# Building Trust to AI Systems Through Explainability. Technical and Legal Perspectives

Grzegorz J. NALEPA[a,1] and Michał ARASZKIEWICZ[a]
Sławomir NOWACZYK[b] and Szymon BOBEK[a]
[a]*Jagiellonian University*
[b]*Halmstad University*

**Abstract.** In this position paper we discuss two perspectives on explainability of AI systems: technical and legal one, and we investigate how the two perspectives should be integrated to develop trust in the AI systems. We consider trust building as a process that should reflected in the design process of AI systems.

**Keywords.** trust, explainability, liability



## 1. Introduction

Providing explanations for decisions made by AI systems (also called Intelligent Systems, IS) is commonly considered as crucial for the trust and social acceptance of AI. In our view explainability does not simply provide/create trust, instead it serves to build trust. In other words, trust building is a sequential, iterative and interactive process that develops over time, and in relation to a specific user.

The catalogue of factors important for the process of trust building is extensive and diversified. According to the Ethics Guidelines for Trustworthy Artificial Intelligence (AI) - the document elaborated by the High-Level Expert Group on Artificial Intelligence (AI HLEG) seven requirements for the trustworthy AI systems should be listed: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being and (7) accountability [1]. We share the view that these requirements are crucial for the design and implementation of any process of building trust to any IS.

Our motivation for this position paper is to consider the role of explainability of AI in the trust building process, in a much needed synergistic perspective, both legal and technical.

[1] Corresponding Author, Grzegorz J. Nalepa, Jagiellonian University, Gołębia 24, 31-007, Kraków, Poland, gjn@gjn.re.

## 2. Explainability of AI Matters

In 2016 DARPA announced the program DARPA-BAA-16-53 on Explainable Artificial Intelligence (XAI) [2]. This was a response to a growing concern regarding the use and development of certain IS. Furthermore, possible legal consequences of their applications were grave. The Agency acknowledges the problems with models built with modern machine learning (ML) techniques. The main challenge is the tension between model performance and explainability. DARPA asserts that there is a clear tradeoff between both. On the other hand, in practical applications, there is a growing need for explainability, consequently there should be a kind of a balance between these two.

Although the DARPA XAI program sparked many discussions in the AI community, the problem of XAI is not new. Not only the early symbolic IS were explainable, but also the research on XAI-related topics has been around for almost 15 years. Paradigms like explanation-aware computing were proposed over a decade ago. In fact now, the ML community faces the challenge that the knowledge-based systems community solved long time ago. Apparently, it is the time now, for these two approaches to work together on delivering hybrid solutions. The process of building of subsymbolic Machine Learning (ML) models for decision making requires a large amount of training data, often prone to implicit biases that have an impact on the resulting system. Furthermore, the operation of many of these models is often difficult to interpret by different actors, including not only the users, domain experts , but even the designers. Therefore, these models are commonly referred to as "black-box AI".

Trust in IS, especially ones that include such black-boxes has become a challenge that needs to be addressed on more than just technical level to contribute to their social acceptance. In next sections we provide a transition from a technical perspective to the legal one.

## 3. Trust Building Through Explainability – Technical Perspective

First of all, we believe that the explainable IS should never be considered as stand-alone, abstract entity. Instead, we should always consider human-in-a-loop setting – moreover, a particular human. As an example, in the medical domain, this human can be a data scientist building a model, a doctor participating in the design of the system, a doctor using the system to gain insights on a diagnosis, a doctor trying to come up with the best treatment, or a patient seeking interpretation of a medical decision. Each of them needs different explanations, and each of them will develop their own trust relationship towards the system [3].

However, these relationships are not independent – anymore than these actors are independent. For example, if the doctor trusts the system and uses the system to collaboratively create a treatment plan, it will be easier for the patient to trust the system as well – and, we hypothesise, it will also be easier for the patient to accept the system's help in follow up on implementing this treatment, receiving adherence advice, etc. Trust relationship takes time and effort to build, but it also brings long-term benefits.

We envision scenarios where a number of human actors collaborate towards a common goal. In such groups, there is an implicit assumption of some level of trust. In

our vision, IS could become a new actor to support the same end goal, which implies they need to "earn" their own share of trust, in order to be able to contribute to the overall common goal.

Effective collaboration between actors is thus necessary to achieve success, and such collaboration depends on the right communication – IS is necessarily part of this. In this regard trust is crucial in two ways. First, it is impossible to communicate clearly and openly without some level of trust; second, trust can only be built through understandable and clear communication – which is a challenge for IS, and requires novel approaches towards explainability. The explanation mechanisms involved in this process should therefore allow for two-way information flow between different parties involved in the design, implementation and exploitation phases. This could be achieved with different knowledge mediation techniques.

Moreover, the needs and expectations of different groups of users should be taken into account. The explanations IS provides should take into account the different needs of groups of users, both in terms of their personal subgoals as well as different levels of knowledge and cognitive abilities. In fact, we must consider different users (e.g., doctors and patients) using and improving the IS in a collaborative manner. The personalisation layer of an IS system must never be "finished", instead it should continuously adapt to the needs of users throughout IS lifetime. This is why, we argue, the notion of the AI-based system must not be considered only on a technical level. To summarize, model building, evolution and explanation provisioning should also be adapted to specific include domain-specific aspects.

We propose the concept of AI-enhanced "collaborative system" which includes a range of technical components for decision making, explanation provisioning, as well as human experts both using and also improving them. The goal is, ultimately, for the IS to offer certain services (e.g. medical diagnosis), but not "to" other (human) users, but "in collaboration" with (human) users. In fact, users could be different groups of patients, but also other doctors, e.g. of another specializations. Both the decisions, and explanations of AI models and human experts contribute to the efficiency of the task execution as well as trust in the system as a whole. In this context it is crucial that, for example, the IS system and human doctors provide explanations that are consistent. In this setting we argue that the primary function of the explanation related to IS is in fact not to explain the very operation of the model. Instead, explanation is the primary means for IS to contribute to trust building.

Explanation provisioning is also an interactive process that involves different actors or stakeholders. As such it should be delivered in an adaptive, contextualized, and personalized manner. Explanation should always be personalized and the explanation building process should take into account the prior interaction with the given user.

We acknowledge the fact that real life operation of IS is specific domains has an important legal dimension. Each practical implementation and deployment of an IS should take legal consideration into account. These legal aspects vary depending on the context of the domain. They might include certain norms, specific professional regulations and user-specific laws (e.g. regarding privacy). Moreover, an important legal dimension regards liability of the system. The assessment and interpretation of liability is in fact introduced where the stakeholders have only limited trust the system (and possibly to each other). As such measuring trust should always be provided together with assessment of possibly legal liability of the system as a whole, but maybe also the individual stakeholders.

We propose a three phase life-cycle approach for such AI-enhanced systems. During the initial design of the system different operation scenarios should be developed. They include both the decision making aspect as well as corresponding feasible explanations. Explanation elicits not why the system made a decision but how this decision improves trust of the user in the decision-making process. Moreover, specific requirements of all the possible stakeholders should be taken into account and modeled properly. The main design phase includes building the models for domain-specific decision making and support, built with the collaboration of domain experts as well as AI engineers, as well as corresponding models for explanation suitable for expected groups of users. The operation phase involves not only the use of models of both types, but also their iterative evaluation and improvement developed in a process of collaborations of experts and the users.

In this process we assess the effectiveness and usefulness of the explanation at the human and the technical levels, by evaluating how efficiency of decision making, but also transparency, and trust are enhanced. We embrace the fact that explanation is required at different levels and in different dimensions for different stakeholders with different levels of technical knowledge, and in different application domains. In the whole cycle lawyers can be included to identify duties and liability of actors, participate in the certification of certain models, and iteratively assess the liability of the system during its development and operation.

## 4. Legal Aspects

During the recent years the legal issues concerning the development and operation of IS has been recognized as requiring solutions. Determining the solutions to the emerging problems is a necessary prerequisite of building trust towards AI systems. One should begin with an observation on the role of trust in the system of law. Disregarding the numerous differences across legal cultures and jurisdictions, it is a generally held opinion that the law should promote and protect trust. The principle of protection of trust is applicable in both horizontal and vertical legal relations. In civil law, a party may generally rely on another party to a legal relationship and if this relation of trust is breached, it may lead to legal liability of the breaching party. In public law, the citizen is entitled to hold trust in the State and the law enacted by it: if the legal regulation is overly vague, unpredictable or subject to surprising or too frequent amendments, legal consequences may follow, including declaring the trust-breaching regulation unconstitutional. The principle of protection of trust is particularly important in legal relations characterized by asymmetry of knowledge or power among the parties.

A question arises, how it is determined that the relation of trust is breached in particular legal relationship. The general answer to this question concerns the notion of reasonable expectations of an entity who relies on another party, within the constraints that are characteristic for the given domain of law and sphere of societal life. Undoubtedly, the concept of reasonable expectations has strong normative underpinnings and involves a vast amount of commonsensical knowledge concerning what patterns of behavior are deemed typical or acceptable in a given context. An informed party to a legal relationship will also typically assume that another party shall adjust its behavior to avoid legal liability. Therefore, we may state that the process of

building trust among the parties to legal relationships involve three important legally relevant prerequisites: (1) assumptions concerning the typical, or expected behavior in a given situation type, (2) normative criteria serving a tools of evaluation of either party's behavior, and their expectations and (3) appropriate liability rules becoming effective in case of breach of trust. These prerequisites play the role of constraints on the process of trust-building between the parties. They may not be sufficient for the development of actual (rather than assumed) trust relation, but they are typically necessary conditions therefor. The problem for the process of trust-building in case of operation of the IS is that each of these prerequisites may be deemed problematic.

To begin with, the problems of legal liability resulting from the operation of the IS are the subject of vivid debate. Whilst the very idea of ascription of liability to autonomous agents is currently regarded as one of viable options [4], the specific issues concerning the chosen regime of liability and the choice and interpretation of applied liability conditions. The classical legal categories such as fault, negligence and adequate causal link need reinterpretation in the context of operation of IS [5]. It should also be emphasized that liability actualizes itself in case of breach of certain norm following from legal regulation or from a contract. Therefore, it is necessary to investigate the content of applicable legal norms in order to determine the potential grounds of liability. In this connection one of the most important topics is whether the subjects of law are vested with a right to explanation and how the content of such right should be understood.

Some authors point out that the right to explanation is expressed in the GDPR, where a few options are indicated as the source of this right, while another authors openly contest this claim [6]. The issue of actual legal source of right to explanation (if any) is therefore currently a subject of debate. It is more fruitful to consider what is the potential content of this right and what claims could follow from its breach. Certain important distinctions have been already discussed in the literature of the subject, like the difference between the explanation of the systems functionality vs. explanation of a specific decision, and the difference between explanation ex ante and ex post [6]. However, more attention is needed to address the notion of explanation used in the context of AI explainability. Obviously, explainability and explanation have already attracted so much attention in different communities that they begun to function, to certain extent, as hermeneutical concepts, used by the member of community to better understand their own actions and attitudes. Therefore, it would not be reasonable to postulate one "right" definition of explanation used in the context of AI operation. However, the formulation of right to explanation requires delimiting its scope, at least in certain respects. In our view, in this connection the technical explanation - i.e. the description of the systems' functionalities and mechanisms of inference, should be distinguished from the normative explanation: presentation of rules and value the system is (or should be) bound to follow. In other words, normative explanation may be understood as potential justification of the system' operation (e.g. automated decision). In addition, normative explanation should encompass the normative boundaries of the systems' operation and the information on the consequences of breaching of these norms.

The notion of normative explanation should serve as the basis for the forming of reasonable expectations of users and other stakeholders. The design of the IS in order to meet the expectations would be a considerable factor to the process of trust building on both general and particular level. The constraints following from the normative explanation should serve as the criteria of evaluation of typical and non-typical

behavior of the IS and as the basis for introducing appropriate modifications. The notion of normative explanation would also foster accountability of the systems' operators and the compliance with fairness and nondiscrimination requirements. Arguably, normative explanation is a necessary condition for the process of trust-building between the IS and the non-technical users of systems as well as the general public.

## 5. Summary and Outlook

In this short position paper we considered the relation between trust and explainability We consider trust or trustworthiness not a property of an AI system that can be provided. Instead we propose to consider a trust building process, related to the life-cycle of AI system involving different actors, such as designers, users, etc. In this is iterative process, contextualized explanation provisioning and normative explanation play a crucial role.

## References


[1] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI,* https://ec.europa.eu/futurium/en/ai-alliance-consultation, 2019

[2] DARPA, Broad Agency Announcement – Explainable Artificial Intelligence (XAI), DARPA-BAA-16-53, August 10, 2016.

[3] Grzegorz J. Nalepa, Sławomir Nowaczyk, *Towards building trust between users and Artificial Intelligence systems*, (unpublished draft), 2019.

[4] Jaap Hage, Theoretical foundations for the responsibility of autonomous agents, Artificial Intelligence and Law **25** (2017),, 255-271

[5] Grzegorz J. Nalepa, Michał Araszkiewicz, Legal Responsibility of Intelligent Systems, in: An Introductory Guide to Artificial Intelligence for Legal Professionals, María Jesús González-Espejo, Juan Pavón (Eds), Wolters Kluwer Law Intl. 2020, forthcoming.

[6] Sandra Wachter, Brent Mittelstadt, Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law **7** (2017), 76-99.