Trustworthy Artificial Intelligence for Trusted Decision-Making

Kristina Sargsyan
French University in Armenia
Yerevan, Armenia
e-mail: Kristina.Sargsyan@pers.ufar.am

Abstract— **Organizations are increasingly introducing data science initiatives to support decision-making. However, the decision outcomes of data science initiatives are not always used or adopted by decision-makers, often due to uncertainty about the quality of data input. It is, therefore, not surprising that organizations are increasingly turning to data governance as a means to improve the acceptance of data science decision outcomes. In this paper, propositions will be developed to understand the role of data governance in creating trust in data science decision outcomes. The duality of technology is used as our theoretical lens to understand the interactions between the organization, decision-makers, and technology. The results show that data science decision outcomes are more likely to be accepted if the organization has an established data governance capability. Data governance is also needed to ensure that organizational conditions of data science are met, and that incurred organizational changes are managed efficiently. These results imply that a mature data governance capability is required before sufficient trust can be placed in data science decision outcomes for decision-making.**

Keywords— Data lake; data governance; data quality; big data; digital transformation; data science; asset management; boundary condition

I. INTRODUCTION

Data trust means having confidence that your organization's data is healthy and ready to act on.

Trust is the key to making successful use of your data. By ensuring trust in corporate data, an organization provides its teams the ability to design exceptional customer experiences, improve operations, ensure compliance, and drive innovation. But data trust must be earned and quantified. It can't be taken on faith. Before trusting corporate data, you should prove that it can produce reliable analytics to support well-informed business decisions.

We define six dimensions of data quality:

Accuracy: the degree to which data correctly describes the real-world object or event in question

Completeness: the proportion of data stored against the potential for being 100% complete

Consistency: the absence of difference when comparing two or more representations of an item against a definition

Timeliness: the degree to which data is current enough to represent reality as needed to support business functions

Uniqueness: no item, or entity instance, is recorded more than once based upon how that item is identified

Validity or conformity: the degree to which data conforms to the syntax (format, type, or range) of its definition

Bear in mind that data quality is only one dimension of data trust. Analysts also include factors such as reasonability, accessibility, and integrity as important ways to measure organizational data trust. Whatever factors you include, the point is to quantify how usable your data is across the enterprise.

The more highly you can rate the data across each of these dimensions for all tables, records, and fields, the more you can trust it — and the more decision-ready your data will be. Data that performs well in one dimension can't necessarily be 100% trusted. As shown above, you might have information that's valid but not accurate, or accurate but incomplete. It could also be high-quality, but inaccessible.

What matters most will vary depending on the business need. For example, finance teams require a particularly high level of accuracy, while other departments may place a premium on timeliness instead. Data teams must make their own assessments of the metrics that trusted data should meet. They should also quantify that certification of data trust to data users. A combination of trust and transparency gives decision-makers confidence to use the data.

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.

II. Concept Of Thetrustworthy Artificial Intelligence

Artificial intelligence (AI) enables computers to execute tasks that are easy for people to perform but difficult to describe formally [1]. 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 [2]. 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 [3]. 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 [4] or involuntarily tracking individuals over the Internet via the Clearview AI [5]), or the presence of racial bias in widely used AI-based systems [6], 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 [7]).

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 [8]. If, for 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 [9]. 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 [10]. 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 [11]. 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 [12]. 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 [13]. 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 [14]. 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.

III. The Need For Trustworthy Artificial Intelligence

Since the term "artificial intelligence" was conceived at a workshop at Dartmouth College in 1956, the field has experienced several waves of rapid progress. 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. In response to the growing awareness of the challenges that are induced by AI, we have seen multiple calls for *beneficial AI*, *responsible AI,* or *ethical AI* 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 or the White House AI principles.

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. 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.

IV. Explicability And Extant Trust Conceptualizations

Explicability: According to Floridi et al., 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 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" 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). Another prominent stream of research concerned with the explainability of AI encompasses the quantification of uncertainties. Furthermore, there are also first research efforts in the direction of auditing AI. 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.

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. 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.

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. 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. 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. 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. 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. 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*).

V. Conclusion

In this article, we introduced the concept of TAI as a promising research topic for IS research, delineated its background. Further, we drew on a data-driven perspective toward AI to develop the research framework 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.

References

[1] K. D. Pandl, S. Thiebes, M. Schmidt-Kraepelin, and A. Sunyaev. „On the convergence of artificial intelligence and distributed ledger technology: A scoping review and future research agenda." *IEEE Access*, vol. 8, pp. 57075–57095, 2020. https://doi.org/10.1109/ACCESS.2020.

[2] J. Bughin, J. Seong, J. Manyika, M. Chui, and R. Joshi. "Notes from the AI frontier: Modeling the impact of AI on the world economy." *McKinsey Global Institute, Brussels, San Francisco, Shanghai, Stockholm.* , 2018. 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

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

[4] W.Turton and A. Martin "How Deepfakes Make Disinformation More Real Than Ever.", 2020. https://www.bloomberg.com/news/articles/2020-01-06/how-deepfakes-make-disinformation- more-real-than-ever-quicktake

[5] K. Hill "The secretive company that might end privacy as we know it." *The New York times*, 2020. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial- recognition.html

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

[7] T. Harford, "High-frequency trading and the $440m mistake.", 2012. https://www.bbc.com/news/magazine-19214294

[8] "Independent High-Level Expert Group on Artificial Intelligence." *Ethics guidelines for trustworthy AI. Brussels: European Commission* , 2019. https://ec.europa.eu/newsroom/dae/ document.cfm?doc_id=60419

[9] J. Condliffe, "A single autonomous Car has a huge impact on alleviating traffic." *MIT technology review,* 2017. https://www.technologyreview.com/s/607841/a-single-autonomous-car-has-a-huge-impact-on-alleviating-traffic/

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

[11] A. Dakkak, C. Li, S. G. D.Gonzalo, J.Xiong, W. Hwu. "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*, 2019.

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

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

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