, S, G, D., Xiong, J., & Hwu, W. (2019). TrIMS: Transparent and isolated model sharing for low latency deep learning inference in function-as-a-service. Paper presented at the 12th IEEE international conference on cloud computing (CLOUD), Milan, Italy

de Swarte, T., Boufous, O., & Escalle, P. (2019). Artificial intelligence, ethics and human values: The cases of military drones and companion robots. *Artificial Life and Robotics, 24*(3), 291–296. https://doi.org/10.1007/s10015-019-00525-1.

Dillenberger, D. N., Novotny, P., Zhang, Q., Jayachandran, P., Gupta, H., Hans, S., et al. (2019). Blockchain analytics and artificial intelligence. *IBM Journal of Research and Development, 63*(2/3), 5:1–5:14. https://doi.org/10.1147/JRD.2019.2900638.

Dinh, T. N., & Thai, M. T. (2018). AI and Blockchain: A disruptive integration. *Computer, 51*(9), 48–53. https://doi.org/10.1109/MC.2018.3620971.

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence, 1*(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1), 1–15. https://doi.org/10.1162/99608f92.8cd550d1.

Floridi, L., & Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical. *Physical and Engineering Sciences, 374*(2083), 1–5. https://doi.org/10.1098/rsta.2016.0360.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Future of Life Institute. (2017). Asilomar AI Princples. Retrieved from https://futureoflife.org/ai-principles/

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly, 27*(1), 51–90. https://doi.org/10.2307/30036519.

Gibney, E. (2018). The scant science behind Cambridge Analytica's controversial marketing techniques. Nature news explainer. Retrieved from https://www.nature.com/articles/d41586-018-03880-4

Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence, 5*(1), 1–48. https://doi.org/10.2478/jagi-2014-0001.

Goggin, B. (2019). Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts. Retrieved from https://www.businessinsider.com/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12

Goodrich, M. A., & Schultz, A. C. (2007). Human–robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction, 1*(3), 203–275. https://doi.org/10.1561/1100000005.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review, 61*(4), 5–14. https://doi.org/10.1177/0008125619864925.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8.

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer, 51*(9), 28–36. https://doi.org/10.1109/MC.2018.3620965.

Hancock, P. A. (2017). Imposing limits on autonomous systems. *Ergonomics, 60*(2), 284–291. https://doi.org/10.1080/00140139.2016.1190035.

Harford, T. (2012). High-frequency trading and the $440m mistake. Retrieved from https://www.bbc.com/news/magazine-19214294

Hill, K. (2020). The secretive company that might end privacy as we know it. The New York times. Retrieved from https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

Hofman, D., Lemieux, V., Joo, A., & Batista, D. (2019). "The margin between the edge of the world and infinite possibility": Blockchain, GDPR and information governance. *Records Management Journal, 29*(1/2), 240–257. https://doi.org/10.1108/RMJ-12-2018-0045.

Independent High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. Brussels: European Commission Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

Jarvenpaa, S. L., Shaw, T. R., & Staples, D. S. (2004). Toward contextualized theories of trust: The role of trust in global virtual teams. *Information Systems Research, 15*(3), 250–267. https://doi.org/10.1287/isre.1040.0028.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

John, M., Marvin, L. M., Nathaniel, R., & Claude, E. S. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine, 27*(4), 12–14. https://doi.org/10.1609/aimag.v27i4.1904.

Kannengiesser, N., Lins, S., Dehling, T., & Sunyaev, A. (2020). Mind the gap: Trade-offs between distributed ledger technology characteristics. *ACM Computing Surveys, 53*(2), 1–37. https://doi.org/10.1145/3379463.

Kim, M. S., & Chung, J. Y. (2018). Sustainable growth and token economy design: The case of Steemit. *Sustainability, 11*(1), 167–178. https://doi.org/10.3390/su11010167.

Kollmann, T., Hensellek, S., de Cruppe, K., & Sirges, A. (2019). Toward a renaissance of cooperatives fostered by Blockchain on electronic marketplaces: A theory-driven case study approach. *Electronic Markets*, 1–12. https://doi.org/10.1007/s12525-019-00369-4.

Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine, 9*(1), 90–96. https://doi.org/10.1109/MITS.2016.2583491.

Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems, 16*(10), 880–918. https://doi.org/10.17705/1jais.00411.

Lansing, J., & Sunyaev, A. (2016). Trust in cloud computing: Conceptual typology and trust-building antecedents. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems, 47*(2), 58–96. https://doi.org/10.1145/2963175.2963179.

Laskowski, M., Kim, H, M., Zargham, M., Barlin, M., & Kabanov, D. (2019). Token economics in real-life: Cryptocurrency and incentives Design for Insolar Blockchain Network. *arXiv e-prints*, 1–20. arXiv: 1910.02064.

Lee, P. (2016). Learning from Tay's introduction. Retrieved from https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Lewicki, R. J., & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 114–139). Columbus, OH: Sage Publications Inc..

Lins, S., Schneider, S., Szefer, J., Ibraheem, S., & Sunyaev, A. (2019). Designing monitoring systems for continuous certification of cloud services: Deriving meta-requirements and design guidelines. *Communications of the Association for Information Systems, 44*(25), 1–52. https://doi.org/10.17705/1CAIS.04425.

Makridakis, S. (2017). The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms. *Futures, 90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006.

Marella, V., Upreti, B., Merikivi, J., & Tuunainen, V. K. (2020). Understanding the creation of trust in cryptocurrencies: The case of Bitcoin. *Electronic Markets, 30*, 1–13. https://doi.org/10.1007/s12525-019-00392-5.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335.

McKnight, D. H., & Chervany, N. L. (2001). What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International Journal of Electronic Commerce, 6*(2), 35–59. https://doi.org/10.1080/10864415.2001.11044235.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS), 2*(2), 1–25. https://doi.org/10.1145/1985347.1985353.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv e-prints*. arXiv:1908.09635.

Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019). Exploiting Unintended Feature Leakage in Collaborative Learning. Paper presented at the 2019 IEEE symposium on security and privacy (S&P), San Francisco, CA, USA.

Mishra, A, K. (1992). Organizational responses to crisis: The role of mutual trust and top management teams.

Mostafa, S. A., Ahmad, M. S., & Mustapha, A. (2019). Adjustable autonomy: A systematic literature review. *Artificial Intelligence Review, 51*(2), 149–186. https://doi.org/10.1007/s10462-017-9560-8.

Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology, 16*(1), 51–62. https://doi.org/10.1007/s10676-013-9335-0.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.

OECD (2019). OECD Principles on AI. Retrieved from https://www.oecd.org/going-digital/ai/principles/

Ozercan, H. I., Ileri, A. M., Ayday, E., & Alkan, C. (2018). Realizing the potential of blockchain technologies in genomics. *Genome Research, 28*(9), 1255–1263. https://doi.org/10.1101/gr.207464.116.

Özyilmaz, K. R., Doğan, M., & Yurdakul, A. (2018). IDMoB: IoT data marketplace on blockchain. Paper presented at the 2018 Crypto Valley conference on Blockchain technology (CVCBT), Zug, Switzerland.

Pandl, K. D., Thiebes, S., Schmidt-Kraepelin, M., & Sunyaev, A. (2020). On the convergence of artificial intelligence and distributed ledger technology: A scoping review and future research agenda. *IEEE Access, 8*, 57075–57095. https://doi.org/10.1109/ACCESS.2020.2981447.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 30*(3), 286–297. https://doi.org/10.1109/3468.844354.

Paul, D. L., & McDaniel Jr., R. R. (2004). A field study of the effect of interpersonal trust on virtual collaborative relationship performance. *MIS Quarterly, 28*(2), 183–227. https://doi.org/10.2307/25148633.

Preuveneers, D., Rimmer, V., Tsingenopoulos, I., Spooren, J., Joosen, W., & Ilie-Zudor, E. (2018). Chained anomaly detection models for federated learning: An intrusion detection case study. *Applied Sciences, 8*(12), 2663–2684. https://doi.org/10.3390/app8122663.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5.

Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's comments: Next-generation digital platforms: Toward human–AI hybrids. *MIS Quarterly*, 43(1), iii-x. https://doi.org/10.5555/3370135.3370136.

Ray, S., Ow, T., & Kim, S. S. (2011). Security assurance: How online service providers can influence security control perceptions and gain trust. *Decision Sciences, 42*(2), 391–412. https://doi.org/10.1111/j.1540-5915.2011.00316.x.

Robert, L. P., Denis, A. R., & Hung, Y.-T. C. (2009). Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *Journal of Management Information Systems, 26*(2), 241–279. https://doi.org/10.2753/MIS0742-1222260210.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Sarpatwar, K., Vaculin, R., Min, H., Su, G., Heath, T., Ganapavarapu, G., & Dillenberger, D. (2019). Towards enabling trusted artificial intelligence via Blockchain. In S. Calo, E. Bertino, & D. Verma (Eds.), *Policy-based autonomic data governance* (pp. 137–153). Cham: Springer International Publishing.

Sarwate, A. D., & Chaudhuri, K. (2013). Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Processing Magazine, 30*(5), 86–94. https://doi.org/10.1109/MSP.2013.2259911.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377–400. https://doi.org/10.1177/0018720816634228.

Schindler, P. L., & Thomas, C. C. (1993). The structure of interpersonal trust in the workplace. *Psychological Reports, 73*(2), 563–573. https://doi.org/10.2466/pr0.1993.73.2.563.

Selz, D. (2020). From electronic markets to data driven insights. *Electronic Markets, 30*, 1–3. https://doi.org/10.1007/s12525-019-00393-4.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. Paper presented at the 2017 IEEE symposium on security and privacy (S&P), San Jose, CA, USA.

Smith, V., Chiang, C, K., Sanjabi, M., & Talwalkar, A. S. (2017). Federated multi-task learning. Paper presented at the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA,

Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter for information systems users. *European Journal of Information Systems, 25*(3), 274–287. https://doi.org/10.1057/ejis.2015.17.

Stormont, D, P. (2008). Analyzing human trust of autonomous systems in hazardous environments. Paper presented at the Human Implications of Human-Robot Interaction workshop at AAAI, Menlo Park, CA, USA.

Subramanian, H. (2017). Decentralized blockchain-based electronic marketplaces. *Communications of the ACM, 61*(1), 78–84. https://doi.org/10.1145/3158333.

Sunyaev, A. (2020). Internet Computing: Principles of Distributed Systems and Emerging Internet-based Technologies. Springer Nature.

Thatcher, J. B., McKnight, D. H., Baker, E. W., Arsal, R. E., & Roberts, N. H. (2010). The role of trust in postadoption IT exploration: An empirical examination of knowledge management systems. *IEEE Transactions on Engineering Management, 58*(1), 56–70. https://doi.org/10.1109/TEM.2009.2028320.

Thatcher, J. B., Carter, M., Li, X., & Rong, G. (2013). A classification and investigation of trustees in B-to-C e-commerce: General vs. specific trust. *Communications of the Association for Information Systems, 32*(1), 107–134. https://doi.org/10.17705/1CAIS.03204.

Thiebes, S., Schlesner, M., Brors, B., & Sunyaev, A. (2020). Distributed ledger technology in genomics: A call for Europe. *European*

Journal of Human Genetics, 28(2), 139–140. https://doi.org/10.1038/s41431-019-0512-4.

Tramer, F., & Boneh, D. (2019). Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. Paper presented at the International Conference on Learning Representations, New Orleans, LA

Turton, W., & Martin, A. (2020). How Deepfakes Make Disinformation More Real Than Ever. Retrieved from https://www.bloomberg.com/news/articles/2020-01-06/how-deepfakes-make-disinformation-more-real-than-ever-quicktake

UK House of Lords. (2017). AI in the UK: ready, willing and able? Retrieved from https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm

Université de Montréal. (2017). Montreal Declaration for a Responsible Development of AI. Retrieved from https://www.montrealdeclaration-responsibleai.com/the-declaration

Vance, A., Elie-Dit-Cosaque, C., & Straub, D. W. (2008). Examining trust in information technology artifacts: The effects of system quality and culture. *Journal of Management Information Systems, 24*(4), 73–100. https://doi.org/10.2753/MIS0742-1222240403.

Vought, R, T. (2020). Guidance for Regulation of Artificial Intelligence Applications Retrieved from https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., & Qi, H. (2019). Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning. Paper presented at the 2019 IEEE conference on computer communications (IEEE INFOCOM 2019), Paris, France.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., et al. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine, 25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6.

Xiong, W., & Xiong, L. (2019). Smart contract based data trading mode using blockchain and machine learning. *IEEE Access, 7*, 102331–102344. https://doi.org/10.1109/ACCESS.2019.2928325.

Yan, Z., Jouandeau, N., & Cherif, A. A. (2013). A survey and analysis of multi-robot coordination. *International Journal of Advanced Robotic Systems, 10*(12), 399–416. https://doi.org/10.5772/57313.

Zhang, K., & Jacobsen, H. (2018). Towards dependable, scalable, and pervasive distributed ledgers with Blockchains. Paper presented at the IEEE 38th international conference on distributed computing systems (ICDCS), Vienna, Austria.

Zhao, Y., Yu, Y., Li, Y., Han, G., & Du, X. (2019). Machine learning based privacy-preserving fair data trading in big data market. *Information Sciences, 478*, 449–460. https://doi.org/10.1016/j.ins.2018.11.028.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.