5.3.9.1 Human-Artificial Intelligence System

A human-AI system includes both computational elements and a human user who must come together to accomplish a purpose.

5.4 What is Explainable AI?

Explainability is a concept that stands at the crossroads of numerous fields of active AI research, with an emphasis on the following domains.

5.4.1 Fairness

Can we verify that choices taken by an AI system were done consistently?

5.4.2 Causality

Can one learn a system from facts that not only makes the right predictions but also offers some understanding of the core events?

5.4.3 Safety

Can we have confidence in the reliability of our AI system without recognizing how it makes its presumptions?

5.4.4 Bias

How can we be confident that the AI system hasn't picked up a distorted view of the world due to flaws in the training data or objective function?

5.4.5 Transparency

Everyone has a right to be informed about changes that influence us in ways, formats, and languages that we comprehend.

An XAI, also known as a "Transparent AI" or "Interpretable AI", is an AI whose activities are simple for humans to comprehend and evaluate. A civic privilege to explain can be implemented using XAI.

5.5 Need for Transparency and Trust in AI

The black box AI systems have found their way into many of today's modern implementations. Transparency and explainability are not critical requirements for machine learning models used as long as the overall efficiency of these systems is adequate. Even if these systems fail, the implications are unexceptional. As a result, the necessities for trust and openness in these types of AI systems are relatively low. The scenario is different in safety–critical applications. In this case, the opaqueness of ML techniques may be a restricting or indeed rejecting component. Particularly when a single misjudgment can endanger human life and health or lead to significant revenue damages, depending on an information system with unintelligible logic will not be an alternative. This lack of transparency is among the causes why the application of machine learning to areas such as healthcare is extremely careful than its application in the consumer, electronic commerce, or media industries.

5.6 The Black Box Deep Learning Models

The method of developing interpretations for AI system behavior will vary based on the type of ML techniques used: techniques that produce implicitly decipherable models vs deep learning algorithms that are intricate information and understanding methods and produce models that are implicitly indecipherable to actual users.

ML techniques such as Bayesian classifiers, decision trees, spare linear models, and additive models produce decipherable models in the sense that model components can indeed be instantly examined to comprehend the model's inferences. These technique makes use of relatively small internals, and also provide visibility and traceability in their decision-making. As long as the model is precise for the classification process, these strategies offer awareness of the AI system's decision-making.

Deep learning algorithms, one on either side, are a class of machine learning technique that sacrifices clarity and interpretability for the predictability. These techniques are now used to create applications such as consumer behavioral forecasting associated with high inputs, voice recognition, natural language processing, and computer vision.

The lack of transparency and understandability in the Deep Learning Algorithms makes them a black box. The black box model is a model which performs its predictions on its own without explaining anything for humans to understand.

The Black Box Problem occurs when artificially intelligent processor architectures are vague.

Chapter 11
Human-AI Interfaces are a Central Component of Trustworthy AI

Markus Plass, Michaela Kargl, Theodore Evans, Luka Brcic, Peter Regitnig, Christian Geißler, Rita Carvalho, Christoph Jansen, Norman Zerbe, Andreas Holzinger, and Heimo Müller

Abstract This chapter demonstrates the crucial role that human-AI interfaces play in conveying the trustworthiness of AI solutions to their users. Explainability is a central component of such interfaces, particularly in high-stake domains where human oversight is essential: justice, finance, security, and medicine. To successfully build and communicate trustworthiness, a user-centered approach to the design and development of AI solutions and their human interfaces is essential. In this chapter, we explain how proven methods for stakeholder analysis and user testing from human-computer interaction (HCI) research can be adapted to human-AI interaction (HAII) in support of this goal. The practical implementation of a user-centric approach is described within the context of AI applications in computational pathology.

11.1 Introduction

The prevalence of Artificial Intelligence (AI) in daily life is ever-increasing. It is integrated into smartphones and consumer goods, transforming the role of the user (Harper et al. 2020) and the human-machine interface. While the traditional human-computer interface simply represents the input-output (I/O) surface of a device (Holzinger 2004), or web page (Ebner et al. 2007), human-AI interfaces transcend the simple I/O paradigm. Besides enabling intelligent interaction via voice or facial recognition, human-AI interfaces can learn from users' behavior, react adaptively, and make predictions about future actions (Holzinger et al. 2022). Accordingly, the scope and challenges of Human–AI Interaction (HAII) research (Xu et al. 2021) differs from that of the traditional field of Human–Computer Interaction (HCI) (Dix et al. 1993). For example: AI chatbots can express human-like communication behavior (Przegalinska et al. 2019); AI-based natural language translation systems show contextual understanding (LeCun et al. 2022); AI-based programs for music co-

M. Plass (✉) · M. Kargl · T. Evans · L. Brcic · P. Regitnig · C. Geißler · R. Carvalho · C. Jansen · N. Zerbe · A. Holzinger · H. Müller
Medical University Graz, Graz, Austria
e-mail: markus.plass@medunigraz.at

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_11

creation can generate non-deterministic output (Louie et al. 2020); AI systems can collaborate with humans in teams (Calero Valdez et al. 2012; Robert et al. 2016), augment human intelligence (Crisan and Correll 2021; Holzinger 2016), and continuously learn from user behavior (Ortigosa et al. 2014).

AI has the potential to bring about a range of benefits to society, support individual and social well-being, enhance innovation and progress, and help to realize sustainable development goals (European Commission, Directorate-General for Communications Networks, Content and Technology, 2019). Regarding the case in point, AI applications in healthcare support personalized and precision medicine, drug development, critical surgeries, clinical decision and diagnosis support, medical image processing, and early detection of disease (Rajpurkar et al. 2022).

However, alongside these opportunities, the broadening application of AI brings novel risks and side effects. Fear of negative consequences, whether misplaced or valid, may also result in underuse and/or over-regulation of AI systems, leading to opportunity costs for individuals and societies (Floridi et al. 2018). Therefore, both benefits and risks must be addressed adequately to give people and societies the confidence to accept AI-based solutions, and to trust in their development, deployment, and usage, even in areas where stakes are high, such as medicine, justice, finance, and security. The trustworthiness of AI systems is a prerequisite for their uptake (European Commission 2021).

According to the *High-Level Expert Group on AI*, established by the European Commission in 2018, trustworthy AI has three components (European Commission, Directorate-General for Communications Networks, Content and Technology 2019):

(a) it should be compliant with the law
(b) it should be robust (i.e., safe, secure, and reliable to not cause unintentional harm)
(c) it should be in alignment with the four ethical principles respect for human autonomy, prevention of harm, fairness, and explicability.

This book chapter illustrates the central role that explainability and human-AI interfaces play in realizing, communicating, and verifying the trustworthiness of AI systems and the importance of a user-centered approach to the design and development of these components. The next section describes regulatory requirements for trustworthy AI and the role of human-AI interfaces in fulfilling these. Section 11.3 discusses explicability as one of the core components of trustworthy AI, and demonstrates that explainable AI is key to building trustworthiness. Section 11.4 explains why a user-centered approach is essential for achieving highly explainable and trustworthy AI systems and introduces stakeholder analysis, personas, and user-testing as valuable methods aiding user-centered design and development of AI solutions. Section 11.5 shows, with the aid of the use-case of AI applications in computational pathology, how these methods can be applied to develop human-AI interfaces that support trustworthiness.

11.2 Regulatory Requirements for Trustworthy AI

As described above, one of the three components of trustworthy AI is compliance with the law (European Commission, Directorate-General for Communications Networks, Content and Technology 2019). The *Artificial Intelligence Act* (European Commission 2021) proposed by the European Commission in 2021 is the first legal framework aimed specifically at fostering AI trustworthiness. It sets out requirements that are mandatory for all AI systems that pose significant risks to the health and safety or fundamental rights of persons (European Commission 2021). Communication to the users is a recurring feature in many of the requirements stipulated. Thus, human-AI interfaces play a central role in the fulfillment of these requirements, as illustrated in the following paragraphs:

Communicate the AI system's intended purpose and associated risks: Point 4 of article 9 'Risk management system' of the Artificial Intelligence Act specifies: "*The risk management measures ... shall be such that any residual risk ... is judged acceptable, provided that the AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse. Those residual risks shall be communicated to the user*" (European Commission 2021). To meet this requirement, human-AI interfaces must clearly communicate to users the intended purpose of an AI system as well as the residual risks associated with its usage.

Communicate the AI system's result and all information needed for its interpretation: Point 1 of article 13 'Transparency and provision of information to users' of the Artificial Intelligence Act states: "*AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and the provider ....*" (European Commission 2021). To support these demands, human-AI interfaces must clearly communicate to users the AI system's output together with all information needed for the correct interpretation of this output.

Communicate instructions for use of the AI system: Point 2 of article 13 of the Artificial Intelligence Act demands: "*AI systems shall be accompanied by instructions for use ... that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users*" (European Commission 2021); and point 3 of article 13 of the Artificial Intelligence Act specifies in detail all information that shall be included in these instructions for use, such as for example "*identity and the contact details of the provider ... characteristics, capabilities, and limitations of performance of the high-risk AI system ... human oversight measures ... expected lifetime of the high-risk AI system and any necessary maintenance and care measures to ensure the proper functioning of that AI system ....*" (European Commission 2021). Human-AI interfaces can help to fulfill this requirement either by providing information on how to access the instructions for use or by conveying all information constituting instructions for use to the user directly.

Support human oversight of the AI system: Article 14 of the Artificial Intelligence Act calls for 'human oversight', and explicitly mentions the important role of the human-machine interface as a tool enabling humans to complete this task: "*AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools that they can be effectively overseen by natural persons during the period in which the AI system is in use*" (European Commission 2021). Point 4 of article 14 describes in detail all functionalities and features that human-AI interfaces must provide to support human oversight: According to point 4 of article 14, AI systems *"shall enable the individuals to whom human oversight is assigned to do the following:*

(a) *fully understand the capacities and limitations of the high-risk AI system ...;*
(b) *remain aware of the possible tendency of automatically relying or over-relying on the [AI system's] output ('automation bias'), ...*
(c) *be able to correctly interpret the AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;*
(d) *be able to decide, in any particular situation, not to use the AI system or otherwise disregard, override or reverse the output of the AI system;*
(e) *be able to intervene on the operation of the AI system or interrupt the system"* (European Commission 2021).

Support the AI system's cybersecurity: Article 15 'Accuracy, robustness and cybersecurity' of the Artificial Intelligence Act requires that "*AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of ... cybersecurity*" (European Commission 2021). Human-AI interfaces have important functions with respect to the AI system's vulnerability to cyber-attacks, for example, by enabling user authentication or by conveying security alerts to the user.

Communicate the AI system's accuracy: Article 15 'Accuracy, robustness and cybersecurity' of the Artificial Intelligence Act specifies "*AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy .... The levels of accuracy and the relevant accuracy metrics ... shall be declared in the accompanying instructions of use*" (European Commission 2021). This means that human-AI interfaces shall always provide the user with information about the system's current accuracy so that the user can assess whether or not this level of accuracy is appropriate for the task at hand.

Support robustness of the AI system and prevent user errors Point 3 of article 15 of the Artificial Intelligence Act calls for an AI system's robustness and fault tolerance also specifically with respect to user errors: *"AI systems shall be resilient as regards errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems"*(European Commission 2021). To fulfill this requirement, the human-AI interface on the one hand plays an important role in providing clear but

graceful feedback to the user when a user-error has occurred. On the other hand, as we know from usability research (Norman 2013) that the human-machine interface is also crucial for preventing the user from both conscious mistakes and unconscious slips during their interaction with the system.

11.3 Explicability—An Ethical Principle for Trustworthy AI

Compliance with the law is only one of the three components of trustworthy AI. Another is adherence to ethical principles and values, of which four are explicitly named by the High-Level Expert Group on AI: three traditional bioethics principles (human autonomy, prevention of harm, and fairness), which are in turn based on those described in the Charter of Fundamental Rights of the European Union (European Parliament, the Council and the Commission 2012), and a fourth: *explicability* (European Commission, Directorate-General for Communications Networks, Content and Technology 2019).

Explicability is a new ethics principle specifically relating to AI. It relates to the tendency for AI systems to act on the basis of complex internal processes that are invisible and/or unintelligible to humans (Floridi et al. 2018), rendering their decision-making processes difficult to understand, interpret, and explain (Holzinger et al. 2017). These are crucial issues for trustworthiness, validation, and acceptance of AI (Ziefle et al. 2013). According to Floridi et al. (2018), explicability recognizes the need to understand and hold to account the decision-making processes of AI.

To address the challenge of explicability, the field of *explainable AI* (XAI) research strives to provide insights into how a given AI model works and why it generates a particular result (Holzinger et al. 2018; Longo et al. 2020). There is a jumble of terms related to this concept in the XAI literature: with the terms explainability and interpretability often being used interchangeably (Zhou et al. 2021). Moreover, a variety of terms, including *transparency*, *accountability*, *intelligibility*, *understandability*, and *interpretability*, *comprehensibility* are used, sometimes interchangeably, sometimes with subtle differences in meaning that vary according to author. Other times, these terms are used without defining their specific meaning, or with one same term used for different meanings, or many different terms all referring to the same concept (Lipton 2018).

Gilpin et al. (2018) describe the concept of explainability as a combination of *interpretability* and *fidelity*, both of which are needed to achieve explainability. Here, interpretability refers to how understandable an explanation is for a human, and fidelity describes how accurately an explanation depicts the behavior of the AI model over the entire feature space. However, this often entails a trade-off between these two qualities, whereby it is difficult to simultaneously achieve both high interpretability and high fidelity: The most comprehensive explanation may not be easily interpreted by a human, and an intuitive explanation may not be sufficiently complete in its

With High-Risk Systems, Ensure Explanations Do Not Overrepresent Their Validity

If your AI is being used in a high-risk system, e.g., in the medical domain or justice system, be careful to ensure your target audience does not place more emphasis on the explanation than is warranted by the technique. In such cases, we often observe that explainability is used with an intent to improve trust in the system with end users. However, these explanations are often treated as absolute truths by ML consumers and used to justify their conclusions, rather than a *probable* explanation for a model's behavior.[8]

For other, more primitive, explanation techniques such as PDP plots (discussed in Chapter 3), the risk of conveying inaccurate explanations through poor visualizations is much lower. Issues to be aware of include changing scales in the axes between explanations, which consumers may not notice, and plots that are so small they can be overinterpreted.

Build on the ML Consumer's Existing Understanding

The most useful explanations for consumers are those that build on their existing knowledge, either of the inputs, prediction, or ML model, to gain a more sophisticated understanding. To understand why this is, we must first understand a few aspects of human-computer interaction: mental models, situational awareness, and satisficing.

Mental models are very similar to ML models: they represent a framework that a person has learned that lets them quickly and efficiently reach conclusions and make decisions. Like ML, mental models often do not truly represent the actual system they model. As an example, most people's mental model of how to drive a car is that pressing the gas pedal makes the car go faster. In reality, the gas pedal controls the amount of air and fuel flowing into a car engine, which then causes a larger combustion, and in turn, causes the engine to exert more force through gears in a transmission. The gearing in this transmission allows the engine to exert more or less force depending on the speed the wheels are rotating at. While the true model of the car is more accurate, and explains why pressing a gas pedal at different speeds does not make the car accelerate the same amount, reasoning through this process every time would make driving much more onerous. Most of the time, it is sufficient to use a simpler mental model that pushing the gas pedal makes the car go faster.

8 An *absolute truth* explanation would be to trace an explanation throughout the entire ML's decision process/layers and annotate each weight and variable to reference its influence. It would also be absolutely incomprehensible.

For Explainable AI, the best explanations match the consumer's mental model of the ML system, or are able to help them build a sufficiently accurate mental model. It's most effective to evaluate how well the frameworks of the explanation and the mental model match each other when deciding between different families of XAI techniques. Determining user's mental models for different ML systems is done primarily through conducting user interviews and research. This is a time-consuming process that requires a trained UX researcher. Assuming your ML system is replacing an existing process, one shortcut to discovering this pairing is to ask users how they build confidence in decisions without an AI. For example, when classifying cancer in cell tissue slides, pathologists often justify their decision by referencing textbook images that represent canonical examples of cancers in cell tissue. In this case, using an example-based (or counterfactual) technique would be the best pairing for the pathologist end user, as shown by the SMILY app in Figure 7-8, developed by Google Health.

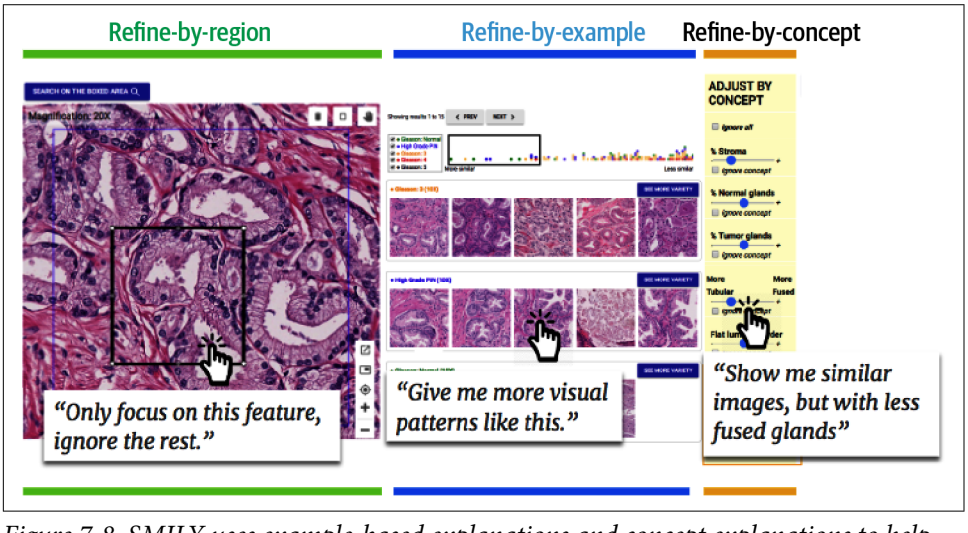

Figure 7-8. SMILY uses example-based explanations and concept explanations to help pathologists understand cell tissue slides.

Situational awareness describes how well a person understands a given scenario, and it is also a process by which people try to make decisions or arrive at conclusions when confronted with new circumstances. For ML consumers, they are often trying to improve their situational awareness when they ask the question, *Why did the ML model behave that way?*

The three steps of situational awareness are:

1. Perceive
: Determine what new information to gather and collect this information.

2. Comprehend
: Using new and existing information, build an understanding of the current situation.

3. Project
: With the current understanding, create assumptions about what will happen in the future. This could be which actions should be taken to change the situation, or how the situation itself will continue to evolve.

Situational awareness also heavily relies on users having a correct mental model; otherwise, it is very difficult to accurately comprehend and project, or even know what information to gather. Building situational awareness is a large field of research in its own right in human-computer interaction, so we will only briefly discuss how this pertains to ML consumers of explanations. Explanations are most influential when a person is perceiving or comprehending. As the ML practitioner, your choice of explanation technique also dictates the information available to the consumer when they are trying to gather new information about the rationale behind a prediction. Similarly, useful explanation techniques are those that best help an ML consumer comprehend the ML. Unfortunately, Explainable AI currently does little, beyond setting the stage, to help users as they project. In the future, we are looking forward to XAI that is closer to an interactive dialogue to help consumers explore different scenarios as part of improving their ability to project the future behavior of an ML model.

Satisficing is a common human behavior that might be best described as "good enough for the amount of time I've got right now." More formally, people are extremely good at deriving a generally optimal solution to a problem, doing it far faster than would be expected given the time it takes them to determine the truly optimal solution, but at the cost of relying on heuristics and stereotypes. Satisficing has been observed across all professions and, counterintuitively, as someone becomes more of an expert at their job and is faster at finding the best solution, they are no less likely to satisfice. For you, this is an important consideration of how you can expect ML consumers to interact with explanations—quickly and forming conclusions that are generally correct but may miss nuances. Anecdotally, we have seen this often arise across many model modalities and explanation techniques.

For example, consumers of feature attribution charts often summarize that features with minimal contributions (even if negative) for an individual prediction are not true for the entire dataset. For example, in a classification model, it may be that most predictions are not influenced by a feature, but for those along a decision boundary, the feature is highly influential. Another example of this approach is when one quickly assigns erroneous, semantic meaning to explanation heatmaps in images, when it is clear the model does not have the capacity for semantic representation; e.g., "clearly the model learned to recognize a cat because the cat ears and eyes are highlighted" rather than the more precise "this region of pixels contains many edges that are indicative of a cat." Satisficing is useful, and with it your ML consumers will probably arrive at the right conclusion most of the time more quickly. However, it also causes failures in the cases that may be of most interest to you and these consumers, where the model is not behaving as expected.

To counter satisficing, try to carefully curate the information presented in the explanation to the user. For example, many feature- and pixel-attribution techniques do not show negative attributions because it helps ML consumers be more focused on signals that drive the model toward the prediction (rather than away from it). It is also useful to design explanations to help answer a specific question for a consumer, rather than just being a general dashboard of the model's health.

With this discussion of mental models, situational awareness, and satisficing, it is also worth asking if explanations can be used to teach users how an ML model works. There is little research into this area of explainability, but we also expect that this would be a difficult use of explanations in their current form. By definition, the techniques we've presented in this book seek to explain a model *after* it has made a prediction, in a post hoc fashion. By asking if explainability can teach a user about how an ML model works, we are also asking if explanations could be used to create a surrogate, interpretable ML model. Whether this is feasible is still an active area of research.

Common Pitfalls in Using Explainability

In using explanations, we find there are a few common pitfalls for ML consumers. These situations occur not because there is something wrong with the explanation or ML model but come from consumers improperly understanding or using explanations. Most result from overreliance or overconfidence in the explanation technique but are also driven by how explainability results are packaged and delivered to consumers. The three most common pitfalls are assuming causality, overfitting "intent" to a model, and leveraging additional explanations in an attempt to augment the original explanation.