Towards a User-Empowering Architecture for Trustability Analytics

Sebastian Bruchhaus[1], Thoralf Reis[2], Marco X. Bornschlegl[3], Uta Störl[4], Matthias Hemmje[5]

Abstract: Machine learning (ML) thrives on big data like huge data sets and streams from Internet of Things (IOT) devices. Those technologies are becoming increasingly commonplace in our day-to-day existence. Learning Autonomous Intelligent Actors (AIAs) impact our lives already in the form of, e.g. chat bots, medical expert systems, and facial recognition systems. Doubts concerning ethical, legal, and social implications of such AIAs consequently become increasingly compelling. Our society now finds itself confronted with decisive questions: Should we trust AI? Is it fair, transparent, and respecting privacy? An individual psychological threshold for cooperation with AIAs has been postulated. In Shaefer's words: "No trust, no use". On the other hand, ignorance of an AIA's weak points and idiosyncracies can lead to overreliance. This paper proposes a prototypical microservice architecture for trustability analytics. Its architecture shall introduce self-awareness concerning trustability into the AI2VIS4BigData reference model for big data analysis and visualization by borrowing the concept of a "looking-glass self" from psychology.

Keywords: Trust; Machine Learning; Digital Humanities; Foundation Model; Transparency; XAI

1 Introduction and Motivation

Individuals in our modern society are arguably compelled to accept the presence of Autonomous Intelligent Actors (AIAs) in their everyday environment. In literature AIAs are also referenced as "AI systems", "artificial agents", "autonomous systems", and sometimes "robots". Applied AI promises tremendous benefits, ranging from such diverse fields like healthcare to meteorology [Rei+22a; Rei+22b; Hig20]. This begs the question how society is going to integrate AIAs. A formal verification of their code is basically not unfeasible in most cases, because their ML models tend to be quite complex. GPT-3 was built in 2020 and has around 175 billion parameters [Bro+20]. On top of that, comprehension of its

[1] FernUniversität, DBIS, Universitätsstr. 1, 58097 Hagen, Germany, sebastian.bruchhaus@fernuni-hagen.de, https://orcid.org/0000-0002-7783-2636
[2] FernUniversität, MMIA, Universitätsstr. 1, 58097 Hagen, Germany, thoralf.reis@fernuni-hagen.de, https://orcid.org/0000-0003-1100-2645
[3] FernUniversität, MMIA, Universitätsstr. 1, 58097 Hagen, Germany, marco-xaver.bornschlegl@fernuni-hagen.de, https://orcid.org/0000-0003-3789-5285
[4] FernUniversität, DBIS, Universitätsstr. 1, 58097 Hagen, Germany, uta.stoerl@fernuni-hagen.de, https://orcid.org/0000-0003-2771-142X
[5] FernUniversität, MMIA, Universitätsstr. 1, 58097 Hagen, Germany, matthias.hemmje@fernuni-hagen.de, https://orcid.org/0000-0001-8293-2802

doi:10.18420/BTW2023-60

training data is key to understanding a model's behavior. In practice such knowledge is often sketchy at best. These properties effectively turn all but the simplest AIAs into black boxes [Rud19]. The so-called "value alignment problem" results from this fundamental uncertainty [Had21]. Risk is a necessary prerequisite of trust [Jac+21]. Common sense forbids overreliance on automated decisions with the penalty of catastrophic consequences [Lif15]. Stakeholders resort to trust when they choose to accept the risks of an AIA although its benevolence or robustness cannot be proven rigorously. Trustworthy AIAs in data analytics and cyber-physical systems will arguably have an empowering effect on their human users [Rei+21]. They will be an important intermediate step towards real digital empathy between humans and AIAs [Bon+19].

1.1 Problem Statement and Research Questions

Trust is a necessity when working with AIAs but it is also a somewhat elusive concept. AI architects and engineers need standardized architectures and best practices that facilitate qualified trust. These architectures will serve as a foundation and yardstick for trustworthy AI software systems. This paper identifies the following challenges from the current scientific debate on trust and AIAs as research problems:

(i) *Trust in AI has no canonical definition and lacks an universal vocabulary.*
Terms like "transparent AI", "explainable AI", and even "responsible AI" or "ethical AI" may have subtly differentiated connotations with different authors. Researchers ought to refrain from using their own ad-hoc definitions. A common framework for trustability analytics needs to be established. There should be an unequivocal language and a sound understanding with rigorous models of trust as a solid foundation for future debate [Jac+21; Mil19; Lip18; Rud19].

(ii) *Stakeholders must still rely on their intuition or educated guessing in order to determine the trustworthiness of an AI system.*
The concept of trust has been described as "diffuse", "disappointing", and even "useless" by authors like O. Williamson [Wer18]. This paper intendeds to demonstrate that the latter verdict is an exaggeration. Trust clearly is an important asset for human society, which empowers us to cooperate and reach otherwise unachievable goals [Luh14].There is, however, not yet a generally accepted metric for trustworthiness of AI. Trust is almost universally described as a highly individual and situational [KCW05]. Therefore it is a desirable feature for AIAs to address users individually and adjust to feedback.

(iii) *There are no actionable guidelines on the practical engineering of trustworthy AIAs.*
Despite of the many guidelines for ethical or trustworthy AI on the one hand, and a growing number of algorithms for ostensibly explainable, robust ML on the other, there is still hardly any practical, systematic advice on the implementation of trustworthy AIAs. Existing solutions seem somewhat insular. AIAs should be able to prove their

adherence to ethical and legal principles in the light of ongoing efforts to regulate ML for critical domains such as healthcare.

Foundation models are an excellent show case for these open problems, because of their immense practical usefulness despite a fundamental opacity [Bom+21]. Three questions shall be addressed in the following:

1. How can trustability of AIAs be practically modeled and analyzed?
2. How can a user empowering AIA maximize its trustability?
3. How can we design systems with regard to trustability?

The goal in answering those questions is to establish qualified trust or trustability in AIAs and an AIA architecture for big data analytics and visualization that proactively anticipates and maximizes its user's level of trust. Therefore it works out an actionable model of trust in section 3.1 after surveying the state of art of explainable AI (XAI) in section 2.3 and digital trust in section 3.2. It lays out a prototypical microservice architecture for trustability analytics in 3.2. This will introduce self-awareness concerning trustability into the AI2VIS4BigData reference model (sec. 2.2) for big data analysis and visualization in analogy to the "looking-glass self" theory from psychology described in 2.1.

2 State of the Art

A fair amount of literature is devoted to trust in AIAs. Yet there is still a shortage of practical research results such as complete and ready for use mathematical models. Stenton and Jensen come close to this with their blueprint model for trust in AI [SJ21]. Abbass lists a number of mathematical models of trust in [ALM16]: statistical models, Bayesian analysis, discrete models, belief models, fuzzy models, flow models, and optimization models. Some of which take the reputation of an AIA into consideration.

2.1 The Looking-Glass Self Theory

The psychologist Cooley developed a theory of an individual's self that he labeled "the looking-glass self". The eponymous "looking-glass" is an archaic term for a mirror. His theory is opposite the "self-verification theory" which states that we want others to see as we see ourselves. The self evolves by forming assumptions about the way that other individuals perceive us according to Cooley. His theory describes the human tendency to understand oneself through the judgements that others supposedly make In short, humans form a mental model of their peers and how they perceive them. Then they adjust their self-view accordingly[McI07]. This happens in a three-step process [Sha04]:

1. Actors imagine how others must perceive them.
2. Actors consider how those others think of them.

3. Actors feel an affective reaction, e.g. pride or shame.

Consider an artist painting a self-portrait using a mirror, e.g. as depicted in the famous self-portraits of Johannes Gumpp and Norman Rockwell [Gos10; Roc60]. The artist (a stand-in for the AIA) uses a mirror (looking glass, the putative user's perspective on the AIA) to paint an image of himself. A point of note here is that the affective reaction spurs a reaction and thus stimulates a feedback loop. This is sometimes is known as the "Michelangelo phenomenon" in sociology. It causes individuals in romantic relationships to adjust their behavior in the direction to their assumed ideal self. In the context of AIAs, users' mental processes in relation to the AIA must be emulated by a mathematical model of trust. Such a user models may be based on actual input data from the users or from first principles. The AIA can then adapt itself according to its supposed trustability score.

2.2 Big Data Analytics and AI2VIS4BigData

Training of ML models usually involves "big data", i.e. data of high volume, variety, and velocity. Such a process tends to be lengthy and highly cost-intensive [Bom+21]. Many of the popular pre-trained foundation models, e.g. GPT-3, DALL-E 2, and BERT [Ram+22; Vas+17] do not enable end-users to inspect their training data. This can lead to underappreciation of algorithmic bias [Meh+21]. Applying the FAIR principles is arguably a step in the right direction [Jac+20], but this alone is insufficient to empower humans to a trustful collaboration with AIAs [Mon+20]. Data engineering for end-to-end machine learning usually happens in ML pipelines.

Reis developed a generic reference model called AI2VIS4BigData for such pipelines [Rei+22a]. It is an extension to the IVIS4BigData framework by Bornschlegl [BH21], using AI both directly and indirectly for analytics and user empowerment, e.g. by recommending analytics algorithms to subject matter experts without a background in data science [Rei+22b]. This abstract and versatile rendition of a generic data analytics workflow makes AI2VIS4BigData a suitable stand-in for a broad class of AIAs. Its focus on visual analysis for big data in combination with a strong connection to AI predisposes it for XAI use cases in particular.

AI2VIS4BigData describes a circular process repeating the following four steps of a big data analytics pipeline: Data Management & Curation, Analytics, Insight & Effectuation, and Interaction & Perception.

2.3 Trustability, Explanability, and Transparency of AI

Trust is primarily an interpersonal phenomenon. It has been studied extensively in psychology, sociology, and philosophy. Kracher gives an overview of the theories about trust in sociology with regard to computer science in [KCW05]. Trust is a prerequisite for human society

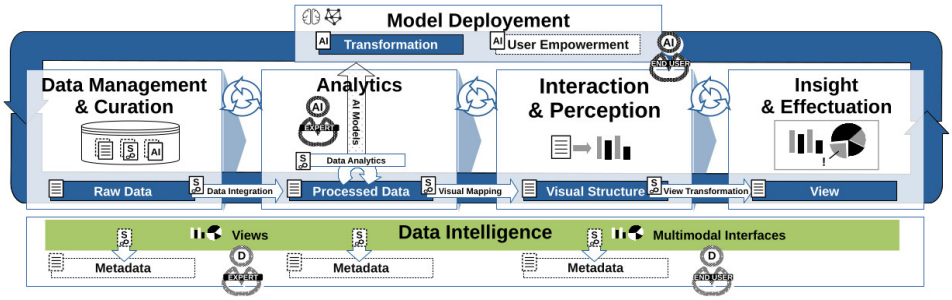


Fig. 1: Reis' AI2VIS4BigData reference model [Rei+22a]

in Niklas Luhman's opinion, as it reduces complexity for the individual [Luh14]. The sources of trust are twofold according to K. Wehrbach [Wer18]: "On one side is a belief rooted in some combination of rational and emotional factors; on the other is acceptance of uncontrolled risk." Three things are essential to trust: a trustor x, a trustee y which may be an inanimate object or AIA, and an element of uncertainty that introduces a risk. Its stakeholders must resort to trust. Hoff and Bashir describe a three-layered model of trust in automation that comprises dispositional, situational, and learned trust [HB14]. Better yet, users ought to be empowered by informed consent. Hence, they should depend on qualified trust that deals with risks transparently. Transparency has been identified as a fundamental prerequisite for trust [HBS11; Jac+21; Rud19; Bon+19]. This insight gave rise to the explainable AI (XAI) branch of AI research [GA19; Bar+20]. It is a truism in the field of explainable AI that understandability begets trust [HBS11; Mil19]. This paper follows the definitions in Barredo Arrieta's paper [Bar+20]: transparent systems are understandable by themselves while explainable systems present explanations as an accurate proxy to their users, etc. Yet the distinction between transparent and non-transparent ML models still lacks a satisfactory differentiation. The field of XAI is a very active research topic [GA19]. Several ostensibly transparent ML algorithms have been developed in recent years [DR20; BB21]. These are complemented by post-hoc explainers like SHAP and LIME for the analysis of black-box models [LL17; RSG16]. As all ML algorithms, these have individual strengths and weaknesses that can affect their trustability. These are hardly assessable – even by users with a background in ML. While it is reasonable that stakeholders are prone to put faith in systems they understand well, this idea should be taken with a grain of salt. There seems to be a trade-off with explanation completeness, as too complete explanations can even discourage users' trust in them [Kul+13; Pap+22]. Sceptics like Rudin and Lipton criticize current XAI methods and post-hoc explainers for black box models in particular [Lip18; Rud19].

3 Model Building

Abbass proposes a trust bus that has a Belief-Desire-Intention-Trust-Motivation (BDI-TM) architecture [ALM16]. The trust bus learns by passing messages between its six components. The constituent modules are these: actors and entities memory, trust production, identity management, intent management, emotion management, risk management, and complexity management. While Abbass' BDI-TM architecture is quite abstract and intended also for interactions between different AIAs, this paper will only consider it in the context of AI2VIS4BigData. Hence it focuses on AIA-user interactions as trustee and trustor. Therefore this paper adapts and simplifies the architecture significantly according to its use case, leading to a prototypical implementation.

3.1 Trustability Analytics

The overall theme of this paper are trustability analytics for AI. Even if a satisfactory model for trust can be found, it is still not quite clear what the terms "trustability of AI" and by extension "trustability of AIA" exactly mean. In order to find an answer for research question 1, i.e. an actionable definition of trust, trustability analytics must be delineated. In the following, this paper will follow McCarthy's working definition [McC07]: "Intelligence is the computational part of the ability to achieve goals in the world.", considering also the ISO's definition of risk as the "effect of uncertainty on an objective" in this context [ISO18]. Without a rigorous definition of AI the subsequent definition of its trustability, this paper must, however, remain preliminary.

AIAs are complex, emergent systems prone to unintended behavior [SY19]. The time-tested practice of software verification for critical settings guarantees that a program's execution aligns with its programmer's codified intentions. This is not yet feasible for AIAs. When the outcome of a computation involves uncertainty, the users knowingly or unknowingly accept some risk. That is inherently the case with self-modifying software such as any ML system during its training phase. A certain risk is consequently characteristic for AIAs.

Gerck offers different definitions of trust [Ger02]. The simplest form is: "Trust is to rely upon actions at a distance." This describes the concept of trust as reliance. Another is "qualified reliance on information". Gerck's definitions connect information theory and – by extension probability theory – with risk and its role as a precondition for trust. These thoughts on trust are rather insightful but too abstract to be directly applicable for software implementation. In summary, uncertainty begets risk and by extension (dis-)trust. Transparency on the other hand reduces uncertainty, or rather epistemic uncertainty, for the user (see also fig. 2).

If trust is "qualified reliance", how is it to be qualified conveniently? In line with the research question 1 Stanton and Jensen give an answer to the question: "How does our evolutionarily ingrained and socially conditioned trust mechanism respond to machines?" They present a mathematical model for AI trustability $T(u, s, a)$ depending on a specific user u, an AI

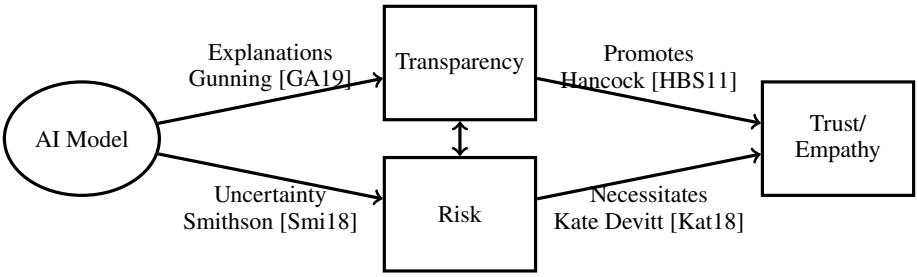


Fig. 2: Risk and transparency as influence factors for trust

system s, and a specific context a [SJ21]. They define the User Trust Potential $UTP(u)$. This subsumes the cultural, individual and otherwise subjective attributes of a user u. An AI system designer has little influence over UTP beyond precluding certain potential users from using his system. Of course, predominantly positive experiences with AI will have a positive impact on society-wide UTP in the long run.

There is also a less arbitrary and subjective factor of trustability that is significantly determined by the design of software architecture. This factor is Perceived System Trust Potential $PST(u, s, a)$ in a system s within a context a. This has an extensive overlap with learned and situational trust in Hoff's and Bashir's model. There is arguably also an additional objective component of qualified trust depending on s and a. The overall likelihood of trust is defined as:

$$T(u, s, a) = UTP(u) \cdot PST(u, s, a). \tag{1}$$

PST itself is defined as an arbitrary function g of user experience UX and perceived technical trustworthiness PTT: The latter is the sum of nine system characteristics ptt_c: accuracy, reliability & resiliency, objectivity & security, explainability, safety & accountability, and privacy. Thus

$$PST = g(UX, PTT), \quad PTT = \sum_{c=1}^{9} ptt_c. \tag{2}$$

Each characteristic is the product of its pertinence p_c and sufficiency s_c:

$$ptt_c = p_c \cdot s_c. \tag{3}$$

Each of the nine characteristics is assigned a relative Pertinence

$$p_c = \frac{q_c}{\sum_{k=1}^{9} q_k}. \tag{4}$$

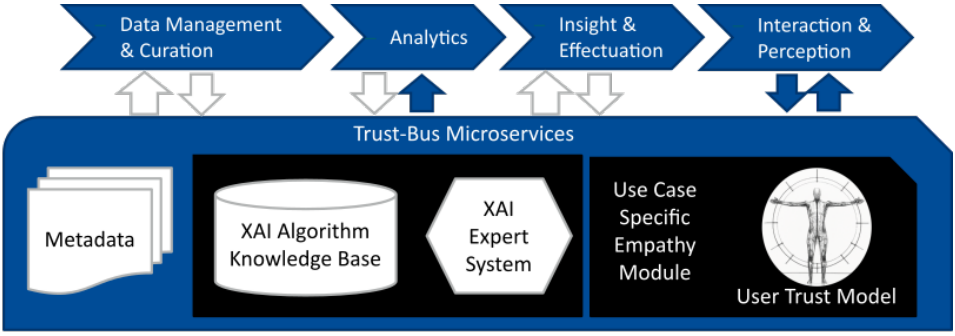
for an absolute pertinence q_c and a sufficiency s_c for a trustworthiness metric m_c and a perceived risk r_a in the given context:

$$s_c = m_c / r_a. \tag{5}$$

Obviously, Stanton's and Jensen's model of trust in AI is far from complete. UTP, g, q_c, m_c, and r_a remain uncertain. They address this by posing a number of research questions. Most are concerned with deducing parameter values from first principles. As this is not yet possible, a prototypical implementation must resort to preliminary measures like user and expert surveys or estimation in lieu of deduction. This should happen on a case by case basis. This paper proposes an XAI knowledge base (KB) that compiles these values for different ML algorithms and will be updated regularly.

The parameters must be deduced from first principles whenever possible. But what are good candidates for those principles? Subjective trust is not enough to build robust and beneficial AIAs. Attributes like reproducibility, transparency, and fairness must be monitored and quantified, if objective facts are to guide the decision to trust an AIA. Users must be made aware of risks that they have to accept [HBS11; Kat18; Smi18]. AIAs not only need to demonstrate the "what?" to their stakeholders but must also be able to explain the "how?" of their decisions, in order to achieve transparency. The answer to "how?" will come from systematically captured and analyzed meta-data presented in unison with the model's main results. Concerning the quality of training data, AIAs' purpose is to receive complex tasks from a human user. They represent data in action [ASR18] while FAIR is more concerned with data reusability for scientific knowledge mining and synthesis. Other metrics, e.g. for fairness or robustness, can be chosen according to the individual use case.

Fig. 3: Trust bus architecture for AI2VIS4BigData, adds a feedback loop that facilitates self-assessment through empathy with its users (Parts of this image were created with the assistance of DALL-E 2.)



3.2 Trust Bus

The research question 2 asks for the implementation of a software service for trustability analytics. This service cannot be a data sink, i.e. the meta-data must be processed somehow. This paper suggests a trust bus architecture to synthesize qualified trust in the trustor. The trust bus architecture was introduced in section 3.1. See figure 3 for its structure. The authors of this paper envision a rule based expert system (ES) for this task. This system will be akin to that in [Rei+22b]. This ES can query an analogue to the looking-glass self for AIAs.

Thus, the service can self-assess its trustworthiness in the eyes of its users. It computes the trust value $T(u, s, a)$ from eq. 1. The goal is to empower users by suggesting trustworthy, transparent, and explainable algorithms according to their individual needs. Therefore, the ES needs knowledge about the specific user u and the situation a that it addresses.

BDI-TM Module	Function
Actors and Entities Memory	stores information about past interactions between trustor and trustees for later evaluation by the trust production module
Trust Production	mechanisms to gauge trustworthiness of the AIA
Identity Management	identifies users or groups of users for appropriate handling by emotion and intent management
Intent Management	estimates users' intent, predicts possible consequences of trust and informs the trust production module accordingly
Emotion Management	models and learns affective states of the user from interaction
Risk Management	manages a task-specific risk registry and analyzes uncertainties in relation to possible risks
Complexity Management	estimates task complexities and the users' ability to make reasonable decisions under these circumstances

Tab. 1: Components of the trust bus BDI-TM architecture from [ALM16]

ES takes care of the *trust production* in table 1. It requests data from the knowledge base KB that stores information about explainability and the specific risks encodes as parameters of ML algorithms. The task specific empathy module holds information about *complexity management* that it passes on to ES and evaluates whether a given task has special requirements concerning fairness, transparency, etc. The metadata store (MS) holds the *actors and entities memory*, but also deals with *identity management*.

The trust bus is intended to be a reference architecture for trustability analytics and hence ought to be fairly universal. Therefore, it needs to interact with all steps of the AI2VIS4BigData analytics pipeline in figure 1. It shall recommend analytics, pre-processing and visualization methods in order to maximize the users' trust $T(u, s, a)$ from equation 1, which it provides for different users and situations by means of modularization. The system gauges its own value for $T(u, s, a)$ with a use case specific "empathy module" (EM). This module represents the mirror or looking-glass in the looking-glass self analogy. It has a user trust model that estimates $UTP(u)$. It also predicts the values for $PST(u, s, a)$ to the ES. The EM's input data come from ES, MS, and KM. Cues from its end users are taken in the last step of AI2VIS4BigData: "insight and effectuation". This is feedback is taken into account along with the other metadata when updating this trust model in a feedback loop. It is possible that the system asks the user to intervene when it passes a predefined uncertainty level.

3.3 Towards a Prototypical Implementation

Steps towards a prototypical implementation of the reference architecture are underway. An experimental system based on AI2VIS4BigData for the explanation of sentiment analysis of natural language texts was implemented. This system uses the visual presentation of feature importance calculated by the post-hoc explanation algorithms LIME and SHAP on variants of the transformer model BERT [Vas+17]. However, it turns out to be highly sensitive to certain data errors like word permutation and duplication, sometimes resulting in incomprehensible explanations. A later and structurally similar system is used for the recognition of emergent, formerly unknown named entities (NER) in medical texts. It forgoes explicit explanations but adds training on adversarial examples to its ML-pipeline for improved robustness. More use-cases involving foundation models for semi-autonomous visual information extraction of highly non-uniform documents and the generation of ontologies from biomedical research documents are being developed. A commonality between these systems is their emphasis on interaction with a human in the loop for result quality control. These systems and more provide input for the knowledge base KB on different approaches to XAI and test beds for the trust bus. As a next step and future research, a data schema for the knowledge base KB must be modeled. Concurrently an expert system will be implemented using miniKanren [Wil20; FBK05]. Together, these will build a first iteration of a prototypical trust bus as outlined above. The rules for this expert system ES will be evaluated in cooperation with users and domain experts.

4 Summary and Outlook

This paper has posed a number of open research questions concerning technical solutions of trust issues arising from the emergence of AIAs in big data analytics. It seeks to answer these questions by the nascent field of trustability analytics which is multidisciplinary and is located in digital humanities. The authors proposed a software service that takes cues from the theory of a looking-glass self. Such a service may fill the role of an architectural blueprint for future designs of trustworthy AIAs in big data analytics.

Many details of such a trustability service remain for future research because of the preliminary condition of the scientific debate. In particular, parameters cannot yet be derived from first principles. First and foremost, a knowledge base for XAI algorithms with parameters for Stanton's and Jensen's trust model has to be compiled before the trust bus architecture for AI2VIS4BigData can be implemented.

"Trustability analytics" stand for a reductionist concept of trust that can be reasoned about rigorously and which leads towards practical software engineering. This paper's outlook on digital trust will undoubtedly leave something to be desired for humanities' scholars, but as George Box famously quipped: "All models are wrong, but some are useful."

The authors of this paper hope that the proposed architecture will be a step towards practical application and a software prototype in the near future.

References

[ALM16] Hussein A. Abbass, George Leu, and Kathryn Merrick. "A Review of Theoretical and Practical Challenges of Trusted Autonomy in Big Data". In: *IEEE Access* 4 (2016), pp. 2808–2830. DOI: 10.1109/access.2016.2571058.

[ASR18] Hussein A. Abbass, Jason Scholz, and Darryn J. Reid. "Foundations of Trusted Autonomy: An Introduction". In: *Foundations of Trusted Autonomy*. Springer International Publishing, 2018, pp. 1–12. DOI: 10.1007/978-3-319-64816-3_1.

[Bar+20] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012.

[BB21] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis. Explore, Explain, and Examine Predictive Models*. CRC PRESS, 2021. ISBN: 9780367135591.

[Bom+21] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2021.

[Bon+19] Raymond Bond et al. "Digital empathy secures Frankenstein's monster". In: *Proceedings of the 5th Collaborative European Research Conference (CERC 2019)*. Vol. 2348. Apr. 2019, pp. 335–349.

[BH21] Marco Xaver Bornschlegl and Matthias L. Hemmje. "Supporting Data Science in Automotive and Robotics Applications with Advanced Visual Big Data Analytics". In: *Advances in Data Science: Methodologies and Applications*. Ed. by Gloria Phillips-Wren, Anna Esposito, and Lakhmi C. Jain. Cham: Springer International Publishing, 2021, pp. 209–249. ISBN: 978-3-030-51870-7. DOI: 10.1007/978-3-030-51870-7_11.

[Bro+20] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. DOI: 10.48550/ARXIV.2005.14165.

[DR20] Arun Das and Paul Rad. "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): a Survey". In: *CoRR* (2020).

[FBK05] Daniel P. Friedman, William E. Byrd, and Oleg Kiselyov. *The Reasoned Schemer*. The MIT Press, 2005. ISBN: 0262562146.

[Ger02] Ed Gerck. "Trust as Qualified Reliance on Information, Part I". en. In: (2002). DOI: 10.13140/RG.2.2.22646.04165.

[Gos10] Helena Goscilo. "The Mirror in Art: Vanitas, Veritas, and Vision". In: *Studies in 20th &; 21st Century Literature* 34.2 (June 2010). DOI: 10.4148/2334-4415.1733.

[GA19] David Gunning and David Aha. "DARPA's Explainable Artificial Intelligence (XAI) Program". In: *AI Magazine* 40.2 (June 2019), pp. 44–58. DOI: 10.1609/aimag.v40i2.2850.

[Had21] Dylan Hadfield-Menell. "The Principal-Agent Alignment Problem in Artificial Intelligence". PhD thesis. EECS Department, University of California, Berkeley, Aug. 2021.

[HBS11] P. A. Hancock, D. R. Billings, and K. E. Schaefer. "Can You Trust Your Robot?" In: *Ergonomics in Design: The Quarterly of Human Factors Applications* 19.3 (July 2011), pp. 24–29. ISSN: 1064-8046. DOI: 10.1177/1064804611415045.

[Hig20] High-Level Expert Group on AI. *White Paper on Artificial Intelligence. a European approach to excellence and trust*. eng. Report. Brussels: EU Kommission, Feb. 2020.

[HB14] Kevin Anthony Hoff and Masooda Bashir. "Trust in Automation". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57.3 (Sept. 2014), pp. 407–434. DOI: 10.1177/0018720814547570.

[ISO18] ISO Central Secretary. *Risk management – Guidelines*. en. Standard ISO 31000:2018. Geneva, CH: International Organization for Standardization, 2018.

[Jac+20] Annika Jacobsen et al. "FAIR Principles: Interpretations and Implementation Considerations". In: *Data Intelligence* 2.1-2 (Jan. 2020), pp. 10–29. ISSN: 2641-435X. DOI: 10.1162/dint_r_00024.

[Jac+21] Alon Jacovi et al. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI". In: FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 624–635. ISBN: 9781450383097. DOI: 10.1145/3442188.3445923.

[Kat18] S. Kate Devitt. "Trustworthiness of Autonomous Systems". In: *Studies in Systems, Decision and Control* (2018), pp. 161–184. ISSN: 2198-4190. DOI: 10.1007/978-3-319-64816-3_9.

[KCW05] Beverly Kracher, Cynthia Corritore, and Susan Wiedenbeck. "A foundation for understanding online trust in electronic commerce". In: *Journal of Information, Communication and Ethics in Society* 3 (Aug. 2005), pp. 131–141. DOI: 10.1108/14779960580000267.

[Kul+13] T. Kulesza et al. "Too much, too little, or just right? Ways explanations impact end users' mental models". In: *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC* (2013), pp. 3–10. DOI: 10.1109/VLHCC.2013.6645235.

[Lif15] Future of Life Intitute. *An Open Letter on AI*. https://futureoflife.org/ai-open-letter/. Accessed: 2020-8-25. 2015.

[Lip18] Zachary C. Lipton. "The mythos of model interpretability". In: *Communications of the ACM* 61.10 (Sept. 2018), pp. 36–43. DOI: 10.1145/3233231.

[Luh14] Niklas Luhmann. *Vertrauen – Ein Mechanismus der Reduktion sozialer Komplexität*. 5th ed. utb GmbH, Feb. 2014. DOI: 10.36198/9783838540047.

[LL17] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774.

[McC07] John McCarthy. *WHAT IS ARTIFICIAL INTELLIGENCE?* 2007. URL: http://www-formal.stanford.edu/jmc/whatisai.html (visited on 04/01/2021).

[McI07] L. McIntyre. *The Practical Skeptic: Core Concepts in Sociology*. McGraw-Hill Companies,Incorporated, 2007. ISBN: 978-0-07-340415-8.

[Meh+21] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Comput. Surv.* 54.6 (July 2021). ISSN: 0360-0300. DOI: 10.1145/3457607.

[Mil19] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (Feb. 2019), pp. 1–38. DOI: 10.1016/j.artint.2018.07.007.

[Mon+20] Barend Mons et al. "The FAIR Principles: First Generation Implementation Choices and Challenges". In: *Data Intelligence* 2.1-2 (Jan. 2020), pp. 1–9. ISSN: 2641-435X. DOI: 10.1162/dint_e_00023.

[Pap+22] Andrea Papenmeier et al. "It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI". In: *ACM Trans. Comput.-Hum. Interact.* 29.4 (Mar. 2022). ISSN: 1073-0516. DOI: 10.1145/3495013.

[Ram+22] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. DOI: 10.48550/ARXIV.2204.06125.

[Rei+21] Thoralf Reis et al. "Towards Modeling AI-based User Empowerment for Visual Big Data Analysis". In: *BIRDS+WEPIR @ CHIIR 2021*. Ed. by Ingo Frommholz et al. Vol. 2863. CEUR Workshop Proceedings. CEUR-WS.org, Mar. 2021, pp. 67–75.

[Rei+22a] Thoralf Reis et al. "A Service-based Information System for AI-supported Health Informatics". In: *2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI)*. 2022, pp. 99–104. DOI: 10.1109/BDAI56143.2022.9862611.

[Rei+22b] Thoralf Reis et al. "Supporting Meteorologists in Data Analysis through Knowledge-Based Recommendations". In: *Big Data and Cognitive Computing* 6.4 (Sept. 2022), p. 103. DOI: 10.3390/bdcc6040103.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[Roc60] Norman Rockwell. "Triple self-portrait". In: *The Saturday Evening Post* (Feb. 1960).

[Rud19] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

[SY19] P.J. Scott and R.V. Yampolskiy. "Classification Schemas for Artificial Intelligence Failures". In: *Delphi - Interdisciplinary Review of Emerging Technologies* 2.4 (2019), pp. 186–199. DOI: 10.21552/delphi/2019/4/8.

[Sha04] Leigh S. Shaffer. "From mirror self-recognition to the looking-glass self: Exploring the Justification Hypothesis". In: *Journal of Clinical Psychology* 61.1 (2004), pp. 47–65. DOI: 10.1002/jclp.20090.

[Smi18] Michael Smithson. "Trusted Autonomy Under Uncertainty". In: *Studies in Systems, Decision and Control* (2018), pp. 185–201. ISSN: 2198-4190. DOI: 10.1007/978-3-319-64816-3_10.

[SJ21] Brian Stanton and Theodore Jensen. *Trust and Artificial Intelligence*. en. Mar. 2021. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087 (visited on 01/09/2023).

[Vas+17] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[Wer18] Kevin Werbach. *The blockchain and the new architecture of trust*. Information Policy. London, England: MIT Press, Nov. 2018.

[Wil20] Brandon T. Willard. *miniKanren as a Tool for Symbolic Computation in Python*. 2020. DOI: 10.48550/ARXIV.2005.11644.