Build confidence and acceptance of AI-based decision support systems - Explainable and liable AI

Claire NICODEME, PhD
Innovation & Recherche
SNCF
Saint Denis, FRANCE
claire.nicodeme@sncf.fr

***Abstract*— Artificial Intelligence has known an incredible development since 2012. It was due to the impressive improvement of sensors, data quality and quantity, storage and computing capacity, etc. The promises AI offered led many scientific domains to implement AI-based decision support tool. However, despite numerous amazing results, very serious failures have raised Human mistrust, fear and scorn against AI. In Industries, staff members cannot afford to use tools that might fail them. This is especially true for Transportation operators where security and safety are at risk. Then, the question that arises is how to build Human confidence and acceptance of AI-based decision support system. In this paper, we combine different points of view to propose a structured overview of Transparency, Explicability and Interpretability, with new definitions arising as a consequence. Then we discuss the need for understandable information from the AI system, to legitimate or refute the tool's proposal. To conclude we offer ethical reflexions and ideas to develop confidence in AI.**

Keywords—explainable AI, liable AI, decision support system, confidence, technology

I. INTRODUCTION

Artificial Intelligence or AI is a wide-ranging branch of computer science. It aims at building systems capable of recognizing a situation or event, and taking decisions of the form "IF this situation exists THEN recommend or take an action" [3]. After decades of lack of interest in a technology that "just didn't work" (AI winter*), Artificial Intelligence awoke again in 2012. It was due to the impressive improvement of sensors, data quality and quantity, storage and computing capacity, etc. It is nowadays a buzz word that media have spread and that triggered academic and industrial interest. The results achieved with those new techniques and technologies have since led numerous professions to implement AI-based tools and systems. It is used today in many fields such as healthcare [5], criminal justice, human resources, finance [1], education, transportation [15], and more.

However, AI-based systems have proved multiple times to have just as much flaws as human beings. Here are three examples from 2018. In China, an AI-based system wrongly identified an advertisement on a passing bus as a jaywalker. The given explanation was that it is difficult for long range live detection to differentiate a real person's face and an image [8]. In the USA, a self-driving Uber car struck and killed a pedestrian.

*An AI winter is a time when support for and interest in Artificial Intelligence research and commercial ventures dries up. It happened at the end of the 1980's and terminated AI research for decades.

The software detected the pedestrian but decided not to take any action. It has been shown that, due to an incompatibility, the Uber autonomous mode disabled the constructor factory-installed automatic emergency braking system [12].

At Amazon, AI-enabled recruiting software helped review applicants' resumes and make recommendations. It was found that the tool was gender-biased and downgraded women's profiles.

Failures in AI have since continued to disappoint and worry people, either with deadly incidents, ethically dramatic events (for example CLEARVIEW AI) or just malfunctioning technology.

The technology that first amazed the world is now facing its limits. AI algorithms have shown they are neither flawless nor 100% reliable. Moreover, the automation of formerly human-specific tasks raises a number of questions. It leads to the rise of human mistrust, fear and scorn against AI. This situation is slowing down the development of new applications. It is especially true for applications in the domain of public transportation, where bad decision-making could cost the life of numerous people.

In this paper, we present the requirements for actors working at the French Railway Company (SNCF) to accept and rely on AI-based innovative systems. Actors will also be referred to as Agents. The first Section gives a structured overview of Transparency, Explicability and Interpretability, and discusses whether or not "trust" can be applied to AI. Section two offers an overview of decision-making tools in the industry and recalls the four major subjects to address to build confidence and acceptability. Finally, Section three concludes the paper with recommendations to give confidence in AI decision-making support tools.

II. DEFINITIONS

A. Trust in AI

According to [13], speaking of trust in AI and technology is an error. It makes the assumption that AI algorithms belong to a group of objects that can be trusted. However, trust implies placing something of value in the responsibility of another being in circumstances of vulnerability. It involves human thoughts, motives and action lying beyond technical characteristics. To date AI do not have motives or character. Moreover, if we were to conflate trust with reliability and accuracy, as the performance of AI improves, this would decrease trust in experts whose technical accuracy might end

978-1-7281-7392-4/20/$31.00 ©2020 IEEE

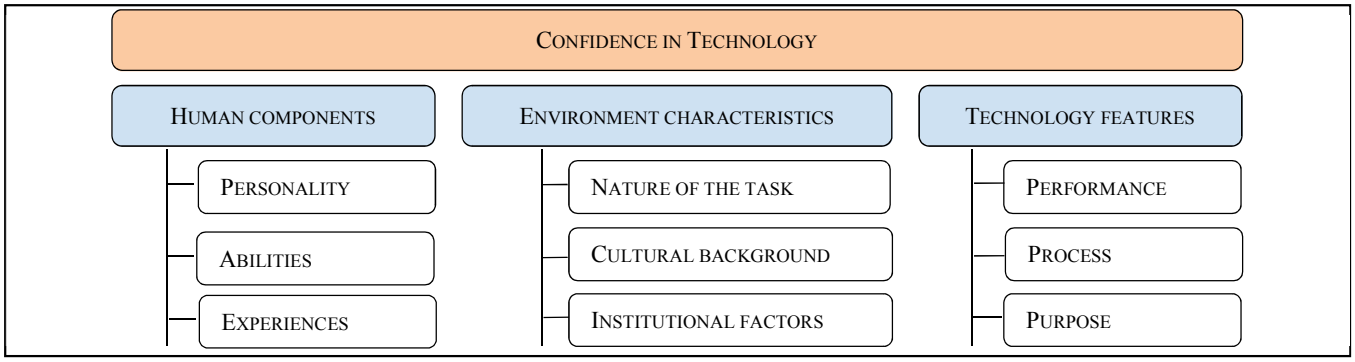

Fig. 1. Factors of confidence in technology [11]

up being inferior to machines. As an entity without feeling or personality the author proposes to express "trust" toward a system or machine as "confidence".

In human relationships, trust is essential to create long lasting bonds. As AI gets more efficient, it strongly impacts and participates in humans' daily lives. Human-AI partnerships need confidence just as Human-Human teams need trust.

Confidence in technology is determined by human components, environment characteristics and technology features as shown in Figure 1.

However, the ways these two kinds of relationships work are reversed: in contrast to Human-Human interactions, Humans start with the assumption that the AI-based system is near perfect. Therefore, at the beginning, faith is the major constituent of the "confidence". Then, the more human and AI interact, the quicker faith is replaced by reliance and predictability. Several factors are at play during this process, as discussed in [11].

B. Transparency, Explicability, Interpretability

Given the widely spread use of AI technology, it is crucial to understand the processes/methods behind it. To build comprehension, different concepts are necessary, according to the purpose of the algorithm, the input data, the expected output data, etc. These concepts are transparency, explicability and interpretability. Even though they are very close in meaning, they have different definitions.

Transparency: *The characteristic of being easy to see through*. Applied to AI algorithms, it is the possibility to visualise the process chain. A transparent AI system should allow answering questions such as: what data were used to train the algorithms, what data were given as input, what is the algorithms architecture, what processing was used on data, etc. Transparency is a necessity as it can help mitigate issues of fairness, discrimination, and confidence. It also helps to calibrate the suitable amount of confidence by providing users with accurate mental models of AI underpinning [10,18]. Note that transparency does not clarify the way the algorithms work.

The interpretability and explicability of algorithms are key issues of Artificial Intelligence. They are especially needed to enlighten the outcome of an AI and guide the decision makers' choice toward the adaptated operational or strategic orientations.
Data scientists clearly distinguish explicability and interpretability, defined as follows.

Explicability: *The characteristic of being able to explain*. For an AI-based system, it means providing the users with knowledge about how results were obtained. Explanation is about reasoning and making the reasoning explicit. Explicability highlights the variables or data that led to a given conclusion (for example LIME, DEEPLIFT, ELI5, INTERPRETML, SHAP). Explicability is mandatory to understand how the system works, its limitations and the possibilities it offers. Note that explicability will vary immensely industry to industry. It also depends on the audience. Users will neither need the same amount nor the same kind of explanations according to their profession.

Interpretability: *The characteristic of being able to find a particular meaning in something*. For an AI tool, interpretability is the degree to which a human can consistently predict the model's result, without trying to know the reasons behind the scenes [2, 4, 6]. An algorithm is said to be interpretable if users understand how it works and how it learns. This is typically the case of a linear regression whose result can be expressed in a simple analytical form.

The mathematical difference between explicability and interpretability, expressed by [9], is as follows: explainable AI is using a black box and explaining what is inside, interpretable AI is intended to use a model that is not a black box

As a synthesis, TABLE I shows the questions that transparency, explicability and interpretability aim at answering.

TABLE I. TRANSPARENCY, EXPLICABILITY, INTERPRETABILITY

Confidence Drivers	**Question concepts have to answer**
Transparency	What is inside my AI-based system?
Explicability	What variables did the system use to give a result?
Interpretability	How does the system work?

C. Liability

Human beings have a tendency or a need to understand how their tools work before using them. This comprehension goes through transparency, explicability and interpretability.

Many people hope that AI is going to augment rather than replace human decision-making. To achieve this result, explicability is a key factor. It becomes a prerequisite for building confidence and favouring the adoption of AI systems.

It is especially true in high stakes domains, requiring reliability and safety such as automated transportation or critical industrial applications with significant economic implications (e.g. predictive maintenance). As a consequence, AI actors have focused their attention on explainable and interpretable AI to help them increase confidence and understand models at scales [16].

However, the understanding of a tool by a user is not enough. The question of liability is a very important part in the acceptance of AI-based tools. The law has long regulated the causal relation between people and things. However, AI is different: the "thing" is increasingly complex (a difference in degree) and its agency is continuously changing (a difference in kind) [16].

Lawyer firms in France such as [19], presents the subject as follow. The arrival of "intelligent" robots in our world is equivalent to the birth of a new species. To date, AI-based systems are not comparable to humans. They debate on the need to introduce a specific legal framework for AI.

AI is not yet personified enough to have its own crime and punishment regime. Yet, the separation between makers and machines' liability is widening. Causation and fault are more and more opaque too. This is linked to the increase of human-machine interaction in a widening spectrum of AI usage.

III. Decision-Making Support Tools in the Industry

Decision making is an inherent part of human activity that can have significant impacts. Researchers have attempted to improve the quality of humans' decisions by developing computer technologies to augment and extend human capabilities. In recent years, AI tools have advanced sufficiently to be integrated into decision-making support systems for industrial applications.

They can be used to extend human capabilities by, for example, surveying and selecting relevant information from extremely large and distributed data sources, applying analytical tools to unstructured data, creating generalized solutions from rulesets and probabilities, and finding associations in information from multiple sources that may influence a decision. Algorithms such as artificial neural networks, fuzzy logic, evolutionary computing and probabilistic reasoning improve decision support systems drastically, allowing them to evaluate and select better alternatives. Decision support tools are thus impacting decision-making in substantial ways. It offers time savings and efficiency gains. It also allows less experienced Agents to handle difficult or complex situations.

Those support tools are aimed at improving the decision-making process. Nonetheless the acceptance of future users is questioned as well as their will to rely on those new, often unproven, tools. In the railway industry, people have their expertise. They are very keen to preserve their current behaviour to ensure safety and efficiency. Current processes have worked very satisfactorily for a long time, specific profession gestures and actions are engraved in Agents' habits. Providing a new tool in this context could be a little complicated. Four key elements are to be studied in order to guide users toward acceptance.

A. Ethics and responsablities

In Europe, common regulation (GDPR) sets out the rights and obligations around the use of automated decision making. Meaningful information must be given [20], explaining the logic involved as well as the significance and envisaged consequences of such a processing [7].

However, specificities are left for countries to choose. For example, the French Ministry of Defence and the European Commission [14] considers that human is always responsible, whatever his involvement in the decision-making process. Ethics managers in major companies wonder about the fairness of this choice. "Did the user have all the elements to be held responsible for an algorithm's decision?" It seems a bit insincere to put liability on the humans when AI usually takes over. Indeed, it has been shown that humans tend to give in to the machine [17]. What happens when humans' experiences do not fit with the systems recommendations? Table II summarises the possible cases that can be faced when using a decision-making support system

TABLE II. Human – AI-based tool interaction

The Agent (A) and the System (S) agree on the action to take	
1 *2*	• *A & S are right:* **The tool learned the profession's expertise properly** • *A & S are wrong:* **The incident was either too occasional or too complicated**
The Agent and the System do not agree on the action to take	
	A does not take into account the recommendation from S
3 *4*	• *A is right:* **The Human is better than the Machine, AI lacks business expertise** • *A is wrong:* **The Human needs to justify his action, liability is an issue here**
	A takes into account the recommendation from S
5 *6*	• *A is right:* **The system performed better than the Human, it could be used to help train new Agents** • *A is wrong:* **The Human needs to justify his action, liability is an issue here**
The System is unable to give a recommendation	
7 *8*	*A knows how to handle the event:* **Same as case 3_a** *A doesn't know how to react to the event:* **Same as case 2**
The Agent is unable to give a recommendation	
9 *10*	*S knows how to handle the event:* **Same as case 3_b, liability issue** *S doesn't know how to react to the event:* **Same as case 2**
Neither the Agent nor the System is able to give a recommendation	
11	***Same case as 2***

The five inventoried categories include eleven situations. In this lot, four situations seem to imply a lack of training or experience from both the Agent and the AI system (2, 8, 10, 11). Three situations present liability issue (4, 6, 9). When the experienced Human has to justify his action and compare his choice to the machine, it questions his ability to perform his duties. The confidence and acceptance of the tool decrease. At this point it become crucial to understand the reasoning of the tool.

B. Understand the tools

In 2019, a survey [21] of the UK population showed that to date, the most common feeling towards the impact of AI is anxiety. Results highlighted a "markedly negative view of this technology" [21] due to a lack of in technical knowledge. The omission of interaction with futures users when designing AI-based tools is also a strong contributing factor in human mistrust.

The first obstacle to the quick adoption and acceptance of a new tool is therefore the understanding of the AI-based system. The user will have to learn the possibilities offered by

these algorithms, the limitations, the use case they were developed for, etc.

Overall, users must understand the purpose of the tool. Then, the question is about how the result was obtained. In cases of liability issues, this knowledge brings arguments to question or assert the legitimacy of the algorithms output.

The second impediment is the concern of an Agent about how the tool will transform his missions. The decision-making support systems are here to help people. They must not replace the Human (their recommendations are more important), except maybe for low added value tasks. In that case, the tool helps the user to focus on more interesting and challenging tasks, his interest is maintained and so is his concentration. Hence, the Human become more efficient thanks to the AI-based system. In that case, the introduction of new tools is facilitated.

C. Common training for technical teams and managers

Providing news tools to people usually requires training them. The users will benefit from explanations and it will answer the need to "understand the tool". Their comprehension will grow faster and they will be efficient and confident more quickly.

However, they are not the only ones that need to be trained. Managers must also be in the loop of training. In case of an incident, it is them who will attribute the blame. To be fair and save Agents' faith in their tools, they have to know the accuracy of the models, the limitation of the systems, take into account the possible biases, etc. They also need to learn how the tool reacts in real environment. Being trained with their units would help them better apprehend their teams' operations.

D. Profession-adapted Human-System interface

Though the same AI algorithms could be used in different situations, the visualisation of the system's output should be adapted to every application [22]. For example, an AI-based maintenance tool should not show the same information given that the user wants to intervene on the rail network or the electrical installation Easy visualisation, recommendation understanding and data interpretability are essential for a field user.

IV. Conclusion

AI is a new factor of production and unveils unprecedented opportunities for value creation. It has the potential to double economic growth rate across some of the world's economic giants. One of the biggest challenges in the application of intelligent decision support systems to real problems is confidence in autonomous systems. Now, industrials and researchers need to work on questions such as: "What decisions are we willing to permit computer systems to make autonomously?", "Will we allow autonomous systems to make decisions and act on that decision, and under what conditions?", "Do we really believe autonomous systems to act in our best interests?"

The development of an ethical framework for AI systems is mandatory and would introduce and boost confidence in products and services developed for industries. In addition, the design of AI-based tools should incorporate transparency, explicability and interpretability. This process would also improve AI social acceptance.

References

[1] N. Chen, B. Ribeiro and A. Chen, "Financial credit risk assessment: a recent review", Artificial Intelligence Review, col.45, n°1, pp.1-23, 2015

[2] B. Kim, R. Khanna and O. Koyejo, "Examples are not Enough, Learn to Criticize! Criticism for Interpretability", Advances in Neural Information Processing Systems 29, NIPS, 2016

[3] J. Kingston, "Artificial Intelligence and Legal Liability", Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV, pp. 269–279, 2016

[4] H. Lakkaraju and C. Rudin, "Learning Cost-Effective and Interpretable TreatmentRegimes for Judicial Bail Decisions", 30th Conference on Neural Information Processing Systems NIPS, 2016

[5] F. Jiang et al., "Artificial intelligence in healthcare: past, present and future", Stroke and Vascular Neurology, doi:10.1136/svn-2017-000101, vol.2, pp.230-243, 2017

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences", arXiv Preprint arXiv:1706.07269

[7] E. Thelisson, "Towards Trust, Transparency, and Liability in AI/AS Systems", Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), pp. 5215–5216, 2017

[8] L. Dodds, "Chinese businesswoman accused of jaywalking after AI camera spots her face on an advert", The Telegraph, 25/11/2018

[9] C. Rudin and S. Ertekin, "Learning customized and optimized lists of rules with mathematical programming", Mathematical Programming Computation, vol.10, n°4, pp.659-702, 2018

[10] A. Sethumadhavan, "Trust in Artificial Intelligence", Ergonomics in Design, https://doi.org/10.1177/1064804618818592, vol.27, n°2, pp.34–34, 2018

[11] K. Siau and W. Wang, "Building Trust in Artificial Intelligence, Machine Learning, and Robitics", Cutter Business Technology Journal vol. 31, n°2, pp. 47–53, 2018

[12] D. Wakabayashi, "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam", The New York Times, 19/03/2018

[13] M. DeCamp and J. C. Tilburt, "Why we cannot trust artificial intelligence in medicine", The Lancet Digital Health, https://doi.org/10.1016/S2589-7500(19)30197-9, vol.1, n°8, PE390, Dec. 2019

[14] European Commission, "Liability for Artificial Intelligence and other emerging digital technologies", Report from the Expert Group on Liability and New Technologies, 2019

[15] A. M. Nascimento et al., "A Systematic Literature Review About the Impact of Artificial Intelligence on Autonomous Vehicle Safety", IEEE Transactions on Intelligent Transportation Systems, pp.1-19, 2019

[16] D. Tobey, "Explainability: where AI and Liability meet", DLA PIPER, https://www.dlapiper.com/fr/france/insights/publications/2019/02/explainability-where-ai-and-liability-meet/, 2019

[17] B. W. Smith, "Ethics of Artificial Intelligence in Transport", (24/02/2019). The Oxford Handbook of Ethics of Artificial Intelligence (Markus Dubber, Frank Pasquale & Sunit Das, eds., 2020 Forthcoming).

[18] Y. Zhang, Q.Vera Liao and R. K. E. Bellamy, "Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making", Conference on Fairness, Accountability, and Transparency (FAT* '20), 2020

[19] Alain Bensoussan, Colloque Justice et Sécurité, les défis de l'IA, 2019

[20] S. Wachter, B. Mittelstadt and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR", Harvard Journal of Law & Technology, 2017

[21] S. Cave, K. Coughlan, and K. Dihal, (2019, January). " 'Scary robots': examining public responses to AI". AIES Proceedings Conference on AI, Ethics, and Society, pp.331-337, 2019

[22] J. J. Dudley, and P. O. Kristensson, "A review of user interface design for interactive machine learning", ACM Transactions on Interactive Intelligent Systems (TiiS), vol.8, n°2, 8.