Explainability as a User Requirement for Artificial Intelligence Systems

Mladan Jovanović, Singidunum University
Mia Schmitz, EVOTEK

As the capabilities of artificial intelligence (AI) systems constantly grow, so too does their complexity. The explainability toward their users is gaining attention, becoming a requirement that these systems should satisfy. We articulate user requirements for explainable AI systems.

Digital Object Identifier 10.1109/MC.2021.3127753
Date of current version: 14 February 2022

Humans excel at common sense, empathy, flexibility, and creativity, whereas machine algorithms are faster, cheaper, more efficient, scalable, and consistent in dealing with large amounts of data. Currently, there is a consensus among researchers from different disciplines that cooperative[12] or hybrid intelligence[16]—that being human–machine symbiosis—is a viable alternative to a successful artificial intelligence (AI) system.

The mutual understanding between humans and machines is critical for achieving symbiosis, that is, that machines should better learn about humans and humans should better understand what machines have learned about them. However, this goal appears far away, especially as humans are also subject to various biases. For example, our everyday decisions depend on how information relevant to the decision is presented to us (known as the *framing effect*).[13] In this perspective, AI's explainability is emerging as a bottleneck to human–AI integration: The famous Edvard I. Koch quote, "I can explain it to

0018-9162/22©2022IEEE

EDITOR **MARK CAMPBELL**
EVOTEK; mark@evotek.com

you, but I can't comprehend it for you," epitomizes current machine-generated explanations.

"There is a notable transparency gap in machine learning (ML)-supported tools, such as in the financial industry, from banking transactions to money laundering and fraud detection. Visual tools for bridging such gaps with explanations for businesspeople and data scientists are a necessity," posits Krishna Gade, founder and CEO at Fiddler AI, a leader in explainable monitoring for AI systems.[17]

As the capabilities of AI systems constantly grow, so too does their complexity. In parallel, humans began struggling to understand their automated decisions. The explainability toward their users is gaining attention, becoming a requirement that these systems should satisfy. It is challenging to develop a clear yet shared definition of explainability because the term does not originate from computer science (namely, AI). Instead, it is tackled by multiple disciplines such as social science, psychology, ethics, philosophy, and medicine, each having its own perspective of the term.[1,7–9] Some practices already recognize explainability, such as the European Union's General Data Protection Regulation Article 13-2-f,[4] which declares the right for comprehensible information about the logic of automated decisions (namely, providing explanations).

Different approaches have focused on the qualities that an AI explanation should exhibit. For instance, Miller[1] highlights contrast (presenting with counterfactual and/or counterintuitive events), selectivity (containing a few necessary and sufficient reasons), sociability (considering beliefs of both the explainer and the explained), and causality (describing causes, not their likelihoods).

On the other side, Sundararajan et al.[2] list desirable visual properties of explanations such as integrity (displaying features that contribute the most to predictions, including positive and negative attributions), coverage (showing a significant fraction of the most important features), clarity (presenting essential features clearly), and separation (the visual detachment of distinct features).

DATA-DRIVEN EXPLAINABILITY

Medical AI acknowledges that explainability depends on data,[3] and scientists stress establishing quality recommendations and standards for creating training data sets that involve diverse participants at scale. Similarly, Schneiderman[5] advocates for user interfaces that manually explore AI systems' decisional spaces by modifying influencing variables. They could reduce uncertainty, prevent confusion, and make AI systems trustworthy to their users. Different users should manipulate input parameters at desired levels of detail to reach a satisfactory understanding of the outcomes.

Like Schneiderman, Sheth et al.[11] propose adding explicit knowledge on explainability during the design and development of AI systems to facilitate their adoption. Thus far, the focus has been on explaining the system's decisions in the production phase. From now on, developers should bring structured domain knowledge (understandable by humans) into an AI system design (for example, through model features and algorithm steps).

ALGORITHM-DRIVEN EXPLAINABILITY

There is a known tradeoff between automated learning's performance (accuracy) and explainability.[6] For example, the higher accuracy of deep learning models makes it more difficult to understand how inputs produce outputs.

AI algorithms ultimately depend on training data concerning performance, accuracy, and the origins of explanations. Nevertheless, the visualization and analytical tools used for data collection, cleaning, and labeling can help discover and understand errors and missing data, clusters, gaps, nonuniform distributions, and anomalies. "ML model monitoring visual analytics provide our customers, interactively, with various insights in the production phase, including performance, feature and outcome comparison, bias exploration for fairness, and decision explanation," describes Gade.

It is challenging to develop a clear yet shared definition of explainability because the term does not originate from computer science (namely, AI).

The related issues are algorithmic fairness and bias, which require careful composition and analysis of the training data. To ensure equity, data categories should be balanced and treated equally (that is, gender, age, race, skin color, income, origin, and education). All users should benefit; minority groups should not be underrepresented. On the other hand, AI algorithms discriminate by their very nature. Their design premise is to personalize services to specific groups or individuals to improve outcomes. The result is that they treat different users differently. The situation requires careful consideration of the

aforementioned aspects, with iterative testing and limitations transparent to end users.

OPERATIONALIZING EXPLAINABILITY

Instead of just defining it, we should operationalize explainability through a set of attributes as a prospective AI system's properties. This approach can express and quantify explainability for a specific type of system under development or a problem domain. Seeing explainability as a requirement for an AI system development, we do not aim for how it should be formulated but focus on its consequences and why it is necessary to introduce it initially in an AI system's development.

Explainability is context dependent, and properties of applications determine the approaches to designing explainable AI systems. Accordingly, we propose explainability-related attributes to characterize the application domains and guide their design and development:

- *Risk*: describes the unwanted consequences that a specific AI decision can cause with end users (namely, failure in task execution or behavior contrary to user goals).
- *User*: concerns primary target users' levels of the application's domain knowledge and skills.
- *Timeline*: defines the timeliness of AI decision making, whether the system should make real-time decisions or expose a relaxed response time.
- *Automation*: relates to the level of autonomy in decision making, that is, the degree of human assistance or intervention needed to support the activity.

We extracted common AI application domains[7] and described their decision making with the attributes' values (see Table 1). The AI domain landscape reveals expert and nonexpert groups being targeted equally.

Some applications (autonomous driving and clinical diagnosis) introduce considerable risk, where a failure in decision making can lead to severe consequences. In contrast, others (language translation and web searches) do not cause adverse effects in the occurrence of errors. Specific applications are expected to deliver immediate decisions (fall prevention and conversational agents), whereas others do not respond in real time

TABLE 1. An AI decision-making design space concerning explainability-related attributes.

Attribute domain	User		Risk		Timeline		Automation	
	Nonexpert	Expert	Low	High	Real time	Relaxed time	Low	High
Autonomous driving	✓	—	—	✓	✓	—	—	✓
Financial decision making	—	✓	—	✓	—	✓	—	✓
Clinical diagnosis	—	✓	—	✓	—	✓	✓	—
Targeted advertising	✓	—	✓	—	—	✓	—	✓
Delivery drones	✓	—	✓	—	—	✓	✓	—
Conversational agents	✓	—	✓	—	✓	—	—	✓
Language translation	✓	—	✓	—	—	✓	—	✓
Law decision making	—	✓	—	✓	—	✓	✓	—
Crime decision making	—	✓	—	✓	—	✓	✓	—
Fall prevention	✓	—	—	✓	✓	—	—	✓
Remote education	✓	—	✓	—	—	✓	—	✓
Digital manufacturing	—	✓	✓	—	✓	—	—	✓
Contact tracing	✓	—	—	✓	✓	—	—	✓
Cyberdefense	—	✓	—	✓	✓	—	—	✓
Agriculture	—	✓	✓	—	—	✓	—	✓
Web search	✓	—	✓	—	✓	—	—	✓
Environmental decision making	—	✓	—	✓	—	✓	—	✓

(law decision making and agriculture). Finally, certain applications are highly autonomous (digital manufacturing and targeted advertising), while others require human supervision (delivery drones and crime decision making).

Presented with the attributes' values for applications, we can define suitable explainability strategies to address the related issues. For example, different target groups (expert versus nonexpert) will require usable explanations concerning their language and coverage (breadth and depth of information).

Higher risk applications should help users understand their decisions by initiating explanations to decrease a potential risk, while lower risk applications may explain on demand. Real-time applications should explain while not interfering with the decision-making process as they pursue strict timeliness (that is, upon completion of a set of required actions). Those with a relaxed response time can provide explanations on the fly alongside communicated decisions.

Highly automated applications can bring uncertainty to their users, so they should have richer explanations of the logic behind automated elements. A lower level of automation assumes higher user involvement. Therefore, the explanations should focus on the automated, nontransparent parts of the decision-making process.

IMPLICATIONS FOR FUTURE AI SYSTEMS

Addressing explainability-related challenges requires a collective effort concerning related and emerging phenomena in AI. We formulate user requirements for explainable AI systems based on these implications.

Collaborative workflow for valid, AI-generated explanations

Humans should evaluate performance and explainability along with an AI system's development and deployment. However, as humans, we are highly susceptible to biases (for example, confirmation bias, peak-end rule, recall bias, and prior beliefs) when reasoning about phenomena that surround us[14] due to cultural and personality differences. To mitigate such biases, different actors with a balanced mix of knowledge and skills can cross-check system-generated explanations iteratively to reconcile different perspectives and reach a consensus in explaining the system's decisions.

Transparent explainability

Transparency assumes a certain level of access to the AI system's data and algorithmic logic. There should be a tradeoff between openness and privacy in proposing satisfactory explanations. A successful balance requires structured knowledge about the specific data set. It is critical to understand the nature and interdependencies among data items so that exposing some through explanation does not violate any level of privacy.

Bias explainability

In general, biases exist in different stages of AI system development, including problem statements, data collection, algorithms, and testing.[12] The examples include favoring certain data instances compared to others (data related) or making a sample that does not represent the analyzed population (algorithm related). Avoiding such pitfalls will ensure fairness so that all categories are equally represented (in data) and analyzed (by algorithms). Constructing and documenting representative yet diverse data sets could reduce these issues. At the same time, identifying and describing the presence of biases through explanations could raise the awareness of such phenomena in the existing data sets and algorithms. Explainability as a service is a significant step toward this goal.[18]

Ethical explainability

Before offering explanations, an AI system should consider to whom and how it should communicate them, which is essential for vulnerable and sensitive user groups. For example, persons with special needs should not be harmed or stigmatized with the form and content of presented explanations.

> Explainability is context dependent, and properties of applications determine the approaches to designing explainable AI systems.

Open explainability for accountability and trust

We build trust with familiar and understandable things. For instance, explainable AI in health care is crucial for medical applications' accountability.[15] Knowing how the recommender personalizes and why it suggests certain items (for example, a treatment or a medicine) strengthens the users' bond with the application. The system should formulate explanations in a user-friendly language. Additionally, the system could feature an open dialogue with the users if they do not understand or are not satisfied with explaining a particular decision. This dialogue can facilitate the AI system's learning skills about users and make it more trusted from their voice being heard.

AI needs discoveries from cognitive sciences during its search to build systems

capable of explaining their decisions, like humans. Although AI need not replicate human reasoning, a deeper understanding of the human mind can critically advance the explainability of machines. A step forward in producing intuitive and satisfactory explanations is implementing the core mechanisms of human reasoning, including knowledge about objects and events and their causal relations.[11,13] The applications should be teachable and learnable using scarce and incomplete knowledge to display human-friendly explanatory behavior. In such a future, we will learn from machines as we do from other people.

It is critical to understand the nature and interdependencies among data items so that exposing some through explanation does not violate any level of privacy.

REFERENCES

1. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.
2. M. Sundararajan, S. Xu, A. Taly, R. Sayres, and A. Najmi, "Exploring principled visualizations for deep network attributions," in *Proc. 2019 Joint ACM IUI Workshops Co-Located 24th ACM Conf. Intell. User Interfaces*, p. 11.
3. H. Muller, M. T. Mayrhofer, E.-B. Van Veen, and A. Holzinger, "The ten commandments of ethical medical AI," *Computer*, vol. 54, no. 7, pp. 119–123, 2021, doi: 10.1109/MC.2021.3074263.
4. "General data protection regulation," Intersoft Consulting, Fremont, CA, USA. Accessed: Sep. 15, 2021. [Online]. Available: https://gdpr-info.eu/art-13-gdpr/
5. B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, pp. 1–31, 2020, doi: 10.1145/3419764.
6. M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, pp. 68–77, 2019, doi: 10.1145/3359786.
7. G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," Inf. Fusion, vol. 76, pp. 89–106, Dec. 2021, doi: 10.1016/j.inffus.2021.05.009.
8. A. Adadi and M. Berrada, "Peeking inside the black box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52,138–52,160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.
9. A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Rev., Data Mining Knowl. Discovery*, vol. 9, no. 4, p. e1312, 2019, doi: 10.1002/widm.1312.
10. A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel, "Cooperative AI: Machines must learn to find common ground," *Nature*, vol. 593, no. 7857, pp. 33–36, 2021, doi: 10.1038/d41586-021-01170-0.
11. A. Sheth, M. Gaur, K. Roy, and K. Faldu, "Knowledge-intensive language understanding for explainable AI," *IEEE Internet Comput.*, vol. 25, no. 5, pp. 19–24, Sep./Oct. 2021. doi: 10.1109/MIC.2021.3101919.
12. R. Srinivasan and A. Chander, "Biases in AI systems," *Commun. ACM*, vol. 64, no. 8, pp. 44–49, 2021, doi: 10.1145/3464903.
13. S. Plous, *The Psychology of Judgment and Decision Making*. New York, NY, USA: McGraw-Hill, 1993.
14. J. B. T. Evans, *Bias in Human Reasoning: Causes and Consequences*. New Jersey, USA: Lawrence Erlbaum Associates, 1989.
15. B. Babic, S. Gerke, T. Evgeniou, and I. G. Cohen, "Beware explanations from AI in health care," *Science*, vol. 373, no. 6552, pp. 284–286, 2021, doi: 10.1126/science.abg1834.
16. Z. Akata *et al.*, "A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence," *Computer*, vol. 53, no. 8, pp. 18–28, 2020, doi: 10.1109/MC.2020.2996587.
17. M. Jovanovic, M. Schmitz, and M. Campbell, Interview with K. Gade, Aug. 24, 2021.
18. "Explainability service," Fiddler AI, Palo Alto, CA, USA. Accessed: Sep. 15, 2021. [Online]. Available: https://poweredby.fiddler.ai/

MLAĐAN JOVANOVIĆ is an assistant professor at Singidunum University, Belgrade, 11000, Serbia. Contact him at mjovanovic@singidunum.ac.rs.

MIA SCHMITZ is with EVOTEK, San Diego, California, 92121, USA. Contact her at mschmitz@evotek.com.