A Survey on methods of Trustworthiness towards Artificial Intelligence

Hiralal B. Solunke
Asst. Professor, (CSE & IT Department)
G.H. Raisoni Institute of Business Management, Jalgaon (Maharashtra, India)
Gate No.57, Shirsoli Road, Mohadi Jalgaon 425002

Sonal P. Patil
Asst. Professor, HOD (CSE & IT Department)
G.H. Raisoni Institute of Business Management, Jalgaon (Maharashtra, India)
Gate No.57, Shirsoli Road, Mohadi Jalgaon 425002

Shital S. Jadhav
Asst. Professor, (CSE & IT Department)
G.H. Raisoni Institute of Business Management, Jalgaon (Maharashtra, India)
Gate No.57, Shirsoli Road, Mohadi Jalgaon 425002

ABSTRACT

This is the survey paper based on the role of trustworthiness of data analytics from the data quality and privacy concern perspectives in artificial intelligence. Science fiction movies like ‘The Terminator’ and ‘I, Robot’ have exhibited what might happen in case artificial intelligence goes rogue. Such dystopian fantasies about AI are widely discussed by experts and researchers in the field of AI as well. Many of these experts believe that super-intelligent AI systems will pose a significant threat to humanity in the near future. And, considering the untold potential of AI, this may soon become a reality. Artificial Intelligence System Developers need to understand society concerns over the development of Artificial Intelligence System.

There have been many reported instances where developers neglected these warnings and created AI systems that went rogue and which is harmful to society. This survey paper describes the risks and challenges AI; also AI can be used to enhance the trustworthiness of a system. It is noteworthy that the same technologies that can lead to trust concerns may also be applied to improve the trust in systems and to mitigate risks. Safety, security and reliability can be improved through the appropriate use of AI technologies since they can enable faster response and adaptability of a system to unforeseen situations.

Keywords

Artificial Intelligence (AI), trustworthiness, super-Intelligent, adaptability, predictability, explainability, Trust, E-Trust.

1. INTRODUCTION

The advancement of artificial intelligence (AI) presents humanity with opportunities as well as challenges. AI could contribute to increases in efficiency that radically impact productivity and that may someday mitigate or even eliminate scarcity—offering abundance and ending wealth disparity. However, this future is not inevitable. AI is a powerful technology, and its power continues to grow. Like any powerful tool, it can be used to positive or negative ends. In developing and deploying AI systems, we must carefully consider the implications of this advancement and emphasize collaboration between humans the technology [1]. Real engines of the artificial intelligence (AI) revolution, machine learning (ML) models, and algorithms are embedded nowadays in many services and products around us. As a society, we argue it is now necessary to transition into a phronetic paradigm focused on the ethical dilemmas stemming from the conception and application of AIs to define actionable recommendations as well as normative solutions. However, both academic research and society-driven initiatives are still quite far from clearly defining a solid program of study and intervention.

In this contribution, we will focus on selected ethical investigations around AI by proposing an incremental model of trust that can be applied to both human-human and human-AI interactions. Starting with a quick overview of the existing accounts of trust, with special attention to Taddeo’s concept of “e-trust,” we will discuss all the components of the proposed model and the reasons to trust in human- AI interactions in an example of relevance for business organizations [2].

1.1.1 AI, ML, and Algorithms

In this section, we briefly introduce a minimum amount of notation on AI and ML in order to guide the reader through the paper and then move on to a quick analysis of society-relevant AI, ML.

1.1.1.2 Definitions

AI is the multidisciplinary endeavor to build machines that can learn, take decisions, and act intelligently in the environment (Russell and Norvig 2009). Decisions can be outputs of a learning process, as well as inputs to generate new ones. Machines are technological artifacts comprised of a combination of software and hardware components. The rising interest on AI in domains like healthcare, retail, marketing, and financial services is due to the ability of AIs to endow products and services with “cognitive functions” through their capability to learn and suggest decisions from digital data. The effectiveness of AIs in delivering performance (e.g., supporting financial growth and cost savings or beating human experts in computer vision tasks) has further boosted their penetration in modern societies. An important component of AI’s success is represented by ML, which is the discipline that combines statistical modelling and science of algorithms to create computer systems able to automatically generate predictions and support decision-making by learning inductively from input data (Mitchell 1997; Vapnik 2000). Following Mitchell: “[e]ach ML problem can be precisely defined as the problem of improving some measure of performance P when executing some task T, through some type of training experience E" (Mitchell 1997).

Therefore, given a ML problem at hand (where the task T is, for example, the classification of an email into "spam" or "not spam," the computation of personalized premiums for a given insurance product or the assessment of the risk level of a bank customer), the statistical model defines the theoretical structure of its solution. It comprises mathematical constructs describing the task T, the performance measure P to assess results, and the structure of the set of input data encoding experience E. On the other hand, the algorithm3 is the procedure implemented into computer-understandable language to generate the solution itself, for example, by computing the parameters of the chosen ML model using available input data.

This is what technically is referred to as "training the ML model": it is the core process to deploy those AIs which learn inductively on data. In most applications, the result of training a ML model is an object in a given programming language that can be used to generate predictions to support decision-making, once it is fed with new data and embedded in an ad hoc IT architecture [2].

This dynamic infrastructure is a key component of the design of AI powered products and services; depending on their technical complexity (e.g., the number of customers they are supposed to reach, the number of transactions per second) and the structure of the organization promoting them (start-up vs. well established organization), such infrastructure can be fully cloud-based, hybrid, or developed and managed in-house. AIs typically use multiple ML models, algorithms, and automated data processing pipelines to generate predictions5: on the other hand, communication with end users is driven by interfaces that can take the form of apps, web-based applications, and, more recently, even augmented reality devices [2].

1.1.1.3 Trust, e-trust, and an Incremental Model of Trust

People use the expression trust in a manifold of different ways to describe a variety of human affairs. People talk easily about trusting their friends, peers, or even strangers; people trust their own intuitions or themselves; they trust science and possibly the scientific community. Sometimes they (even) trust politicians or institutions; in certain cases, they can also express a trusting attitude towards non-human agents. Trust is thus a construct enriching a wide variety of relationships people establish, nurture, and interrupt in everyday life. When talking about trust, people refer to specific goals and contexts: for example, I trust my friend's competence as a scholar versed in microbiology, but I would never trust him to post my letter.

Similarly, I trust my accountant to accurately file my tax return respecting the fixed deadline, but not as a political advisor. Therefore, people do not trust each other in every possible way, but assess the competence of others in a specific context, with respect to a predefined goal they care about. Sometimes, they express trust in someone or something, sometimes they simply trust someone or something, and sometimes they even trust that something is the case. In most of our experiences, trust is interpersonal and "face-to-face" or it is mediated by technology, as in the case of digital environments and the phenomenon of e-trust (Taddeo 2009). Therefore, trust is a relevant construct to be considered in everyday life, at different levels; however, "there is not yet a shared or prevailing, and clear and convincing notion of trust" (Castelfranchi and Falcone 2010) [2].

E-Trust

E-trust is an interesting approach to trust in digital environments and in the presence of artificial agents (of which AIs are part of). E-trust is "trust applied to digital context and/or involving artificial agents" (Taddeo and Floridi 2011) and occurs "in environments where direct and physical contacts do not take place, where moral and social pressures can be differently perceived, and where interactions are mediated by digital devices" (Taddeo 2009). Therefore, e-trust becomes relevant in the presence of interactions with electronic commerce platforms, group chats and online communities, technology-mediated self-services, multi-agent systems contexts, and, in general, whenever humans or artificial agents, or both interact in a digitally mediated environment. As starting point, an account of e-trust is necessarily required to define trust, since the latter is the reference point, and to tackle the question whether e-trust is an occurrence of trust or an independent phenomenon. As the existence of a shared and institutional background and the certainty about the trustee's identity are usually identified as necessary conditions to develop any trust theory (Taddeo and Floridi 2011), detractors of e-trust discuss the impossibility of having trust in digital environments due to the absence of such conditions (Pettit 1995; Nissenbaum 2001) [2].

2. METHODS

2.1 An Incremental Model of Trust: Definition

A model of trust, which takes care of both cognitive and non-cognitive accounts incrementally and in a finite sequence of steps. The model explains the many ways in which people—both ordinary people and scholars—talk about trust, as well as the relation between trust and trustworthiness. It can be applied to relationships between humans and artificial agents, with focus on those artificial agents endowed with cognitive capabilities stemming from ML algorithms. Model of trust T consists of the triple [2].

T= (Simple Trust, reflective trust, paradigm trust)

Whose elements are constructed on the 5-tuple (X, Y, A, G, C), where X and Y denote interacting agents and A the action to be performed by the agent Y to achieve a goal G of relevance for X in a given context C. For simplicity of exposition, this latter is kept fixed in the forthcoming discussions. The remainder of this section is devoted to the discussion of the triple T.

2.2 Incremental Model of Trust in Human-AI Interactions: a Simple Example

Let us now move to a simple example of human-AI interactions. To do so, let us imagine a proactive company pushing for AI-powered solutions to generate business performance through innovation and new technologies. The top management of the company recently decided to approve the design and test launch of a new product for a selected (and high priority) portfolios of customers. At the core of the product lies a two-staged cognitive engine, i.e., an AI comprising of multiple ML algorithms and automated data processes. The business decisions are taken based on both machine generated predictions and human expertise (for example, considering a nested IF/ELSE logic structure). Health insurance solutions collecting wearable's data, telemetric products, or personalized marketing campaigns based on credit card usage are all examples of the above. The conception, design, development, testing, and deployment of the solution is a highly complex endeavour, involving a mix of domain expertise and technical capabilities in product development, data science, IT engineering, and project management. For this means, the company decides to hire a team of seasoned consultants to support in-house resources (in primis the project sponsor and his manager) to finalize and launch the AI-powered solution [2].

2.3 Analyzing the trust issues with AI

It is highly likely that the negative implications of AI will not be as bad as sci-fi movies and books depict. However, the possible negative consequences of AI can still pose significant threats to the human race. Experts have discussed consequences like losing jobs to AI, where humans may soon be replaced by AI in various roles. The competence of AI can already be witnessed in different industries such as healthcare, retail, aviation, manufacturing, and many more as AI-enabled applications have transformed and streamlined various business procedures. Hence, well-established businesses are leveraging AI for automating core tasks.

2.4 Developing Trustworthy AI

Tech companies and developers can consider the following factors for building trustworthy AI:

2.4.1Explainability

AI has a serious black box problem, where AI systems make crucial decisions based on machine learning algorithms instead of big data. Hence, end-users and developers may not understand why an AI system made a specific decision. Due to the lack of explanation, users may doubt the accuracy of results generated by AI systems. Hence, developers need to build explainable AI systems. For this purpose, companies that utilize AI have to open the black box and understand how AI systems make crucial decisions and generate results.

2.4.2 Integrity

Machine learning integrity is a necessary condition for developing trustworthy AI systems. Machine learning integrity can help ensure that AI systems generate the output according to a developer's predefined operational and technical parameters. With machine learning integrity, developers can make sure that AI systems work as they are intended to. Also, developers can set up certain limitations for AI systems that can be used to regulate the usage of AI. In this manner, developers can design trustworthy AI systems that produce accurate results by following predefined conditions.

2.4.3 Conscious development

While developing AI systems, developers need to ensure that the decisions made by AI will benefit humans. For this purpose, AI systems need to be aligned with human principles and values. Hence, the objectives designed for AI systems must align with human values and focus on making human life better. Using this mindset, developers can consciously design applications that will benefit the human race.

2.4.4 Reproducibility

Reproducibility ensures that every outcome generated by an AI system can be reproduced. If an outcome is not reproducible, there is no clear way to understand why a result was generated. Also, the outcomes generated by an AI system can be affected by multiple factors such as algorithms, artifacts, system parameters, different versions of code, and various datasets. Hence, ensuring reproducibility can be immensely challenging.

Multiagent Systems, Trust, and AI [11]

It must be noted here that I am not excluding a trust directed towards individual human beings behind the development, deployment, and integration of AI, or the possibility of trusting the organisations developing, deploying and integrating AI.

There are positions in the field that try to include AI as something that can be trusted in a very weak sense, often tying this trust to a trust in'multi-agent systems', where AI is one of these agents.

Buechner and Tavani (2011), using Walker's (2006) difuse/default model of trust, claim that one can trust multi-agent systems that include humans, groups of humans, and also artifcial agents—'such as intelligent software agents and physical robots' Walker stated that we trust particular zones and groups of people. She discusses larger groups or communities, such as cities, whereby people can follow practices appropriate for that place. There is a normative expectation on people to act in a certain way. This behaviour becomes habitual and 'one simply engages in that behavior, with little or no conscious refection' Buechner and Tavani claim that this difuse/default model of trust may be applied to AI, because it allows for distributing responsibility over a diverse network of human agents and artifcial agents. As many acts of trust are grounded in non-interpersonal relationships, or mixed-relationships (i.e. interpersonal and non-interpersonal), then we should establish a type of trust that takes this into account. These mixed trust relationships, or multi-agent trust relationships, may take the form of trusting groups of individuals, organisations, and perhaps, AI technologies within that network of trust (Buechner and Tavani 2011).

Within the literature on the philosophy of trust, there is often disagreement over trust in organisations, institutions, and groups. Some argue that one can indeed place a trust in organisations as entities themselves, as they have a normative commitment towards us or we believe they are acting out of goodwill towards us. Others propose that these organisations are only a very complex form of interpersonal trust. When we refer to trusting an organisation, we are implicitly trusting the entire composition of individuals in that group to commit to the normative standards of their organisation. I will evaluate Buechner and Tavani's position that we can trust AI in multiagent systems, with these two positions in mind.

Firstly, Buechner and Tavani (2011) provide an example of an auction on eBay. However, this is still a zone of default trust in the organisation itself, and/or the other moral agents in these exchanges, regardless of their proximity or relationship to us.It is a trust in eBay as a company to ensure that we are not scammed, and there are appropriate responses to those who do not respect their normative commitment to users. It is a trust in the individuals working in eBay who are designing

3. CONCLUSION

AI is a phenomenon affecting individuals and their lives, organizations, and societies as a whole. The ability to perform complex tasks and support decision-making thanks to ensembles of ML models and algorithms prima facie supports the adoption of AI in multiple domains. Therefore, it is necessary to discuss the nature and dynamics of trust in the presence of human-AI interactions, with focus on the properties of trustworthy AI. In this paper we discussed different models & factors which are considered for trustworthiness of Artificial Intelligence (AI). Those using the concept of trustworthy AI to indicate a moral goal or objective should carefully define what they mean by trustworthiness. Sequence of past, successful interactions, as they have been working together on similar engagements for quite some time.

4. REFERENCES

[1] Andrew B. Ware, "Algorithms and Automation: Fostering Trustworthiness in Artificial Intelligence" spring 2018.

[2] Andrea Ferrario, Michele Loi & Eleonora Viganò," In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions".

[3] Buechner, J., & Tavani, H. T. (2011). Trust and multi-agent systems: applying the "diffuse, default model" of trust to experiments involving artificial agents. Journal Ethics and Information Technology, 13(1), 39–51.

[4] Anderson, J., & Rainie L. (2018). Artifcial intelligence and the future of humans, Pew Research Centre, available here: https://www.pewinternet.org/2018/12/10/artificial-intelligence-and-the-future-ofhumans/. Accessed 25 Sept 2019.

[5] Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., et al. (2018). Trusting intelligent machines: Deepening trust within socio-technical systems. IEEE Technology and Society Magazine, 37(4), 76–83.

[6] Asaro, P. M. (2019). AI ethics in predictive policing: From models of threat to an ethics of care. IEEE Technology and Society Magazine, 38(2), 40–53. https://doi.org/10.1109/MTS.2019.2915154. Baier, A. (1986). Trust and antitrust. Ethics, 96(2), 231–260.

[7] Blumberg Capital. (2019). Artifcial Intelligence in 2019: Getting past the adoption tipping point. Blumberg Capital. 2019. https://www.blumbergcapital.com/ai-in-2019/. Accessed 21 Nov 2019. Bryson, J. (2018). AI & Global Governance: No one should trust AI. United Nations.

[8] Bryson, J. J., & Kime, P. P. (2011). Just an artifact: Why machines are perceived as moral agents. In Twenty-second international joint conference on artifcial intelligence.

[9] Buechner, J., & Tavani, H. T. (2011). Trust and multi-agent systems: Applying the "difuse, default model" of trust to experiments involving artifcial agents. Ethics and Information Technology, 13(1), 39–51.

[10] Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral signifcance of the appearance, perception, and performance of artifcial agents. AI & SOCIETY, 24(2), 181–189. Coeckelbergh, M. (2012). Can we trust robots? Ethics and Information Technology, 14(1), 53–60.

[11] Mark Ryan (2020) ,In AI We Trust: Ethics, Artifcial Intelligence, and Reliability