What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research ☆,☆☆

Markus Langer [a,*,1], Daniel Oster [b,1], Timo Speith [b,c,*,1], Holger Hermanns [c,d], Lena Kästner [b], Eva Schmidt [e], Andreas Sesing [f], Kevin Baum [b,c]

[a] *Department of Psychology, Saarland University, Saarbrücken, Germany*
[b] *Institute of Philosophy, Saarland University, Saarbrücken, Germany*
[c] *Department of Computer Science, Saarland University, Saarbrücken, Germany*
[d] *Institute of Intelligent Software, Guangzhou, China*
[e] *Institute of Philosophy and Political Sciences, Technical University Dortmund, Germany*
[f] *Institute of Legal Informatics, Saarland University, Saarbrücken, Germany*

A R T I C L E   I N F O

A B S T R A C T

Previous research in Explainable Artificial Intelligence (XAI) suggests that a main aim of explainability approaches is to satisfy specific interests, goals, expectations, needs, and demands regarding artificial systems (we call these "*stakeholders' desiderata*") in a variety of contexts. However, the literature on XAI is vast, spreads out across multiple largely disconnected disciplines, and it often remains unclear *how* explainability approaches are supposed to achieve the goal of satisfying stakeholders' desiderata. This paper discusses the main classes of stakeholders calling for explainability of artificial systems and reviews their desiderata. We provide a model that explicitly spells out the main concepts and relations necessary to consider and investigate when evaluating, adjusting, choosing, and developing explainability approaches that aim to satisfy stakeholders' desiderata. This model can serve researchers from the variety of different disciplines involved in XAI as a common ground. It emphasizes where there is interdisciplinary potential in the evaluation and the development of explainability approaches.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Background, motivation, and related work

Explainable Artificial Intelligence (XAI) is – once again [1–4] – a burgeoning multidisciplinary area of research. In general, XAI can be perceived as the topic or research field concerned with developing approaches to explain and make artificial systems understandable to human stakeholders [5–7].

This puts several central aspects into the focus of XAI research. First, artificial systems are the primary objects of investigation. Such systems can range from systems following a predefined set of rules, to expert and knowledge-based systems, to systems relying on machine learning. Insights from XAI research become important when these systems are too complex to allow for human oversight or are inherently opaque, which precludes human insight [8]. Second, this view on XAI emphasizes the importance of approaches that enable or provide insights into artificial systems, their functioning, and their outputs. These approaches (we call them *'explainability approaches'*) encompass methods, procedures, and strategies to provide explanatory information helping us to better understand artificial systems. Third, there is a decisive need for XAI because there are human stakeholders (e.g., users, developers, regulators)[2] whose interests, goals, expectations, needs, and demands regarding artificial systems (e.g., to have fair or trustworthy systems [9,10]) call for greater understandability of artificial systems. We call such conglomerations of stakeholders' interests, goals, expectations, needs, and demands regarding artificial systems *'stakeholders' desiderata'*.

A large part of previous XAI research was mainly concerned with developing new explainability approaches without evaluating whether these methods are useful to satisfy stakeholders' desiderata (except maybe the desiderata of developers) [9,11–13]. In fact, only a minority of papers concerned with explainability approaches also evaluated the proposed methods [12,13]. In contrast, nowadays an increasing number of researchers strongly suggest putting human stakeholders in the center of attention when evaluating and developing explainability approaches. For instance, researchers have proposed to comprehensively examine the perspectives of all stakeholders involved in the discussions around XAI (e.g., [14–16]) or they have introduced evaluation methods and metrics to systematically and empirically investigate the effects of explainability approaches on human stakeholders and their desiderata (e.g., [9,17]).

This paper reinforces and extends the focus on human stakeholders as well as on the development and evaluation of explainability approaches, and provides three main contributions. First, we propose that when evaluating, adjusting, choosing, and developing explainability approaches, research needs to pay more attention to stakeholders' specific desiderata in given contexts. This is crucial as the success of explainability approaches depends on how well they satisfy these desiderata. Current measures and metrics focus on how well explainability approaches calibrate trust or how much they increase human-machine performance (see, e.g., [17]). However, these are just two of many desiderata driving XAI research and although there is research that investigates which classes of stakeholders hold essential desiderata for XAI (e.g., [14,15,18,19]), there is a lack of research identifying, defining, and empirically investigating these desiderata, let alone research that links them to explainability approaches suitable for their satisfaction. Our paper identifies desiderata of different classes of stakeholders and calls for systematic empirical research investigating how explainability approaches, through the facilitation of understanding, lead to the satisfaction of these desiderata.

Second, we emphasize the central role of understanding as a path through which explainability approaches satisfy stakeholders' desiderata. Although research has highlighted the importance of human understanding to XAI (e.g., [17,20]), understanding sometimes seems to be considered as just one of many important outcomes of explainability approaches [21]. We claim that increasing human understanding is not just one of many important effects of explainability approaches, but crucial for the satisfaction of desiderata in general. For this reason, we introduce a model that emphasizes the critical importance of human understanding as a mediator between explainability approaches and the satisfaction of desiderata (for a related model focused on user performance, see Hoffman et al. [17]).

Third, we propose that our model can be used to guide evaluating, adjusting, choosing, and developing explainability approaches. In particular, our model highlights the main concepts and their relations of how explainability approaches are supposed to lead to the satisfaction of stakeholders' desiderata. Clearly defining, analyzing, and capturing these concepts, as well as clarifying their relations are central for the systematic evaluation of the success of explainability approaches, as this helps to identify potential reasons for why an explainability approach did not satisfy given desiderata. Similarly, considering these concepts and their relations is crucial for the choice between different explainability approaches or for the successful development of such approaches, because this supports the derivation of requirements for an explainability approach that has the potential to satisfy stakeholders' desiderata. Furthermore, our model is useful to detect where input from disciplines outside of computer science (e.g., psychology, philosophy, law, sociology; [22]) is crucial when evaluating or developing explainability approaches. Thus, our model serves to identify interdisciplinary potential and is aimed to establish a common ground for different disciplines involved in XAI. Overall, the current paper is intended for an interdisciplinary readership interested in XAI.

1.2. A conceptual model of the relation between explainability approaches and stakeholders' desiderata

For the purposes of this paper, we introduce a conceptual model (see Fig. 1) that organizes and makes explicit the central concepts of how explainability approaches relate to the satisfaction of stakeholders' desiderata, as well as the relations between these concepts. The main concepts in this model are: 'explainability approach', 'explanatory information', 'stakeholders' understanding', 'desiderata satisfaction', and '(given) context'.

The overall idea of our model is that the success of an explainability approach depends on the satisfaction of stakeholders' desiderata (consisting of the *substantial* and the *epistemic* facet of desiderata satisfaction, see Section 3.1). Desiderata

---

[2] Generally, we speak of single stakeholders here. Since we cannot consider each stakeholder individually, we treat them as representative members of specified stakeholder classes.
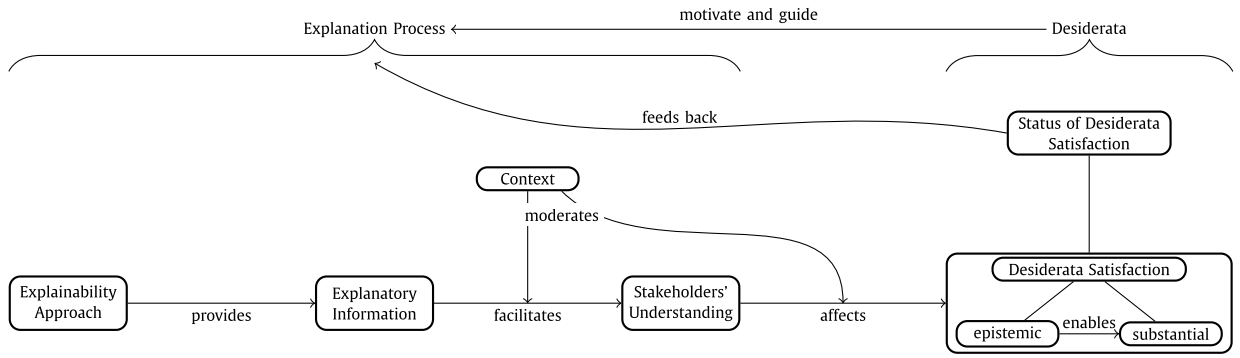
Fig. 1. Our proposed model of how explainability approaches relate to the satisfaction of stakeholders' desiderata.

satisfaction, thus, motivates an explanation process including explainability approaches, explanatory information, and stakeholders' understanding. Specifically, in the explanation process we assume that explainability approaches provide explanatory information to human stakeholders. Human stakeholders engage with the information to facilitate their understanding of an artificial system, its functioning, and outputs. As a consequence, the adjusted understanding of the stakeholders affects the extent to which their desiderata are satisfied. The context in which the human stakeholder and the artificial system operate and interact affects the relations between the other concepts (i.e., influences the relation between explanatory information and stakeholder understanding as well as the relation between understanding and desiderata satisfaction). Identifying, defining, as well as capturing and empirically examining the concepts and their relations should guide evaluating, adjusting, choosing, and developing of explainability approaches that aim to satisfy stakeholders' desiderata.

With a focus on stakeholders' desiderata, the following sections will elaborate on the model's concepts and their relations in more detail, as well as explicate shortcomings of the current view on these concepts and their relations. This paper is structured as follows. We will start on the right side of Fig. 1 and will continue to work backwards from stakeholders' desiderata. In Section 2, we will describe different classes of stakeholders and provide examples of their pertinent desiderata. We will elaborate on the central role of understanding for desiderata satisfaction in Section 3. Evoking understanding, in turn, requires explanatory information, as we will illuminate in Section 4. Section 5 will shed light on the connection between explainability approaches and explanatory information. Throughout these sections, we will point towards interdisciplinary potential that becomes apparent with the transition from one concept to the next in our model. In Section 6, we will, then, exemplify how our model can be used to evaluate, adjust, choose, and develop explainability approaches.

2. Stakeholders' desiderata

Stakeholders' desiderata are one, if not *the*, reason for the rising popularity of XAI (see also [14,19]). Since stakeholders in combination with their concrete desiderata motivate, guide, and affect the explanation process depicted in Fig. 1, we propose that identifying and clarifying desiderata of the various classes of stakeholders related to artificial systems is a crucial first step when evaluating, adjusting, choosing, and developing explainability approaches for an artificial system in a given context.

2.1. Stakeholder classes

The need for explainability starts with the increasing societal impact of artificial systems and the fact that many such systems still have to be operated by humans or affect human lives. This indicates that there are various groups of people with different interests in the explainability of artificial systems: people operate systems, try to improve them, are affected by decisions based on their output, deploy systems for everyday tasks, and set the regulatory frame for their use. These people are commonly called *stakeholders*.[3]

Previous research has discussed varying classes of stakeholders in the context of XAI. For instance, Preece et al. [18] distinguish between four main classes of stakeholders: *developers*, *theorists*, *ethicists*, and *users*. Arrieta et al. [14] categorize the main classes of stakeholders into domain experts/users, data scientists/developers/product owners, users affected by model decisions, managers/executive board members, and regulatory entities (see also [15,19,24]). We follow Arrieta et al. and distinguish five classes of stakeholders: *users*, *(system) developers*, *affected parties*, *deployers*, and *regulators* (see Fig. 2).

---

[3] According to the Merriam-Webstar dictionary, a stakeholder is, among others, someone 'who is involved in or affected by a course of action' [23]. For this reason, we use this as a general term, but refer to specific stakeholder classes where appropriate.
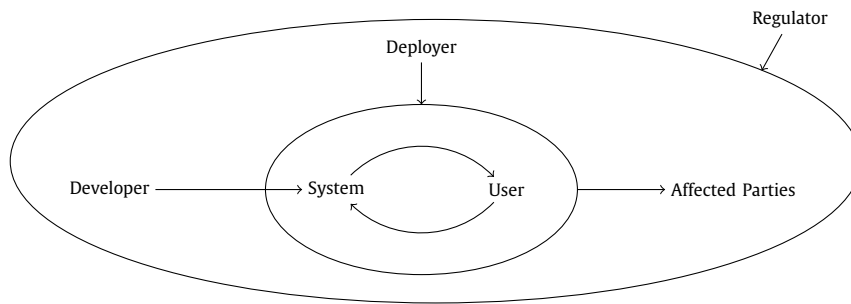
Fig. 2. The different classes of stakeholders associated with artificial systems and their relations.

Clearly, one person can be a member of several stakeholder classes. A user, for instance, can be affected by the outputs of the system she operates. Additionally, these are just prototypical classes of stakeholders and more fine-grained distinctions into sub-classes of stakeholders are possible [15]. For example, there is not one prototypical developer, but developers differ in their expertise and in other factors (e.g., personality). A novice developer may have different desiderata than an expert. In a similar way, a lay user's desiderata might differ from those of an expert user (more on this in Section 4.2). Moreover, we want to emphasize that this list of stakeholders is not necessarily exhaustive because our distinction is based on previous research that mainly comes from a computer scientific background. Thus, it might neglect other classes of stakeholders.

2.2. Exemplary stakeholders' desiderata

The desiderata arising from the five classes of stakeholders are diverse. Based on a term search, we identified more than 100 peer-reviewed journal and conference publications that postulate XAI as being indispensable when it comes to satisfying the different desiderata (see Table 1).

Table 1 presents exemplary desiderata that we have extracted from this literature review. Each row contains a desideratum, partnered with sources that claim, propose, or show that XAI-related research (e.g., on explainability approaches) and its findings and outputs can contribute to the satisfaction of this desideratum. Furthermore, this table presents stakeholder classes that may be most prone to have one or more of these desiderata. Whenever we could not extract stakeholder classes from the respective papers, we did our own (mostly common-sense) mappings. Notably, most of the sources we present in this table only claim that XAI-related research can contribute to satisfy the respective desiderata with only a subset of these papers (for instance, [28,31,45]) providing empirical evidence for their claims (e.g., regarding the mapping of desiderata to stakeholder classes or regarding the relation of explanatory information and desiderata). In what follows, we present two important exemplary desiderata for each class of stakeholders.

Users. Most papers concerning stakeholders in XAI have this class of stakeholders in common (see, e.g., [14,18,19]). Among others, users take into account recommendations of artificial systems to make decisions [24]. Some prototypical members of this stakeholder class are medical doctors, loan officers, judges, or hiring managers. Usually, users are no experts regarding the technical details and the functioning of the systems they use. However, they can work more effectively if they form adequate expectations concerning the systems' functioning. In case they cannot do so, and in cases where their expectations are violated, they need information that goes beyond the knowledge of purely operating the system. This motivates at least the following two central desiderata of users: *usability* [22,36] and *trust* [29,96,115].

In many cases, a system is more usable if it offers meaningful information alongside its outputs. This information can help users to adequately link their knowledge and assessment of a given situation to the information used by a system, can help them to make decisions more quickly, or to increase decision quality [116]. All of this can contribute to the usefulness (another important desideratum of users) of a system and is important in high-stakes scenarios where a user decides on the basis of a system's recommendations.

This is closely linked to the desideratum of adequately calibrating trust in systems. Both undertrust and overtrust can negatively affect the appropriate use of systems [117]. In the case of undertrust, users may constantly try to supervise a system's behavior or even attempt to intervene in a system's processes, thereby undermining the effectiveness of the human-system interaction [118]. In the case of overtrust, people may use a system without questioning its behavior [118–120]. This can again decrease the effectiveness of the human-system interaction, as humans rely on the system's outputs even in situations where they should challenge them [117,121]. Explainability approaches have the potential to provide means to let users adequately calibrate their trust in artificial systems [17].

Developers. Individuals who design, program, and build artificial systems are the developers. Naturally, they count as a class of stakeholders, as without them the systems would not exist in the first place. Generally, developers have a high expertise concerning the systems and an interest in improving them.

An especially important desideratum of developers is *verification*, that is, to check whether a system works as intended [7,122–124]. There are many ML-based classifiers that consider, for instance, irrelevant inputs as relevant (see, e.g., [125, 126]). Increasing insights into the system's decision-making processes by using certain explainability approaches can help

Table 1

An exemplary list of desiderata, stakeholders holding these desiderata, and sources that claim, propose, or show that XAI-related research (e.g., on explainability approaches) and its findings and outputs can contribute to the satisfaction of these desiderata.

Desideratum	Tentative description	Stakeholder	Without empirical (emp.) investigation	Emp. evidence	Emp. mixed	Emp. inconclusive
Acceptance	Improve acceptance of systems	Deployer, Regulator	[5,13,15,25–48]	[49]	[50]	
Accountability	Provide appropriate means to determine who is accountable	Regulator	[7,15,16,20,22,29,51–67]	[43,68]		
Accuracy	Assess and increase a system's predictive accuracy	Developer	[5,25,33,38,46,51,55,69–73]			
Autonomy	Enable humans to retain their autonomy when interacting with a system	User	[15,44,53,56,57,60,74–76]			
Confidence	Make humans confident when using a system	User	[5,14–16,28,32,34–36,38,39,48,49,57,58,69,73, 74,77–81]			[82]
Controllability	Retain (complete) human control concerning a system	User	[6,12,13,22,44,49,56,60,74]	[39]	[43]	
Debugability	Identify and fix errors and bugs	Developer	[5,7,12,13,16,25,30,33,44,47,48,51,56,58,61–64, 67,74,83–88]	[82,89,90]		
Education	Learn how to use a system and system's peculiarities	User	[6,13,24,31,33,36,39,49,74,80,88,91,92]			
Effectiveness	Assess and increase a system's effectiveness; work effectively with a system	Developer, User	[13,26,34,36,39,42,48,49,65,67,74,77,78,93–95]	[28,31,46]		
Efficiency	Assess and increase a system's efficiency; work efficiently with a system	Developer, User	[13,28,39,55,74,78,93–95]	[31]		
Fairness	Assess and increase a system's (actual) fairness	Affected, Regulator	[8,12–14,16,22,24,32,38,40,44,51,53,55,56,58, 60,63,64,67,69,70,78,81,86,87,96–100]		[29,43,68]	
Informed Consent	Enable humans to give their informed consent concerning a system's decisions	Affected, Regulator	[15,56,57]			
Legal Compliance	Assess and increase the legal compliance of a system	Deployer	[5,7,12,15,16,21,22,24,26,27,29,30,34,35,44,47, 55,56,58,62,63,66,68–70,74–76,81,87,91,98, 100–102]			
Morality/Ethics	Assess and increase a system's compliance with moral and ethical standards	Affected, Regulator	[6,12,14,16,29,34,37,44,47,53,55,57,58,60,69,78]			
Performance	Assess and increase the performance of a system	Developer	[15,26,32,33,36,38,40,43,48,50,51,66,72,74,91, 96,100,102]		[34,103]	[49]
Privacy	Assess and increase a system's privacy practices	User	[14,16,78,98]			
Responsibility	Provide appropriate means to let humans remain responsible or to increase perceived responsibility	Regulator	[6,13,20,43,56,57,60,104]	[68]		
Robustness	Assess and increase a system's robustness (e.g., against adversarial manipulation)	Developer	[5,14,55,105]			
Safety	Assess and increase a system's safety	Deployer, User	[44,58,69,70,74,78,105]			
Satisfaction	Have satisfying systems	User	[5,13,15,24,28,33,36,39,43,46,78,79,94,102]			[31]
Science	Gain scientific insights from the system	User	[5,7,12,14,35,37,44,47,60,63,66,67,70,87,97,100, 106]			
Security	Assess and increase a system's security	All	[55,57,78,100]			
Transferability	Make a system's learned model transferable to other contexts	Developer	[14,37,84]			
Transparency	Have transparent systems	Regulator	[13,16,25,26,29,33,34,37,39,44,46,49,55,57,62, 63,67,74,75,78,80,92,94,107,108]	[28,31,43,50]		
Trust	Calibrate appropriate trust in the system	User, Deployer	[5–7,12,13,15,16,20,22,25–29,32–37,42,44–47, 49,51,53,55–59,62,63,67,69–75,77–79,81,83–86– 88,91,93–96,98,100,103,106–113]	[30,39]	[38,40,48]	[21,31,50,82,92, 114]
Trustworthiness	Assess and increase the system's trustworthiness	Regulator	[7,12,33,38,41,60,69,74,77,91,94,104,105]			
Usability	Have usable systems	User	[13,22,28,33,36,43,57,62,63,78,104,108,110]			
Usefulness	Have useful systems	User	[13,34,36,43,63,77]	[31,45]		
Verification	Be able to evaluate whether the system does what it is supposed to do	Developer	[7,14,16,20,32,33,43,44,55,56,62,66,80]			

We classify the sources into those that provide no empirical investigation of their claims, those that show empirical evidence (e.g., an explainability approach affected a desideratum's satisfaction), those that provide mixed empirical evidence (e.g., for some explainability approaches there are positive effects on a given desideratum's satisfaction, whereas for others there are no effects), and those that present inconclusive empirical evidence (e.g., the effect of an explainability approach on a desideratum's satisfaction was not significant).

developers to recognize and correct such mistakes. Accordingly, there are cases where XAI contributes to determine whether a system works as intended and, thus, explainability approaches can support verification of the system.

Another important desideratum for developers is *performance*. There are many ways in which a system can achieve a better performance. For example, the predictive accuracy of an ML algorithm can be seen as a performance measure. Although there are some claims that explainability and accuracy are difficult to combine [40,44], there is also the opposite view, which sees XAI as a way to actually make systems more accurate and, in particular, to help developers estimate system accuracy [70,73]. By means of getting information of what led to a system's outcomes, developers can detect underrepresented or erroneous training data and, thus, fine-tune the learning process to achieve higher accuracy. Another way in which performance can be understood is user-system interaction. The better users can interact with a system, the better they, the system, and the combination of user and system perform. To this end, insights about a system, its functioning, and its outputs are a fruitful way to improve user-system performance [14,17,48].

Affected Parties. The influence of artificial systems is constantly growing and decisions about people are increasingly automated – often without their knowing. Affected parties are such (groups of) people in the scope of a system's impact. They are stakeholders, as for them much hinges on the decision of an artificial system. Patients, job or loan applicants, or defendants at court are typical examples of this class.

Crucial desiderata of affected parties are *fairness* [7,9,11,122] and *morality/ethics* [6,11,127]. These desiderata are closely related. If a system is fair, for instance, the influence of protected attributes (e.g., gender or ethnicity) is adequately limited or controlled in the systems' decision-making processes. In the case of ethical systems, their decision-making processes rely on morally permissible considerations (e.g., according to certain moral theories, an autonomous car in a dilemma situation should never let affected parties' age contribute to its decision-making process, see [128]).

Considerations of fairness and ethics have evolved because there is an increasing number of affected parties. This can lead to discrimination of individuals (e.g., concerning the distribution of jobs, loans, or healthcare), not on the basis of their own actions or characteristics but on the basis of actions or characteristics of social groups to which they belong (e.g., women, ethnic minorities, older people) [54]. One hope of establishing automated decision-processes was to make decisions less prone to human bias [129]. However, it is commonly acknowledged that artificial systems can reproduce and, in this process, even intensify human biases (see, e.g., [54] and [130]). To counteract biases, it is, therefore, crucial to enable their detection. Explainability approaches may aid in this regard by providing means to track down factors that may have contributed to unfair and unethical decision-making processes and either to eliminate such factors, to mitigate them, or at least to be aware of them.

Deployers. People who decide where to employ certain systems (e.g., a hospital manager decides to implement a diagnosis system in her hospital) are deployers. We count them as another class of stakeholders because their decisions influence many other classes of stakeholders. For example, users have to work with the deployed systems and, consequently, new people fall inside of the range of affected parties.

Deployers want the systems they bring into use to be *accepted* [5,11,96,131]. In the eyes of deployers, the worst case in terms of acceptance is that users reject appropriately working systems so that the systems will end up never being used [132]. Therefore, low acceptance undermines what deployers intend to achieve when providing systems to users. Previous research claims that explainability approaches can aid in this case by providing people with more insights into systems, which can improve their acceptance [11,115,131].

Another desideratum of deployers is the system's *legal compliance*. As deployers bear a certain degree of responsibility for systems they bring into use, they have to ensure that these systems comply with legislation. Non-discrimination and safety of a system are two important factors for its legal compliance. Explainability approaches promise to enable deployers and other stakeholders to check whether the system is indeed safe and non-discriminatory. Moreover, the European General Data Protection Regulation (GDPR) and the often discussed *Right to Explanation* [101] (arguably) explicitly require explanations.

Regulators. Finally, there are regulators stipulating legal and ethical norms for the general use, deployment, and development of systems. This class of stakeholders occupies a somewhat extraordinary role, since they have a 'watchdog' function not only with regard to systems, but to the whole interaction process of systems and the other stakeholder classes. This class consists of ethicists, lawyers, and politicians, who must have the know-how to assess, control, and regulate the whole process of using artificial systems.

Regulators call, for instance, for *trustworthy* systems [5,6,10,11,29,96,115,131]. However, the concept of trustworthiness is still only vaguely defined [133]. For example, the High Level Expert group on Artificial Intelligence (HLEGAI) initiated by the European Commission does not provide a common definition for trustworthiness, but it only proposes that trustworthy systems have three properties: they are lawful, ethical, and robust [10]. Without examining trustworthiness more closely, the HLEGAI emphasizes the significance of trustworthy artificial systems by stating that the trustworthiness of systems is imperative for the realization of potentially vast social and economic benefits. Regulators such as the EU, as well as previous research on artificial intelligence that calls for trustworthy systems (e.g., as described in [53]), agree that explainability approaches are one central way to facilitate the trustworthiness of systems [10,53].

Accountability is another important desideratum of regulators [20,56]. Accountability is about being able to identify who is blamable or culpable for a mistake. With increasing use of artificial systems, accountability gaps might emerge [134,135]. For instance, when the use of an artificial system harms a person, it may not be clear who is accountable, as there are many parties that may have contributed to the harm. Opaque artificial systems only amplify this issue. For example, a person acting on the outputs of a system may not (be able to) know that this output was erroneous, so blaming her

for ensuing problems might inadequately ignore the contribution of the artificial systems. Overall, regulators want to avoid situations in which existing legislation is hard to apply or where no one is (or feels) accountable for a mistake. In such cases, explainability approaches may restore accountability by making errors and causes for unfavorable outcomes detectable and attributable to the involved parties.

2.3. Interdisciplinary potential

Artificial systems will continue to influence humans in every part of their lives, thus it is likely that new desiderata will emerge. Further desiderata might evolve from societal, legal, political, philosophical, or psychological needs regarding artificial systems (e.g., for competence, relatedness, or autonomy; [136]). For example, with artificial systems in healthcare [137] there is a pressing need for formulating relevant ethical and legal desiderata. In addition, it is also possible that explanatory information provided by artificial systems does not only aim to improve task achievement but also to entertain users [138].

We conducted a literature review to derive an overview of stakeholders' desiderata, but it will clearly be possible to extend our list in Table 1. In fact, further developing and refining this list of desiderata is an important point that reveals interdisciplinary potential. First, most of the sources referred to in Table 1 only claim that these desiderata are relevant for stakeholders. There needs to be a more thorough empirical investigation, probably done by interdisciplinary teams of psychologists, philosophers, and scholars from law to show the actual importance of these desiderata for certain stakeholder classes.

Furthermore, in our overview, the desiderata's denotations stem (in most cases) directly from the source papers. However, some of these desiderata are closely related and, especially given the interdisciplinary research contributing to XAI, it is plausible that different authors actually mean to refer to the same desideratum but give it a different term or use the same term to refer to different desiderata. Consistent terminology and conceptual clarity for the desiderata are pivotal and there is a need to explicate the various desiderata more precisely. Different disciplines like law, philosophy, and psychology need to come together to discuss their conceptions of various desiderata to agree on common definitions of these desiderata. Without this, insights from different disciplines regarding the respective desiderata (e.g., what kind of explanatory information is required to satisfy a given desideratum) might not be adequately integrated into a common stream of research.

Additionally, we need research that more explicitly analyzes society's stakeholder classes affected by artificial systems. For instance, collaborating with sociologists could offer a broader or more nuanced picture of the classes of stakeholders that have to be considered within the scope of XAI. In any case, in order to comprehensively address the stakeholders' desiderata, we need a more detailed understanding of stakeholder classes and sub-classes. For this, it is promising to consult disciplines outside of computer science focusing on society (i.e., sociology) as well as individual differences within groups of society (i.e., psychology).

Furthermore, researchers from different disciplines may be able to take the perspective of certain stakeholder classes. By doing so, they can help to refine the list of desiderata. For instance, computer scientists can take the perspective of developers. Psychologists can take the perspective of users and affected people. Management scholars could take the perspective of deployers. Philosophers, political scientists, as well as researchers from law can take the perspective of regulators.

Working together in interdisciplinary teams can, thus, contribute to a comprehensive consideration of important desiderata in a given context. However, comprehensiveness is just one side of the coin, the justification of desiderata is another. Concerning this justification, there are two main perspectives: one from ethics and one from jurisprudence. From an ethical perspective, we can judge whether a desideratum is compatible with some, many, or even all established moral theories. Similarly, legislation can be consulted to assess whether there are laws demanding (or prohibiting) to meet certain desiderata. When engaging in thorough moral and legal justification, we might conclude that there will be desiderata that are not justifiable. In high-stakes decisions, for instance, each individual user might want systems to do what is best for her. In the case of autonomous cars, drivers will probably want a car to decide in a way that makes it more likely that they will survive if the car faces an imminent accident [139]. There might be cases where such a decision is neither morally nor legally justifiable. In a less drastic example, users may ask for an explanation of why they received a low score on a personnel selection test. However, providing this explanation might render the given test obsolete because the explanation possibly enables participants to game the test [8]. We suggest that the given context, as well as moral and legal considerations are decisive factors when determining whether certain stakeholders' desiderata can be justified.

3. Desiderata satisfaction requires understanding

In the previous section, we have introduced our claim that the need for XAI arises from stakeholders' desiderata. More precisely, the need arises in cases where certain stakeholders' desiderata are not (sufficiently) satisfied [7,14,15,18,53]. For this reason, we have to take a look at what it means for a desideratum to be satisfied.

3.1. Facets of desiderata satisfaction

We propose that the satisfaction of each desideratum can take two facets. We call these facets *epistemic* and *substantial* desiderata satisfaction, respectively. On the one hand, stakeholders want systems *to have* certain properties that make them

actually fair, transparent, or usable. In line with this, a desideratum (e.g., fairness) is substantially satisfied if a system sufficiently possesses the corresponding properties. On the other hand, stakeholders want *to know* or be able *to assess* whether a system (substantially) satisfies a certain desideratum (i.e., whether the system has the required properties). So, the epistemic facet of the fairness desideratum is satisfied for a stakeholder, if she is in a position to assess or know whether and to what extent the system is fair. Naturally, for XAI the epistemic facet is the most important one, since explanatory information can contribute to the satisfaction of the epistemic facet of every desideratum, whereas this is not the case for the substantial facet.

As an example, take the desideratum of having usable systems. A successful explanation process as depicted in our model may enable users to recognize whether a system is usable, and, optimally, also increase the system's usability to a certain degree. In this case, the epistemic satisfaction consists in the stakeholders being able to check whether a system or its outputs are usable for the task at hand. To a lesser extent, however, an explanation process can also contribute to the substantial satisfaction of the desideratum, since it provides additional knowledge about the system that makes it more usable for the stakeholder. For larger deficits in usability to be addressed, however, explanatory information might not directly help; for this, the entire artificial system may need to be redesigned.

Depending on the desideratum, the two facets are correlated to a certain degree (possibly even completely, when satisfying the epistemic facet also completely satisfies the substantial facet). To illustrate, consider the desideratum of retaining user autonomy in human-in-the-loop scenarios. Let us assume that an explanation process has helped to satisfy the epistemic facet of this desideratum to a certain degree, as it has enabled the user to assess the extent to which she can retain her autonomy in making decisions based on the recommendations of the system. Additionally, the more understanding a user has about a system's output, the more autonomous she can decide based on it. Thus, the explanation process has helped to satisfy the epistemic and the substantial facet of this desideratum. Hence, in this case, the substantial and the epistemic facet of desiderata satisfaction are highly correlated.

Now, consider the desideratum that systems adhere to certain ethical principles. When having sufficient information about a system, regulators can evaluate whether this system complies with ethical standards. Again, the explanation process serves to satisfy the epistemic facet of this desideratum. However, this does not directly make the system's processes and outputs more likely to comply with ethical standards. Consequently, explanation processes can at most indirectly satisfy the substantial facet of this desideratum: based on the understanding obtained by the explanation process, faults can be identified and steps to improve systems regarding their ethical properties can be initiated. In this case, the epistemic and the substantial facet of desiderata satisfaction are only loosely correlated.

On the one hand, the distinction of these two facets shows that explanation processes can contribute to the satisfaction of all epistemic facets of desiderata concerning artificial systems. On the other hand, it shows that an explanation process alone does sometimes not suffice to satisfy the substantial facet of desiderata concerning artificial systems. In many cases, however, the epistemic satisfaction enables the substantial one. This means that even if a better understanding of the systems triggered by explanatory information does not always directly lead to the substantial satisfaction of the desiderata, it can form the necessary basis for achieving it. As the epistemic satisfaction of a desideratum is closely linked to a better understanding of a system, understanding is the pivotal point for all endeavors of satisfying desiderata.

3.2. Understanding

Throughout the history of XAI research, authors have highlighted the central importance of understanding in XAI (e.g., [2, 20,115,140–142]). The overall goal of XAI is to advance human understanding of artificial systems in order to satisfy a given desideratum. There is an ongoing debate in the philosophical literature about what constitutes understanding [143–145], and a comprehensive review of this concept is beyond the scope of the current paper (see, for instance, [144] for a review on understanding, [146,147] for papers on the concept of understanding, [138,148,149] for the relation between explanations and understanding, or [150] for the related topic of cognitive processes in knowledge acquisition; furthermore, see [151] for a broad overview on the theoretical basics of understanding relevant for XAI research). Some aspects of understanding, however, are typically agreed upon: There are different *depths* and *breadths* of understanding (in the following, we will use the term *degree of understanding* to address depth and breadth of understanding, similar to [144,152]), and there are different *kinds* of understanding [20,146].

For the evaluation of explainability approaches it will, thus, be crucial to determine stakeholders' understanding of artificial systems. Examining understanding of software has a long history in human-computer interaction and education [108,153]. Typically, when referring to the understanding of an artificial system, these disciplines also refer to users' *mental models* of systems. A mental model of a system can be understood as a mental representation of this system and its functioning. According to Rouse et al., mental models allow humans 'to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future states' [154]. In other words, mental models allow humans to mentally simulate aspects of a system, for instance, in order to understand the causes of its decision-making [17].

Translated to the case of artificial systems, this means that we could examine understanding by investigating stakeholders' mental models of a system [17,108]: how well does a person's mental model mirror the actual system? Are there gaps in the current understanding of a system's functioning? Are people overconfident that they understand how a system works (when they actually only have an illusion of understanding) [155]? Are there learned misconceptions about systems

and their outputs that need to be revised [156]? For example, it is possible to investigate a stakeholder's mental model through think-aloud techniques where stakeholders are tasked to describe systems and their inner workings, or by letting stakeholders draw their mental model of a given system (for an overview of methods to elicit mental models, see [17]). Similarly, it is possible to measure understanding by capturing what humans' mental models enable them to do in relation to a system and its outputs. For example, Kulesza et al. [157] used an *explanation task* to assess whether participants understood what kinds of information are used to predict outcomes. Other studies used *prediction tasks* to assess, for instance, whether participants can anticipate which predictive model would produce better outcomes [115], whether participants can predict what outcome a predictive model would produce for a person with a given profile [21], or whether participants can foresee the influence of a given feature on an outcome [21]. Another possibility would be to use *manipulation tasks* in order to assess whether people understood what kind of information might add to the predictive accuracy of a model (e.g., like in [115]). Further possible tasks might be *perception tasks* (e.g., naming of recognized characteristics of a model) or *imagination tasks* (e.g., estimating what a model would predict for a given input), all of which would reflect different degrees of understanding. Furthermore, all of these tasks can reveal misconceptions of a system's functioning or knowledge gaps that need to be adjusted or filled with additional or alternative explanatory information. Additionally, what all of these ways to capture stakeholder understanding have in common is that they might help us to examine whether a given desideratum has the potential to be satisfied. For example, if a developer has understood a system in a way that she can imagine situations under which a system might fail, her ability to make the system more robust most likely increases.

Furthermore, there is an initial degree to which stakeholders understand artificial systems. Specifically, stakeholders without any prior experience with a given system will likely start with a degree of understanding that corresponds to their (background) knowledge of artificial systems or arose from initial instructions they have received regarding the system [17, 158]. Thus, they will have an incomplete or even faulty mental model of the given system [20]. For instance, a stakeholder might know (or might be informed) that machine-learning based systems are usually trained on historical data in order to predict new data. This degree of understanding can, then, be augmented (e.g., with explanatory information generated from explainability approaches). With a higher degree of understanding (and, consequently, a more detailed and accurate mental model of a system), a stakeholder might understand what kind of training data underlie a given system, what kind of algorithm is used for a given system, or what kind of output data a system produces [17,63,159]. Thus, with increasing degrees of understanding, stakeholders will be able to assess whether a given system has desired characteristics and adequate processes, or produces expected outcomes. In other words, an increasing degree of understanding will satisfy the epistemic facet of more and more desiderata. For the satisfaction of the desideratum's substantial facet, however, the opposite might sometimes be the case, as we will discuss in the next section.

3.3. The relation between understanding and desiderata satisfaction

In certain cases, a stakeholder's degree of understanding and the extent of desiderata satisfaction are positively correlated. For instance, if the desideratum is to retain autonomy in interaction with a system, usually a higher degree of understanding satisfies the epistemic and substantial facets of this desideratum to a greater extent. However, there are also more complex cases. Assume that the desideratum is to trust a certain system. Acquiring a higher degree of understanding will increase a stakeholder's epistemic satisfaction of this desideratum (i.e. she can better assess whether and to what extent to trust the system), but the substantial facet (i.e., her actual trust) can be influenced in a negative way. When a stakeholder still possesses a low degree of understanding, she is likely to be unaware of problematic features a system has in certain contexts (e.g., in complex environments) or with certain kinds of input data (e.g., noisy inputs). So, with a low degree of understanding, a stakeholder is likely to trust a system (although inadequately) [29,38]. In contrast, with a higher degree of understanding, the stakeholder is able to recognize or even explain the conditions under which a system will tend to fail. Therefore, she is more aware of the system's problematic features, and this may, consequently, decrease her trust in it [21,38].

Additionally, it can happen that understanding contributes to the satisfaction of a single desideratum of a stakeholder to a greater extent, while the satisfaction of other desiderata for the same stakeholder suffers [63]. To illustrate, take the trade-off between transparency and non-gameability of systems. Deployers of systems want to comply with legislation and, consequently, want their systems to be transparent. Understanding is a necessary condition for perceived transparency. However, making systems more transparent can diminish another desideratum of deployers, the systems' non-gameability [8]. In other words, it should not be possible for particular users to manipulate a system in such a way that they can systematically evoke beneficial outputs for them. However, more transparency caused by a higher degree of understanding may enable some people to exploit the system [63]. In a personnel selection test, for example, a better understanding of the selection system may enable participants to game the test, preventing its proper use (i.e., selecting suitable applicants).

This points to another potential trade-off to be considered regarding the relation between an advanced understanding and desiderata satisfaction. Felzmann and colleagues [15] argue that different stakeholders might hold different expectations regarding the extent to which a single desideratum has to be satisfied. It is furthermore possible that, while desiderata of one stakeholder are influenced positively by an increase in understanding, desiderata of other stakeholders suffer. An example of such a case was described by Langer et al. [160]. They provided additional information accompanying an automated

personnel selection system for their participants, which resulted in more perceived transparency, but at the same time reduced acceptance of the system (for a similar finding see [161]). In such cases, it can happen that the two desiderata of transparency and acceptance arise from different perspectives and characteristics of stakeholders. For instance, legislation (i.e., a regulator) might call for transparency of systems, whereas a company using a system (i.e., a deployer) desires the system to be accepted. Explaining the system will, then, lead to the satisfaction of the legal desideratum, but at the price of impairing the company's desideratum.[4]

These examples indicate that the degree and kind of understanding of artificial systems which explainability approaches should evoke may depend on trade-offs between a variety of desiderata from a variety of stakeholders. Consequently, the development, implementation, and use of explainability approaches should go hand in hand with a case-by-case evaluation of the relevant stakeholders' desiderata. While estimating the effects of explainability approaches, it is central to investigate the perspective of not only one, but all stakeholders who potentially have a variety of (conflicting) desiderata with regard to a given system.

3.4. Factors influencing the relation between understanding and desiderata satisfaction

Characteristics of the context in which artificial systems operate moderate the relation between understanding and desiderata satisfaction and may affect the degree and kind of understanding necessary to satisfy a given desideratum. There is no agreed-upon definition of the term 'context' [162] and a deeper dive into the discussion about the term goes beyond what we can achieve in this paper (for discussions on this topic, see [162,163]). Following Dourish [163], we hold that the context is set by a given situation, in the interaction between a stakeholder, an artificial system, a given activity or task, and an environment. This makes it impossible to anticipate all contextual influences that will affect the process of how explainability approaches aim to satisfy stakeholders' desiderata without knowing the concrete situation. Nevertheless, it is crucial to consider and anticipate these contextual influences when evaluating and developing explainability approaches.

For instance, what is at stake in a given situation can affect the relation between understanding and stakeholders' desiderata. That is, whether a context is a high or low stakes scenario may determine the degree of understanding necessary to satisfy a given desideratum. Research indicates that certain situations tend to require a greater understanding of an event than other situations. Specifically, situations where instrumental, relational, moral, or legal values are at stake might be more likely to require extensive understanding [164–166]. Instrumental values are at stake when there are (personal, economical etc.) benefits or losses to expect in a specific situation (e.g., when an artificial system handles financial transactions). Relational values are at stake when important interpersonal relationships might be affected by an event (e.g., when artificial systems are used for employee layoff). Moral values are at stake when moral rights might be violated (e.g., when using artificial systems for sentencing in court). Finally, legal values are at stake when legal rights might be violated (e.g., when an artificial system's outputs conflict with the right of non-discrimination). Depending on the concrete situation (e.g., being in an autonomous car), instrumental, relational, moral, and legal values determine the stakes of a situation. Additionally, identifying which of these values stakeholders regard as relevant or which values are indeed relevant in which situations may allow for drawing inferences about whether a situation is (considered) high or low stakes. Consequently, these values serve as an orientation for when stakeholders are more likely to demand higher degrees of understanding. Specifically, this means that for the satisfaction of a given desideratum in a low stakes scenario, a lower degree of understanding might be sufficient compared to high stakes scenarios.

Furthermore, contexts involving artificial systems may differ in their 'outcome favorability' [167,168]. In the case of an unfavorable outcome, people are more likely to call for additional information in order to understand the reasons for the outcome, and so to be able to better assess and control further similar outcomes [36,167]. On the one hand, it may be that increasing the extent of understanding in the case of favorable outcomes has little potential to positively affect a desideratum's satisfaction. On the other hand, a better understanding can be central to positively affecting a desideratum's satisfaction in the case of a unfavorable outcome. For instance, perceived fairness is a central desideratum for many decision situations [169]. While understanding may have negligible effects on perceived fairness under favorable outcomes, it can improve perceived fairness under unfavorable outcomes [170]. Supporting this claim, Kizilcec [38] found that available explanatory information only affected perceived fairness when people's expectations concerning an outcome were violated. There may, however, be certain conditions under which a better understanding (negatively) affects perceived fairness even when people experience favorable outcomes [166].

The stakes of a situation and outcome favorability are just two of many possible contextual factors that may influence the relation between understanding and desiderata satisfaction. Further candidates are the application context (e.g., at home, at work), time constraints [171], and social circumstances (e.g., whether there are other people present in a given situation; [163]).

---

[4] Note that this example also describes a situation where two desiderata of the company are in conflict: user acceptance and adhering to legislation.

3.5. Interdisciplinary potential

The interdisciplinary potential we see in the relation between understanding and desiderata satisfaction is described in research questions such as: (a) How does stakeholders' initial degree of understanding or prior knowledge of artificial systems relate to their desiderata satisfaction? (b) What are the trade-offs between desiderata of a single stakeholder and/or desiderata of multiple stakeholders and what are the implications of understanding regarding these trade-offs? (c) How does the degree and kind of understanding relate to the satisfaction of desiderata? (d) How do task or context affect the relation between understanding and desiderata satisfaction in a given situation?

Scholars from educational sciences could collaborate with computer scientists in order to investigate how to design adequate instructions to achieve a proper basic understanding or background knowledge of artificial systems that can serve as a general basis to partially satisfy a large variety of desiderata. Furthermore, it is necessary to involve psychologists in order to experimentally examine the relation between understanding and desiderata satisfaction as well as influences affecting this relation. In this regard, it will be central to determine what it means for a certain desideratum to be satisfied. This requires finding an adequate way of measuring the satisfaction of this desideratum (e.g., using self-report measures, expert interviews, legal analyses), as well as understanding the requirements for it to be satisfied (e.g., defining minimum legal standards, enabling stakeholders to perform a specific task successfully). In practice, this involves having an elaborated research design, clear conceptual definitions, appropriate operationalization and measurement methods, and fitting research disciplines for iterative (empirical) research.

Finally, scholars from requirement engineering may help to understand relationships between several desiderata. Based on their results, scholars from law and from philosophy can help to determine which trade-offs are morally and/or legally justified. In general, an interdisciplinary collaboration can contribute to being aware of potential relationships and trade-offs between certain desiderata.

4. Understanding requires explanatory information

Providing explanatory information of a given phenomenon is the default procedure to facilitate its understanding [147,172,173]. Explanatory information helps humans to navigate complex environments by facilitating better understanding, predictions, and control of situations [138,174]. Such information narrows down possible reasons for events, decreases uncertainty, corrects misconceptions, facilitates generalization and reasoning, and enables a person to draw causal conclusions [138,155,174,175].

In the context of XAI, explanatory information puts stakeholders in a position to grasp generalizable patterns underlying the production of a system's outcomes (e.g., LIME [115]). These patterns allow for drawing inferences about (potentially causal) connections between the system's inputs and outputs, or for narrowing down the possible ways in which the system might have failed. Additionally, these patterns may help stakeholders to tell apart correct but unexpected behavior from malfunctions in order to debug the system. In general, all of this can reduce people's uncertainty concerning a system (e.g., uncertainty concerning how to behave towards it, how to react to it, what to think of it, whether to recommend it, or whether to disseminate it; [63]). Overall, explanatory information should lead to a better understanding, which should, in turn, positively affect the satisfaction of stakeholders' desiderata. Depending on the explanatory information, different degrees and kinds of understanding can be acquired [20,152]. For this reason, we have to shed more light on what characteristics of explanatory information are.

4.1. Characteristics of explanatory information

Important characteristics of explanatory information concern its kind and its presentation format. There are various kinds of explanatory information: teleological (i.e., information that appeals to the function of the explanandum; [176]), nomological (i.e., information that refers to laws of nature; [177]), statistical relevance (i.e., information that is statistically relevant to the explanandum; [178,179]), contrastive (i.e., information that highlights why event P happened and not event Q; [6]), counterfactual (i.e., information that appeals to hypothetical cases in which things went differently; [180]), mechanistic (i.e., information that appeals to the mechanisms underlying a certain process; [181]), causal (i.e., information that appeals to the causes of an event; [182,183]), network (i.e., information that appeals to the topological properties of a network model describing a system; [184]), and many more.

Concerning the presentation format, there are also various possibilities. Roughly, we can distinguish between text-based and multimedia presentation [36]. A text-based presentation can be a natural language text, a rule extracted from a rule-based system, an execution trace, or simply the program's source code. A multimedia presentation can include graphics, visualizations, images, and animations. For instance, heat-maps of neural activity are a popular presentation format for explanatory information of neural networks [14].

Aside from the very general characteristics of kinds and presentation format, there are further characteristics of explanatory information that influence how and whether it evokes understanding: soundness (i.e., how accurate the information is), completeness, novelty (i.e., whether the information is new for the recipient), and complexity (e.g., depending on the number of features and on the interrelation between the features the information contains; [171]) are just some of the various examples of further characteristics of explanatory information (for more, see [63]).

The importance of acknowledging the characteristics of explanatory information is highlighted by research in cognitive and educational psychology that shows that effects of explanatory information can vary depending on their characteristics [138,150,166]. Take complexity as an example: studies show that people prefer simpler information (e.g., information mentioning fewer causes; [185]). Another example is the finding that explanatory information that aligns with a stakeholder's goals in a certain situation is preferred [186]. Vasilyeva et al. [186] showed that people evaluate teleological explanatory information compared to mechanistic explanatory information as more useful when asked to name the function of a phenomenon, and, conversely, perceive mechanistic explanatory information as more useful when asked to name the cause of an event.

4.2. Factors influencing the relation between explanatory information and understanding

The relation between explanatory information and understanding is influenced by a variety of characteristics of the stakeholders, the context, and interactions between these characteristics [164,166,187]. Considering these influences is central to evaluate and develop explainability approaches.

Characteristics of stakeholders. Since every stakeholder possesses an individual degree of understanding of a system, an individual ability to understand, and an individual set of desiderata that are or need to be satisfied to a certain extent, the characteristics of stakeholders who receive explanatory information influence the relation between explanatory information and understanding [18]. Some of the characteristics that most obviously influence this relation are the stakeholders' background knowledge, beliefs, learning capacities, and desiderata they have concerning respective systems [63,188–190]. For instance, explanatory information including technical details might increase an expert developer's degree of understanding while technical details can hamper understanding for novice developers or other (non-expert) stakeholders [175].

Furthermore, desiderata that are salient for a respective stakeholder can influence the relation between explanatory information and understanding. Specifically, as stakeholders engage with information in order to advance their understanding of artificial systems, their motivation and prior beliefs may affect how they interpret a given set of information [67]. For instance, if a stakeholder's primary desideratum is to ensure that a system provides fair outputs, they will scrutinize explanatory information for signs of bias that might lead to unfair outcomes. In contrast, if a stakeholder's primary desideratum is to improve a system's predictive accuracy, they might pay special attention to information providing insights on how to improve system performance. This indicates that, depending on the given desideratum, the same amount and kind of information can lead to different degrees and kinds of understanding. Consequently, it is important to provide the appropriate information for the given purpose. These assumptions are supported by research proposing that human reasoning processes can be strongly influenced by explanatory information. Such information has the potential to improve human decision-making but may also hamper it (e.g., when explanatory information attenuates human biases or when humans need to invest too much effort to use the information [36,67]).

The list of individual differences influencing the relation between explanatory information and understanding goes beyond the scope of this article and covers personality traits such as conscientiousness [191], need for cognition[5] [192,193], and need for closure [194,195], as well as people's preferences for detailed versus coarse explanations [196], or age differences between stakeholders [197].

Characteristics of the context. Time pressure can be a relevant contextual influence [63,171]. For a stakeholder under high time pressure, the same explanatory information may lead to less understanding as compared to situations where she is under low time pressure [198]. Workload is another contextual influence [199]. In situations of high perceived workload, the same explanatory information can lead to a different degree of understanding compared to low workload conditions. Similar things are true for situations where it is more likely that stakeholders will experience higher levels of stress [200,201] (e.g., high stake and high risk situations, multitasking environments; [171,202]).

In general, depending on the situation and task at hand, the effects of explanatory information on understanding may differ. Therefore, it is important to investigate the contextual conditions of a stakeholder's interaction with an artificial system in detail when theorizing about how explanatory information can best improve understanding. Given the impact of the context and given the fact that it might not be possible to anticipate all relevant contextual influences [163], it is especially important to assess the effects of explanatory information in laboratory and in field settings [9,63]. Although it has advantages to investigate the effects of, for instance, stress on the relation between explanatory information and understanding in controlled laboratory settings, such results may not translate to field settings. This finding is in line with calls for experiments involving human participants in proxy tasks in order to show that given explanatory information not only elicits understanding when simulating a context, but also under real world conditions (that is, to investigate whether given explanatory information is equally valuable in the wild) [9,63]. Although it will be impossible to fully anticipate how the context will alter the relation between explanatory information and understanding (e.g., because the interpretation of

---

[5] Need for cognition is a personality trait that distinguishes people who like to put effort into cognitive activities from those who prefer less cognitively demanding activities.

contextual influences depends on the relevant stakeholders [162]), at least it is crucial to be aware that contextual influences may be central for the success or failure of explanatory information.

Interactions between characteristics of explanatory information, stakeholders, and the context. Finally, characteristics of explanatory information, stakeholders, and the context can interact in ways that affect how stakeholders engage with explanatory information to advance their understanding of artificial systems. For instance, the level of detail of explanatory information can interact with the prior knowledge of a stakeholder. Whereas expert developers' understanding may benefit to a higher degree from explanatory information with much detail, this level of detail can have the contrary effect for novice developers [36,166,203,204]. At the same time, when novice users want to learn how to use a system, they might want detailed explanatory information whereas when expert users want to use the system for task fulfillment, every unnecessary piece of information could lead to the rejection of the system [36]. Further, if the context changes, the relations between explanatory information and understanding are prone to change as well. For instance, as soon as there is time pressure in the aforementioned situations, it is plausible that neither expert nor novice developers or users pragmatically benefit from too detailed explanatory information.

4.3. Interdisciplinary potential

The following research questions reinforce calls for extensive validation and experimental studies investigating the effects of explainability approaches in different contexts and in relation to different stakeholders [9,63] (we refer readers to Sokol et al. [63] for a description of further important characteristics of explanations, stakeholders, and contexts that affect the relations between explanations and stakeholders' understanding): (a) How should we classify explanatory information? (b) How does different explanatory information lead to understanding? (c) How do different stakeholders engage with explanatory information? (d) How can we optimally evoke understanding through explanatory information? (e) How do stakeholder differences (e.g., background knowledge, personality characteristics) affect understanding? (f) How do contextual influences (e.g., different levels of risk, multi-tasking environments, time pressure) affect the relation between explanatory information and understanding? (g) How should explanatory information be designed? What should it include? What kind of presentation format (e.g., textual, graphical) is appropriate? How can stakeholders interactively engage with this information?

On the one hand, these research questions call for a unified classification of explanatory information. Often, computer scientists classify explanatory information based on its presentation format or based on the explainability approach it originated from (e.g., [205]). Philosophers and psychologists, however, usually classify explanatory information in terms of the kinds mentioned above (e.g., causal or nomological; [206]). In order to prevent the debate from drifting apart, philosophers, psychologists, and computer scientists should collaborate to find standardized ways to classify explanatory information. On the other hand, these research questions call for empirical evaluation of the value of different kinds of explanatory information for different stakeholders, under a variety of contexts and under the consideration of different desiderata. This may primarily call for empirically working psychologists, as they have the tools and expertise to design experimental studies, for philosophers to determine the qualities of good explanatory information, and for computer scientists to adjust the presentation of explanatory information as well as possible. Furthermore, cognitive scientists might need to examine what exactly understanding of artificial systems actually means and how to measure it (see [17] for ideas of how to capture human understanding of artificial systems).

However, deriving explanatory information of artificial systems is a task that in itself requires a lot of interdisciplinary research. The following section completes the specification of the concepts and relations in our model by describing that, in order to get the required explanatory information from the systems, there is the need for fitting explainability approaches.

5. Explanatory information requires explainability approaches

In order to provide explanatory information that facilitates understanding and, thus, affects the satisfaction of the desiderata of the different classes of stakeholders, XAI research has developed a wide variety of explainability approaches. These approaches encompass methods, procedures, and strategies that provide explanatory information to help stakeholders better understand artificial systems, their inner workings, or their outputs. A specific explainability approach is characterized by all steps and efforts that are undertaken to extract explanatory information from a system and to adequately provide it to the stakeholders in a given context. Explainability approaches can take many guises and the literature commonly distinguishes two families of approaches (e.g., [12,14,63,131,205,207]): *ante-hoc* and *post-hoc* approaches.

5.1. Families of explainability approaches

Ante-hoc approaches aim at designing systems that are inherently transparent and explainable. They rely on systems being constructed on the basis of models that do not require additional procedures to extract meaningful information about their inner workings or their outputs. For example, decision-trees, rule-based models, and linear approximations are commonly seen as inherently explainable (given they have a limited size) [7,131]. A human can, in principle, directly extract information from these models in order to enhance her understanding of how the system works or of how the system arrived at a particular output. Unfortunately, this way of deriving explanatory information from transparent models

might only be useful for stakeholders with a certain expertise [16,208]. For this reason, ante-hoc approaches make systems directly explainable only for developers and members of other stakeholder classes that possess enough expertise about artificial systems. Furthermore, ante-hoc explainability can also lead to a loss of predictive power [14,15] and not all systems can be designed inherently explainable.

Post-hoc approaches try to circumvent the aforementioned shortcomings. Such approaches are, in principle, applicable to all kinds of models. The difference to ante-hoc approaches is that post-hoc approaches do not aim at the design-process of a particular system, but at procedures and methods that allow for extracting explanatory information from a system's underlying model, which is usually not inherently transparent or explainable in the first place [11,63,131,207]. Post-hoc approaches are, for example, based on input-output analyses or the approximation of opaque models by models that are inherently explainable.

In many cases, however, post-hoc approaches are restricted with respect to how they present explanatory information. That is, given a specific model or one of its outputs, the information an approach will provide on repeated usage (and the format in which the approach provides the information) will be similar. Hence, for post-hoc approaches the same holds as for ante-hoc approaches: it is not guaranteed that all stakeholders are able to understand the provided information in the given format [16,208]. So, the explanatory information accessible from both, post-hoc and ante-hoc approaches is often only interpretable for developers or other expert stakeholders. This means that this information does not directly facilitate understanding for non-experts [16].

One solution to this is to combine several explainability approaches in order to cover a broad range of different information and presentation modes. Another solution that has received increasing attention is to have recourse on interactive explainability approaches [12,67]. Interactive approaches are based on the idea that the user or some other stakeholder is provided with more in-depth information concerning a system if the information she initially received does not suffice. Based on her needs, the person interacting with the system can call for information about specific aspects of a decision or request a presentation in a different format. To date, however, approaches that are fully interactive remain rare [12,205].

A third solution is to have a human facilitator (e.g., an expert stakeholder) explain a system to other stakeholders. For instance, when regulators want to satisfy their desiderata concerning artificial systems, there will be cases where they do not directly interact with an artificial system. Instead, a human facilitator (e.g., an expert user or a developer) will do so and derive suitable explanatory information for the regulator. In a sense, this process introduces a desiderata hierarchy based on the stakeholders into Fig. 2. In other words, one desideratum might have to be satisfied for one stakeholder class before some other desideratum (for another stakeholder class) can be satisfied. For example, the facilitator's desideratum (in this case: to be able to explain a system to a regulator), must be satisfied first as a precondition for the satisfaction of a regulator's desideratum (e.g., to be able to assess the fairness of a system). Thus, we have the complete processes in Fig. 1 nested within the explainability approach. Having a human facilitator (e.g., a developer) deriving explanatory information (assisted by an explainability approach) for another stakeholder (e.g., a regulator) can be considered a hybrid human-system approach to explainability.

5.2. Factors influencing the relation between explanatory information and explainability approaches

There are further characteristics of explainability approaches that are worth mentioning, since they are likely to influence the provided kind of explanatory information or its presentation format. First, it is important to distinguish post-hoc approaches that work regardless of the underlying model type (so-called *model-agnostic* approaches) from ones that only work for specific (types of) models (so-called *model-specific* approaches). Model-agnostic approaches aim to deliver explanatory information about a system solely by observing input/output pairs [63,115,131,207]. Model-specific approaches do so while also factoring in specific features of the model at hand (e.g., by creating prototype vectors in a support vector machine) [63,131,207]. Model-agnostic approaches have the advantage of working for all types of models, but have the drawback that they tend to be less efficient, less accurate as well as less explanatory powerful (i.e., the explanatory information's level of detail is lower with regard to individual phenomena) than the former.

Second, previous research distinguished the scope of an explainability approach. Some approaches provide information about only single predictions of the model. The scope of these approaches is *local* [63,115,131,207]. Often, they offer visualized prototype outcome examples (e.g., [209,210]). The more general type of approaches has a *global* scope [63,131,207]. These approaches are designed to uncover the overall decision processes in the model. Here, the usual way to provide this information is by approximating complex models with simpler ones that are inherently explainable.

Depending on the explainability approach, the explanatory information provided will differ. Global explainability approaches, for instance, are likely to produce more complex information that requires more background knowledge by stakeholders to be understood. Local explainability approaches, on the other hand, only show a limited picture of a system's inner working and may not be representative of its overall decision-making processes. Based on these differences we can conclude that certain explainabilty approaches are more suitable for the satisfaction of given desiderata of specific stakeholder classes than others.

To elaborate, take users who want to calibrate their trust in a system. They will need a different kind of information compared to users who want to have usable systems. In the first case, the explainability approach will likely need to extract information about the robustness of a system, about conditions under which outputs of the system are trustworthy and situations where users have to be aware that outputs might be misleading. In the case of usable systems, users might

want information that directly matches their specific goals in a given task. Another example: if a desideratum of users is learning how to use a system, they may need a different kind of explanatory information compared to when they simply want to fulfill tasks with the help of an artificial system [211]. This is because users who want to learn how to use a system need more details whereas users who want to fulfill tasks need directly useful information in order not to reduce their productivity through overly detailed information. Differences in the information needed can also be due to differences in the perspective of stakeholders. Take the desideratum of fairness. Affected parties will more likely focus on aspects of individual fairness which may call for explainability approaches that facilitate local explainability. Regulators, however, may focus on more general notions of fairness calling for explainability approaches facilitating global explainability.

5.3. Interdisciplinary potential

The design of explainability approaches and the goal of providing adequate explanatory information with the potential to affect stakeholders desiderata, again, hold untapped interdisciplinary potential, with research questions such as: (a) How can we pinpoint what explanatory information an explainability approach should provide in which case? (b) How can we design interactive explainability approaches? (c) How can explainability approaches involving human facilitators be optimally designed? (d) How can we guarantee that an explainability approach provides the required explanatory information? (e) How should stakeholders' desiderata and their degrees of understanding be taken into account when generating explanatory information or when developing new explainability approaches?

For all of these research questions, we see a potential for collaboration between computer scientists, philosophers, psychologists, and cognitive scientists. For instance, philosophers and psychologist have to determine which information is needed to assess whether a system is fair, whereas computer scientists develop explainability approaches that provide this information, are aware of trade-offs between different approaches as well as of technical constraints. Furthermore, investigating how the examination of stakeholders and their desiderata narrows down the options of possibly successful explainability approaches might be a fruitful area for future research.

A more thorough discussion about which desiderata of what stakeholder class call for what kind of explainability approach goes far beyond what a single paper can achieve. However, in Section 6, we will show that our model could inspire work that is necessary to evaluate the usefulness of an explainability approach in relation to a given desideratum. Furthermore, we will show that the model supports the development of new explainability approaches to satisfy a given set of desiderata. Thus, the next section aims to show how our model can lead to actionable insights for XAI research by analyzing the concepts and relations that we propose in Fig. 1. By means of hypothetical application scenarios, we want to stimulate ideas about specific applications of our model.

6. Bringing it all together: hypothetical application scenarios

Following, we will present how the previous sections come together within hypothetical application scenarios. These scenarios highlight the importance of different stakeholder classes and their desiderata. Furthermore, these scenarios emphasize that understanding affects the satisfaction of these desiderata, that explanatory information provided by explainability approaches facilitate understanding, and that analyzing and investigating these concepts and their relations is central for the aims of XAI as well as for the development of explainability approaches that can successfully satisfy stakeholder desiderata.

The main application of our model is derived from the idea that if an explanation process does not change a certain desideratum's extent of satisfaction, the corresponding explainability approach might not be a suitable means for satisfying the desideratum in the given context. Such a discovery (e.g., resulting from stakeholder feedback or from empirical investigation) can provide feedback regarding which explainability approaches and what kinds of explanatory information work for which desiderata in which contexts. Thus, this feedback can serve as an input for the improvement of explanation processes and, consequently, helps to evaluate, adjust, choose, and develop explainability approaches for a given purpose and context.

We propose that each step in our model (Fig. 1: Explainability Approaches → Explanatory Information → Understanding → Desiderata Satisfaction) allows for drawing inferences about the explanation process involving explainability approach(es), kind(s) of explanatory information, and stakeholder understanding. The following questions, which arise at different points in our model, are of particular interest:

- Who are the relevant stakeholders and what are their specific characteristics? Which are the relevant desiderata in a specific context and are they satisfied?
- Have the stakeholders acquired a sufficient degree and the right kind of understanding that allows for assessing whether given desiderata are satisfied and to facilitate their satisfaction?
- Does the provided explanatory information and its format of presentation facilitate stakeholders' understanding in a given context and in consideration of the stakeholder characteristics?
- Is the explainability approach able to provide the right kind and amount of explanatory information in the right presentation format?
- Are there contextual influences hindering or promoting the satisfaction of desiderata through the explanation process?

Investigating these questions requires empirical research, hypothesis testing and interdisciplinary cooperation, but should, eventually, aid in evaluating, adjusting, choosing, and developing explainability approach(es) and finding explanatory information in order to adequately satisfy stakeholders' desiderata.

With this in mind, we believe that our model is useful for several important application scenarios. First, our model is useful for choosing adequate explainability approaches for novel application contexts of artificial systems and for guiding the development of new explainability approaches. Specifically, our model can be used to inform projects on how to develop explainability approaches to satisfy certain desiderata for a given class of stakeholders. Second, our model is useful for evaluating why and at which stage an explainability process failed to contribute to satisfying the relevant desiderata. Let us assume that the use of an explainability approach does not lead to the satisfaction of a certain desideratum. Why is this the case? Is the explainability approach not suitable for the satisfaction of the desideratum (e.g., the explainability approach provides the false kind of explanatory information) and should be replaced by another approach? Or does the error lie somewhere else in the explanation process? In some cases, we may be able to adjust the explainability approach appropriately to achieve its intended purpose.

6.1. General application scenarios

For structured attempts to evaluate, adjust, choose, or develop explainability approaches for a given context we roughly distinguish two scenarios, which we call the *evaluation* and the *discovery* scenario. In the evaluation scenario, we want to investigate whether the use of a specific explainability approach was adequate, and if not, what is needed to fix its shortcomings. In the discovery scenario, we want to find an adequate explainability approach to satisfy stakeholders' desiderata. This can take one of two forms: choosing among existing approaches or, if no adequate approach is available, developing a new one.

Stakeholder and Desiderata. Both evaluation and discovery scenarios start with examining the stakeholders and clarifying their desiderata in the given context. In the discovery scenario, we have to examine what the relevant classes of stakeholders are and which desiderata they have concerning the application of a system under consideration. In the evaluation scenario, we have to check – even in scenarios where already identified desiderata are satisfied – whether the explanation process fits all relevant classes of stakeholders and all of their desiderata (and ensure that we did not overlook important stakeholders or desiderata).

For this, we need input from a wide variety of disciplines including but not limited to scholars from law, sociology, psychology, philosophy, and computer science. Such a combination of expertise and perspectives helps to identify relevant stakeholder classes and list their desiderata pertaining to a given context. Defining these desiderata falls within the expertise of philosophers. The elaboration of the desiderata's relevant moral and legal aspects is a task for ethicists and scholars from law. Furthermore, assessing contextual peculiarities and stakeholder characteristics relevant in a given context will require psychologists (on the individual level), sociologist (on the group level), as well as domain experts such as judges or personnel managers (for particular application scenarios).

Desiderata Satisfaction. For the discovery scenario, the next step is to determine what it means that stakeholders' desiderata are satisfied in a specific context: to make estimates, to provide guidelines, and to create measures for when the identified desiderata are satisfied. For the evaluation scenario, we have to check whether the relevant desiderata are satisfied or not. If they are, the explanation process was successful and the chosen explainability approach appropriate. If they are not substantially satisfied, there are two possible cases. First, the necessary understanding was not acquired and, thus, the desideratum is also not epistemically satisfied. Then, an improvement of the stakeholders' understanding is required. Second, an adequate degree and kind of understanding is reached and, thus, the epistemic facet of the desideratum is satisfied. In this case, we may conclude that the regarded desideratum is not directly substantially satisfiable by means of explainability approaches. Regardless of the case, at this point we have to move on to investigate stakeholders' understanding.

Determining conditions for desiderata satisfaction will be an interdisciplinary task for psychologists, philosophers, and scholars from law. Furthermore, computer scientists and domain experts can give practical input on satisfaction conditions for the desiderata in their specific domain. Making the satisfaction of these desiderata measurable and examining their extent of satisfaction will be a job for psychologists who develop measures or tasks that help to assess the extent of desiderata satisfaction. Additionally, it can be a task for scholars from law or philosophers to provide clear guidelines for when a desideratum is satisfied.

Stakeholders' Understanding. The discovery scenario continues by investigating and defining requirements for the stakeholders' understanding needed to satisfy the desiderata under consideration. Specifically, we have to determine the appropriate degree and kind of understanding concerning the system and its output that promise to enable the epistemic facet of desiderata satisfaction. At the same time, this means that we need to assess (e.g., in studies using tasks to measure stakeholders' mental models of a system, or in studies using tasks to reveal whether stakeholders were able to, for instance, explain a systems' functioning or predict a systems' behavior and outcomes) stakeholders' actual degree and kind of understanding. This is especially important in the evaluation scenario when deficits in the substantial desiderata satisfaction become apparent. This results from the circumstance that assessing whether the required degree and kind of understanding that has been achieved allows for drawing inferences as to whether there is a fundamental gap between the explanatory process and the substantial desiderata satisfaction, or whether the provided information is not appropriate to evoke under-

standing. This is due to the fact that the explanation process can only serve to maximize understanding. If a desideratum still remains (substantially) unsatisfied, its full satisfaction goes beyond the scope of any explainability approach.

Philosophers can help to investigate and explicate what it means to have a certain degree or kind of understanding. Building on this, psychologists can design ways to empirically assess and measure such understanding (e.g., is it necessary for a given desideratum that stakeholders are able to predict a system's outputs? Is it necessary that stakeholders are able to anticipate situations when systems will likely fail?). Furthermore, scholars from law can contribute conditions for traceability and auditability.

Explanatory Information. The next step in the discovery scenario is to pin down what explanatory information has the potential to facilitate the right kind and degree of understanding in a predetermined context. This implies an evaluation of different dimensions of explanatory information with respect to the expected effects within an explanation process. To do so, we have to pay attention to, for instance, the kind of information, its presentation format [36], its quality [17], its amount [171], its completeness, its complexity, or its adequacy for the given context [171]. In the evaluation scenario, we have to check whether an explainability approach provides explanatory information that sufficiently meets the previously identified requirements. Sometimes there will also be a need to re-evaluate whether the requirements concerning the information are indeed adequate.

In this respect, philosophers and other explanation scientists can help to distinguish between different kinds and features of explanatory information [6,7]. Furthermore, scholars from law can examine current legislation to find out whether it prescribes certain kinds of explanatory information. In the case of the GDPR, for instance, they have to specify what it means to 'provide [...] meaningful information about the logic involved' (GDPR Art. 13 (2)(f); [101]). Finally (and based on this differentiation), educational or cognitive psychologists have the task to characterize the explanatory information that is best suited to facilitate the required kind and degree of understanding for a certain context.

Explainability Approach. Insights from the assessment of the concepts and relations in our model can guide and inform the requirements for explainability approaches that aim to satisfy given desiderata. This is of particular importance for the discovery scenario, where the primary objective is to identify which approach is expected to be most appropriate for providing specific information. Assume that, in the evaluation scenario we are at a point where the desiderata are not satisfied, the adequate degree and kind of understanding is not evoked, and the required explanatory information is not delivered. In this case, it is necessary to investigate whether the explainability approach is even capable of producing explanatory information with the right features at all. All the insights that are available at this point can indicate whether an existing explainability approach provides explanatory information that is sufficient to satisfy stakeholders' desiderata.

Additionally, we can learn whether a given explainability approach has the potential to derive explanatory information that can satisfy stakeholders' desiderata to a certain degree, whether it is necessary to adjust the explainability approach, whether it is sufficient to choose another one, or whether we need to develop an entirely new one. At this stage, computer scientists who can improve, adjust, and design explainability approaches are the main contributors integrating the afore-mentioned insights. First, they have the abilities to assess what is technically feasible and possible. Second, they can actually implement the demands regarding the explainability approach resulting from the previous steps.

6.2. Specific application scenario

We will conclude our thoughts about the application of our model by means of a specific example. Consider a situation where users want a system that produces fair outputs. For this, we first need an explanation process that enables them to assess whether the system produces fair outputs and, consequently, we need to find an adequate explainability approach. We first clarify the relevant user sub-classes and their prototypical characteristics. Will it be, for instance, novice or expert users? Then, we determine whether other stakeholder classes also have to be considered. For example, do we need to consider the perspectives of regulators regarding fairness? Furthermore, we have to anticipate the context. Are we talking about a personnel selection tasks or court cases with completely different contextual peculiarities?

Subsequently, we determine what users mean when they desire fair outputs (i.e., we clarify the satisfaction conditions of the desideratum's substantial facet). What kind of algorithmic fairness do they expect? At the same time, it is important to be aware of other relevant desiderata as the satisfaction of other user desiderata could be affected when trying to satisfy the fairness desideratum. Similarly, there may be unanticipated effects on the stakeholders' desiderata. For instance, will using a specific explainability approach that proves suitable to assess system fairness also affect predictive accuracy of the artificial system?

Next, we consider the desideratum's epistemic facet. Under which circumstances would users be enabled to assess whether the outputs of a system are fair? What do they need to understand with regard to a system's functioning or outputs? To answer these questions, we must determine what degree and what kind of understanding is appropriate for the epistemic satisfaction of the fairness desideratum and we need a detailed investigation of how contextual influences moderate this relation. Furthermore, we must be aware of given stakeholder characteristics (e.g., users' background knowledge).

When we have estimated what degree and kind of understanding is required to enable users to assess whether a system's outputs are fair, we can determine what kind of explanatory information facilitates this understanding. For example, we might need explanatory information about the influence of features based on protected attributes (e.g., race) on the system's outputs. Alternatively, we might need information that contrasts the treatment of different groups of people (e.g., minorities and majorities). With this knowledge, we can determine what explainability approach provides this kind of explanatory

information. Most likely, it would be post-hoc, local approaches that provide explanatory information that either highlights feature relevance or that allow users to compare subsets of instances regarding their predicted outcomes. As a result, we have specific demands regarding the explainability approach for satisfying the fairness desideratum in the case under consideration.

With this knowledge we can either choose an adequate explainability approach from existing ones or design a new one. Afterwards, we can investigate whether the explanatory information resulting from the respective explainability approach leads to a better understanding of the system and its outputs. This means that we are now in a position to empirically evaluate the success of the selected approach and the corresponding explanation process.

For this, we may conduct a stakeholder study. Given our definitions of desiderata satisfaction, we may find in this study that the epistemic facet may be satisfied (e.g., users can actually assess whether the system produces fair outputs; users can explain whether and why the respective system's outputs are fair; users can predict what kind of inputs will lead to fair or unfair outcomes) but not the substantial one (i.e., the system's outputs are actually not fair). Then, we may have to conclude that the substantial facet of the desideratum is not satisfiable by altering the explainability approach. Nevertheless, satisfying the epistemic facet of the desideratum can help us to figure out how to satisfy the substantial facet of the desideratum beyond XAI-related strategies. For instance, we may obtain information that a developer can use to improve system fairness. In this case, the artificial system has to be adjusted or changed to additionally satisfy the substantial facet of desiderata satisfaction.

On the other hand, if the stakeholder study reveals that the epistemic facet of the desideratum is not satisfied (e.g., users fail in explaining whether and why the respective system's outputs are fair), we can conclude that the users' degree or kind of understanding of the system and its outputs does not suffice. In many situations this may be the case because the explanatory information was not suitable to evoke the necessary understanding (e.g., the explainability approach did not provide explanatory information suitable to allow assessing whether the system produces fair outcomes). Depending on the context, the task, and stakeholders' individual characteristics, it may be necessary to iteratively adjust the characteristics of the explanatory information so that stakeholders engaging with this information achieve the right kind and degree of understanding. Alternatively, it may be necessary to adjust the explainability approach if it is not suitable to provide explanatory information useful for facilitating stakeholder understanding.

7. Conclusion

With increasing numbers of people affected by artificial systems, the number of stakeholders' desiderata will continue to grow. Although the focus of XAI research has shifted towards human stakeholders and the evaluation of explainability approaches, this shift still needs to incorporate a comprehensive view of all stakeholders and their desiderata when artificial systems are used in socially relevant contexts as well as empirical investigation of explainability approaches with respect to desiderata satisfaction. This engenders an even more pressing need for interdisciplinary collaboration in order to consider all stakeholders' perspectives and to empirically evaluate and optimally design explainability approaches to satisfy stakeholders' desiderata. The current paper has introduced a model highlighting the central concepts and their relations along which explainability approaches aim to satisfy stakeholders' desiderata. We hope that this model inspires and guides future interdisciplinary evaluation and development of explainability approaches and, thereby, further advances XAI research concerning the satisfaction of stakeholders' desiderata.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D.C. Brock, Learning from artificial intelligence's previous awakenings: the history of expert systems, AI Mag. 39 (3) (2018) 3–15, https://doi.org/10.1609/aimag.v39i3.2809.

[2] W.J. Clancey, The epistemology of a rule-based expert system – a framework for explanation, Artif. Intell. 20 (3) (1983) 215–251, https://doi.org/10.1016/0004-3702(83)90008-5.

[3] W.R. Swartout, Xplain: a system for creating and explaining expert consulting programs, Artif. Intell. 21 (3) (1983) 285–325, https://doi.org/10.1016/S0004-3702(83)80014-9.

[4] H. Johnson, P. Johnson, Explanation facilities and interactive systems, in: Proceedings of the 1st International Conference on Intelligent User Interfaces, IUI, Association for Computing Machinery, New York, NY, USA, 1993, pp. 159–166.

[5] O. Biran, C. Cotton, Explanation and justification in machine learning: a survey, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 8–13.

[6] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

[7] B.D. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in AI, in: Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, 2019, pp. 279–288.

[8] J. Burrell, How the machine 'thinks': understanding opacity in machine learning algorithms, Big Data Soc. 3 (1) (2016) 1–12, https://doi.org/10.1177/2053951715622512.

[9] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, CoRR, arXiv:1702.08608 [abs], 2017.

[10] EU High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, 2019.

[11] Z.C. Lipton, The mythos of model interpretability, Commun. ACM 61 (10) (2018) 36–43, https://doi.org/10.1145/3233231.

[12] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.

[13] I. Nunes, D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, User Model. User-Adapt. Interact. 27 (3–5) (2017) 393–444, https://doi.org/10.1007/s11257-017-9195-0.

[14] A.B. Arrieta, N. Díaz-Rodríguez, J.D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

[15] H. Felzmann, E.F. Villaronga, C. Lutz, A. Tamò-Larrieux, Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns, Big Data Soc. 6 (1) (2019) 1–14, https://doi.org/10.1177/2053951719860542.

[16] L.H. Gilpin, C. Testart, N. Fruchter, J. Adebayo, Explaining explanations to society, in: NIPS Workshop on Ethical, Social and Governance Issues in AI, 2018, pp. 1–6, arXiv:1901.06560 [abs].

[17] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: challenges and prospects, CoRR, arXiv:1812.04608 [abs], 2018.

[18] A.D. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in explainable AI, CoRR, arXiv:1810.00184 [abs], 2018.

[19] A. Weller, Transparency: motivations and challenges, in: W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 23–40.

[20] A. Páez, The pragmatic turn in Explainable Artificial Intelligence (XAI), Minds Mach. 29 (2019) 441–459, https://doi.org/10.1007/s11023-019-09502-w.

[21] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F.M. Harper, H. Zhu, Explaining decision-making algorithms through ui: strategies to help non-expert stakeholders, in: Proceedings of the 2019 chi Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–12.

[22] A. Abdul, J. Vermeulen, D. Wang, B.Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda, in: Proceedings of the 2018 Conference on Human Factors in Computing Systems, CHI, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–18.

[23] M.-W. Dictionary, Stakeholder, https://www.merriam-webster.com/dictionary/stakeholder, 2020. (Accessed 30 July 2020).

[24] M. Hind, D. Wei, M. Campbell, N.C.F. Codella, A. Dhurandhar, A. Mojsilović, K.N. Ramamurthy, K.R. Varshney, Ted: teaching AI to explain its decisions, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society, Association for Computing Machinery, New York, NY, USA, 2019, pp. 123–129.

[25] S. Anjomshoae, K. Främling, A. Najjar, Explanations of black-box model predictions by contextual importance and utility, in: Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer, 2019, pp. 95–109.

[26] M. Atzmueller, Towards socio-technical design of explicative systems: transparent, interpretable and explainable analytics and its perspectives in social interaction contexts information, in: Proceedings of the 2019 Workshop on Affective Computing and Context Awareness in Ambient Intelligence, AfCAI, 2019, pp. 1–8.

[27] I. Baaj, J.-P. Poli, W. Ouerdane, Some insights towards a unified semantic representation of explanation for explainable artificial intelligence, in: Proceedings of the 2019 Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, NL4XAI, Association for Computational Linguistics, 2019, pp. 14–19.

[28] K. Balog, F. Radlinski, S. Arakelyan, Transparent, scrutable and explainable user models for personalized recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, 2019, pp. 265–274.

[29] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, N. Shadbolt, 'It's reducing a human being to a percentage': perceptions of justice in algorithmic decisions, in: Proceedings of the 2018 Conference on Human Factors in Computing Systems, CHI, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–14.

[30] T. Chakraborti, S. Sreedharan, S. Grover, S. Kambhampati, Plan explanations as model reconciliation, in: 14th ACM/IEEE International Conference on Human-Robot Interaction, HRI, IEEE, 2019, pp. 258–266.

[31] L. Chen, D. Yan, F. Wang, User evaluations on sentiment-based recommendation explanations, ACM Trans. Interact. Intell. Syst. 9 (4) (2019) 1–38, https://doi.org/10.1145/3282878.

[32] K. Cotter, J. Cho, E. Rader, Explaining the news feed algorithm: an analysis of the "news feed fyi" blog, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1553–1560.

[33] K. Darlington, Aspects of intelligent systems explanation, Univers. J. Control Autom. 1 (2) (2013) 40–51, https://doi.org/10.13189/ujca.2013.010204.

[34] K. Ehrlich, S.E. Kirk, J. Patterson, J.C. Rasmussen, S.I. Ross, D.M. Gruen, Taking advice from intelligent systems: the double-edged sword of explanations, in: Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI, Association for Computing Machinery, New York, NY, USA, 2011, pp. 125–134.

[35] A.A. Freitas, Comprehensible classification models: a position paper, ACM SIGKDD Explor. Newsl. 15 (1) (2014) 1–10, https://doi.org/10.1145/2594473.2594475.

[36] S. Gregor, I. Benbasat, Explanations from intelligent systems: theoretical foundations and implications for practice, MIS Q. 23 (4) (1999) 497–530, https://doi.org/10.2307/249487.

[37] J. Hois, D. Theofanou-Fuelbier, A.J. Junk, How to achieve explainability and transparency in human AI interaction, in: International Conference on Human-Computer Interaction, HCI, Springer, 2019, pp. 177–183.

[38] R.F. Kizilcec, How much information? effects of transparency on trust in an algorithmic interface, in: Proceedings of the 2016 Conference on Human Factors in Computing Systems, CHI, Association for Computing Machinery, New York, NY, USA, 2016, pp. 2390–2395.

[39] S. Nagulendra, J. Vassileva, Providing awareness, explanation and control of personalized filtering in a social networking site, Inf. Syst. Front. 18 (1) (2016) 145–158, https://doi.org/10.1007/s10796-015-9577-y.

[40] A. Papenmeier, G. Englebienne, C. Seifert, How model accuracy and explanation fidelity influence user trust in AI, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 94–100.

[41] R. Pierrard, J.-P. Poli, C. Hudelot, A new approach for explainable multiple organ annotation with few data, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 101–107.

[42] V. Putnam, L. Riegel, C. Conati, Towards personalized XAI: a case study in intelligent tutoring systems, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 108–114.

[43] E. Rader, K. Cotter, J. Cho, Explanations as mechanisms for supporting algorithmic transparency, in: Proceedings of the 2018 Conference on Human Factors in Computing Systems, CHI, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–13.

[44] A. Rosenfeld, A. Richardson, Explainability in human–agent systems, Auton. Agents Multi-Agent Syst. 33 (6) (2019) 673–705, https://doi.org/10.1007/s10458-019-09408-y.

[45] M. Sato, K. Nagatani, T. Sonoda, Q. Zhang, T. Ohkuma, Context style explanation for recommender systems, J. Inf. Process. 27 (2019) 720–729, https://doi.org/10.2197/ipsjjip.27.720.

[46] J. Vig, S. Sen, J. Riedl, Tagsplanations: explaining recommendations using tags, in: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI, Association for Computing Machinery, New York, NY, USA, 2009, pp. 47–56.

[47] X. Watts, F. Lécué, Local score dependent model explanation for time dependent covariates, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 129–135.

[48] J. Zhou, H. Hu, Z. Li, K. Yu, F. Chen, Physiological indicators for user trust in machine learning with influence enhanced fact-checking, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2019, pp. 94–113.

[49] J.L. Herlocker, J.A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, CSCW, Association for Computing Machinery, New York, NY, USA, 2000, pp. 241–250.

[50] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, B. Wielinga, The effects of transparency on trust in and acceptance of a content-based art recommender, User Model. User-Adapt. Interact. 18 (5) (2008) 455, https://doi.org/10.1007/s11257-008-9051-3.

[51] R.M. Byrne, Counterfactuals in Explainable Artificial Intelligence (XAI): evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 2019, pp. 6276–6282.

[52] P.B. De Laat, Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?, Philos. Technol. 31 (4) (2018) 525–541, https://doi.org/10.1007/s13347-017-0293-z.

[53] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations, Minds Mach. 28 (4) (2018) 689–707, https://doi.org/10.1007/s11023-018-9482-5.

[54] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, P. Vinck, Fair, transparent, and accountable algorithmic decision-making processes, Philos. Technol. 31 (4) (2018) 611–627, https://doi.org/10.1007/s13347-017-0279-x.

[55] S.M. Mathews, Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review, in: Intelligent Computing – Proceedings of the Computing Conference, Springer, 2019, pp. 1269–1292.

[56] B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: mapping the debate, Big Data Soc. 3 (2) (2016) 1–21, https://doi.org/10.1177/2053951716679679.

[57] W. Pieters, Explanation and trust: what to tell the user in security and AI?, Ethics Inf. Technol. 13 (1) (2011) 53–64, https://doi.org/10.1007/s10676-010-9253-3.

[58] G. Ras, M. van Gerven, P. Haselager, Explanation methods in deep learning: users, values, concerns and challenges, in: Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, 2018, pp. 19–36.

[59] M.O. Riedl, Human-centered artificial intelligence and machine learning, Hum. Behav. Emerg. Technol. 1 (1) (2019) 33–36, https://doi.org/10.1002/hbe2.117.

[60] S. Robbins, A misdirected principle with a catch: explicability for AI, Minds Mach. 29 (4) (2019) 495–514, https://doi.org/10.1007/s11023-019-09509-3.

[61] R. Sheh, Different XAI for different HRI, in: AAAI Fall Symposium, AAAI Press, 2017, pp. 114–117.

[62] R. Sheh, I. Monteath, Defining explainable AI for requirements analysis, Künstl. Intell. 32 (4) (2018) 261–266, https://doi.org/10.1007/s13218-018-0559-3.

[63] K. Sokol, P.A. Flach, Explainability fact sheets: a framework for systematic assessment of explainable approaches, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, 2020, pp. 56–67.

[64] K. Sokol, P. Flach, One explanation does not fit all, Künstl. Intell. 34 (2) (2020) 235–250, https://doi.org/10.1007/s13218-020-00637-y.

[65] M. Sridharan, B. Meadows, Towards a theory of explanations for human–robot collaboration, Künstl. Intell. 33 (4) (2019) 331–342, https://doi.org/10.1007/s13218-019-00616-y.

[66] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Comput. Appl. (2019), https://doi.org/10.1007/s00521-019-04051-w.

[67] D. Wang, Q. Yang, A. Abdul, B.Y. Lim, Designing theory-driven user-centric explainable AI, in: Proceedings of the 2019 Conference on Human Factors in Computing Systems, CHI, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–15.

[68] M.K. Lee, A. Jain, H.J. Cha, S. Ojha, D. Kusbit, Procedural justice in algorithmic fairness, Proc. ACM Human-Comput. Interact. 3 (2019) 1–26, https://doi.org/10.1145/3359284.

[69] D. Doran, S. Schulz, T.R. Besold, What does explainable AI really mean? a new conceptualization of perspectives, in: CEUR Workshop Proceedings, vol. 2071, CEUR, 2018, pp. 1–8, arXiv:1710.00794.

[70] M. Krishnan, Against interpretability: a critical examination of the interpretability problem in machine learning, Philos. Technol. (2019) 1–16, https://doi.org/10.1007/s13347-019-00372-9.

[71] S.K. Peddoju, M. Saravanan, S. Suresh, Explainable classification using clustering in deep learning models, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 115–121.

[72] N.F. Rajani, R.J. Mooney, Using explanations to improve ensembling of visual question answering systems, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 43–47.

[73] J. Zhou, F. Chen, Towards trustworthy human-AI teaming under uncertainty, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 143–147.

[74] S. Anjomshoae, A. Najjar, D. Calvaresi, K. Främling, Explainable agents and robots: results from a systematic literature review, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, USA, 2019, pp. 1078–1088.

[75] M. Fox, D. Long, D. Magazzeni, Explainable planning, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 24–30.

[76] S. Jasanoff, Virtual, visible, and actionable: data assemblages and the sightlines of justice, Big Data Soc. 4 (2) (2017) 1–15, https://doi.org/10.1177/2053951717724477.

[77] G. Friedrich, M. Zanker, A taxonomy for generating explanations in recommender systems, AI Mag. 32 (3) (2011) 90–98, https://doi.org/10.1609/aimag.v32i3.2365.

[78] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019) 1–13, https://doi.org/10.1002/widm.1312.

[79] R. Sevastjanova, F. Beck, B. Ell, C. Turkay, R. Henkin, M. Butt, D.A. Keim, M. El-Assady, Going beyond visualization: verbalization as complementary medium to explain machine learning models, in: VIS Workshop on Visualization for AI Explainability, VISxAI, 2018, pp. 1–6.

[80] F. Sørmo, J. Cassens, A. Aamodt, Explanation in case-based reasoning–perspectives and goals, Artif. Intell. Rev. 24 (2) (2005) 109–143, https://doi.org/10.1007/s10462-005-4607-7.

[81] J. Zerilli, A. Knott, J. Maclaurin, C. Gavaghan, Transparency in algorithmic and human decision-making: is there a double standard?, Philos. Technol. 32 (4) (2019) 661–683, https://doi.org/10.1007/s13347-018-0330-6.

[82] A. Lucic, H. Haned, M. de Rijke, Contrastive explanations for large errors in retail forecasting predictions through Monte Carlo simulations, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 66–72.

[83] H.K. Dam, T. Tran, A. Ghose, Explainable software analytics, in: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, ICSE-NIER, Association for Computing Machinery, New York, NY, USA, 2018, pp. 53–56.

[84] J. De Winter, Explanations in software engineering: the pragmatic point of view, Minds Mach. 20 (2) (2010) 277–289, https://doi.org/10.1007/s11023-010-9190-2.

[85] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, F. Doshi-Velez, Explainable reinforcement learning via reward decomposition, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 47–53.

[86] L. Michael, Machine coaching, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 80–86.

[87] K. Sokol, P.A. Flach, Conversational explanations of machine learning predictions through class-contrastive counterfactual statements, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, 2018, pp. 5785–5786.

[88] H. Wicaksono, C. Sammut, R. Sheh, Towards explainable tool creation by a robot, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 63–67.

[89] T. Eiter, Z.G. Saribatur, P. Schüller, Abstraction for zooming-in to unsolvability reasons of grid-cell problems, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 7–13.

[90] T. Kulesza, S. Stumpf, W.-K. Wong, M.M. Burnett, S. Perona, A. Ko, I. Oberst, Why-oriented end-user debugging of naive Bayes text classification, ACM Trans. Interact. Intell. Syst. 1 (1) (2011) 1–31, https://doi.org/10.1145/2030365.2030367.

[91] R.R. Hoffman, G. Klein, S.T. Mueller, Explaining explanation for "explainable AI", Proc. Hum. Factors Ergon Soc. Ann. Meet. 62 (2018) 197–201, https://doi.org/10.1177/1541931218621047.

[92] F. Nothdurft, T. Heinroth, W. Minker, The impact of explanation dialogues on human-computer trust, in: International Conference on Human-Computer Interaction, HCI, Springer, 2013, pp. 59–67.

[93] C. Brinton, A framework for explanation of machine learning decisions, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 14–18.

[94] N. Tintarev, Explanations of recommendations, in: Proceedings of the 2007 ACM Conference on Recommender Systems, Association for Computing Machinery, New York, NY, USA, 2007, pp. 203–206.

[95] R.O. Weber, H. Hong, P. Goel, Explaining citation recommendations: abstracts or full texts?, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 136–142.

[96] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: an overview of interpretability of machine learning, in: IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA, 2018, pp. 80–89.

[97] J.C.-T. Ho, How biased is the sample? reverse engineering the ranking algorithm of facebook's graph application programming interface, Big Data Soc. 7 (1) (2020) 1–15, https://doi.org/10.1177/2053951720905874.

[98] F. Hohman, A. Head, R. Caruana, R. DeLine, S.M. Drucker, Gamut: a design probe to understand how data scientists understand machine learning models, in: Proceedings of the 2019 Conference on Human Factors in Computing Systems, CHI, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–13.

[99] M. Veale, R. Binns, Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data, Big Data Soc. 4 (2) (2017) 1–17, https://doi.org/10.1177/2053951717743530.

[100] C. Zednik, Solving the black box problem: a normative framework for explainable artificial intelligence, Philos. Technol. (2019) 1–24, https://doi.org/10.1007/s13347-019-00382-7.

[101] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI Mag. 38 (3) (2017) 50–57, https://doi.org/10.1609/aimag.v38i3.2741.

[102] E.I. Sklar, M.Q. Azhar, Explanation through argumentation, in: Proceedings of the 6th International Conference on Human-Agent Interaction, HAI, Association for Computing Machinery, New York, NY, USA, 2018, pp. 277–285.

[103] I. Lage, D. Lifschitz, F. Doshi-Velez, O. Amir, Exploring computational user models for agent policy summarization, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 59–65.

[104] E.S. Dahl, Appraising black-boxed technology: the positive prospects, Philos. Technol. 31 (4) (2018) 571–591, https://doi.org/10.1007/s13347-017-0275-1.

[105] B. Ghosh, D. Malioutov, K.S. Meel, Interpretable classification rules in relaxed logical form, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 14–20.

[106] M.T. Stuart, N.J. Nersessian, Peeking inside the black box: a new kind of scientific visualization, Minds Mach. 29 (1) (2019) 87–107, https://doi.org/10.1007/s11023-018-9484-3.

[107] J. Clos, N. Wiratunga, S. Massie, Towards explainable text classification by jointly learning lexicon and modifier terms, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 19–23.

[108] J. Zhu, A. Liapis, S. Risi, R. Bidarra, G.M. Youngblood, Explainable AI for designers: a human-centered perspective on mixed-initiative co-creation, in: IEEE Conference on Computational Intelligence and Games, CIG, IEEE, 2018, pp. 1–8.

[109] M.-A. Clinciu, H. Hastie, A survey of explainable AI terminology, in: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, NL4XAI 2019, Association for Computational Linguistics, 2019, pp. 8–13.

[110] C. Henin, D. Le Métayer, Towards a generic framework for black-box explanation methods, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 28–34.

[111] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, USA, 2019, pp. 1033–1041.

[112] M.L. Olson, R. Neal, F. Li, W.-K. Wong, Counterfactual states for Atari agents via generative deep learning, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 87–93.

[113] Z. Zeng, C. Miao, C. Leung, J.J. Chin, Building more explainable artificial intelligence with argumentation, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI-18, AAAI Press, 2018, pp. 8044–8046, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16762.

[114] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 73–79.

[115] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1135–1144.

[116] M.R. Endsley, From here to autonomy, Hum. Factors, J. Hum. Factors Ergon. Soc. 59 (1) (2017) 5–27, https://doi.org/10.1177/0018720816681350.

[117] J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance, Hum. Factors 46 (1) (2004) 50–80, https://doi.org/10.1518/hfes.46.1.50.30392.

[118] R. Parasuraman, V. Riley, Humans and automation: use, misuse, disuse, abuse, Hum. Factors, J. Hum. Factors Ergon. Soc. 39 (2) (1997) 230–253, https://doi.org/10.1518/001872097778543886.

[119] K.A. Hoff, M. Bashir, Trust in automation, Hum. Factors 57 (3) (2014) 407–434, https://doi.org/10.1177/0018720814547570.

[120] R. Parasuraman, D.H. Manzey, Complacency and bias in human use of automation: an attentional integration, Hum. Factors 52 (3) (2010) 381–410, https://doi.org/10.1177/0018720810376055.

[121] A. Kunze, S.J. Summerskill, R. Marshall, A.J. Filtness, Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces, Ergonomics 62 (3) (2019) 345–360, https://doi.org/10.1080/00140139.2018.1547842.

[122] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models, CoRR, arXiv: 1708.08296 [abs], 2017.

[123] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digit. Signal Process. 73 (2018) 1–15, https://doi.org/10.1016/j.dsp.2017.10.011.

[124] S. Becker, M. Ackermann, S. Lapuschkin, K. Müller, W. Samek, Interpreting and explaining deep neural networks for classification of audio signals, CoRR, arXiv:1807.03418 [abs], 2018.

[125] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, W. Samek, Analyzing classifiers: fisher vectors and deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2912–2920.

[126] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2015, pp. 1721–1730.

[127] K. Baum, H. Hermanns, T. Speith, From machine ethics to machine explainability and back, in: International Symposium on Artificial Intelligence and Mathematics, ISAIM, 2018, pp. 1–8, http://isaim2018.cs.virginia.edu/papers/ISAIM2018_Ethics_Baum_etal.pdf.

[128] C. Luetge, The german ethics code for automated and connected driving, Philos. Technol. 30 (2017) 547–558, https://doi.org/10.1007/s13347-017-0284-0.

[129] S.L.S. Purkiss, P.L. Perrewé, T.L. Gillespie, B.T. Mayes, G.R. Ferris, Implicit sources of bias in employment interview judgments and decisions, Organ. Behav. Hum. Decis. Process. 101 (2) (2006) 152–167, https://doi.org/10.1016/j.obhdp.2006.06.005.

[130] A. Caliskan, J.J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186, https://doi.org/10.1126/science.aal4230.

[131] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2019) 1–42, https://doi.org/10.1145/3236009.

[132] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis, User acceptance of information technology: toward a unified view, Manag. Inf. Syst. Q. 27 (3) (2003) 425–478, https://doi.org/10.2307/30036540.

[133] C. McLeod, Trust, in: E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, fall 2015 edition, Metaphysics Research Lab., Stanford University, 2015, pp. 1–43.

[134] I.D. Raji, A. Smart, R.N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT*, Association for Computing Machinery, New York, NY, USA, 2020, pp. 33–44.

[135] A. Matthias, The responsibility gap: ascribing responsibility for the actions of learning automata, Ethics Inf. Technol. 6 (3) (2004) 175–183, https://doi.org/10.1007/s10676-004-3422-1.

[136] E.L. Deci, A.H. Olafsen, R.M. Ryan, Self-determination theory in work organizations: the state of a science, Ann. Rev. Organ. Psychol. Organ. Behav. 4 (1) (2017) 19–43, https://doi.org/10.1146/annurev-orgpsych-032516-113108.

[137] C. Longoni, A. Bonezzi, C.K. Morewedge, Resistance to medical artificial intelligence, J. Consum. Res. 46 (4) (2019) 629–650, https://doi.org/10.1093/jcr/ucz013.

[138] F.C. Keil, Explanation and understanding, Annu. Rev. Psychol. 57 (1) (2006) 227–254, https://doi.org/10.1146/annurev.psych.57.102904.190100.

[139] J.-F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles, Science 352 (2016) 1573–1576, https://doi.org/10.1126/science.aaf2654.

[140] B.G. Buchanan, E.H. Shortliffe, Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison-Wesley, 1984.

[141] J.S. Dhaliwal, I. Benbasat, The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation, Inf. Syst. Res. 7 (3) (1996) 342–362, https://doi.org/10.1287/isre.7.3.342.

[142] M.A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, D. Bohlender, Explainability as a non-functional requirement, in: IEEE 27th International Requirements Engineering Conference, RE, 2019, pp. 363–368.

[143] H.W. De Regt, Understanding Scientific Understanding, Oxford University Press, 2017.

[144] C. Baumberger, C. Beisbart, G. Brun, What is understanding? An overview of recent debates in epistemology and philosophy of science, in: S.G.C. Baumberger, S. Ammon (Eds.), Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science, Routledge, 2017, pp. 1–34.

[145] F.I. Malfatti, On understanding and testimony, Erkenntnis (2019) 1–21, https://doi.org/10.1007/s10670-019-00157-8.

[146] C. Baumberger, Types of understanding: their nature and their relation to knowledge, Conceptus 40 (98) (2014) 67–88, https://doi.org/10.1515/cpt-2014-0002.

[147] K. Lambert, On whether an answer to a why-question is an explanation if and only if it yields scientific understanding, in: G.G. Brittan (Ed.), Causality, Method, and Modality, vol. 48, Springer, Dordrecht, Netherlands, 1991, pp. 125–142.

[148] T. Lombrozo, S. Carey, Functional explanation and the function of explanation, Cognition 99 (2) (2006) 167–204, https://doi.org/10.1016/j.cognition.2004.12.009.

[149] M.T. Chi, N. De Leeuw, M.-H. Chiu, C. Lavancher, Eliciting self-explanations improves understanding, Cogn. Sci. 18 (3) (1994) 439–477, https://doi.org/10.1207/s15516709cog1803_3.

[150] R.E. Mayer, Cognition and instruction: their historic meeting within educational psychology, J. Educ. Psychol. 84 (4) (1992) 405–412, https://doi.org/10.1037/0022-0663.84.4.405.

[151] S.T. Mueller, R.R. Hoffman, W. Clancey, A. Emrey, G. Klein, Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI, CoRR, arXiv:1902.01876 [abs], 2019.

[152] C. Kelp, Understanding phenomena, Synthese 192 (12) (2015) 3799–3816, https://doi.org/10.1007/s11229-014-0616-x.

[153] P.J. Feltovich, R.L. Coulson, R.J. Spiro, Learners' (mis)understanding of important and difficult concepts: a challenge to smart machines in education, in: Smart Machines in Education: The Coming Revolution in Educational Technology, MIT Press, Cambridge, MA, USA, 2001, pp. 349–375.

[154] W.B. Rouse, N.M. Morris, On looking into the black box: prospects and limits in the search for mental models, Psychol. Bull. 100 (3) (1986) 349–363, https://doi.org/10.1037/0033-2909.100.3.349.

[155] L. Rozenblit, F. Keil, The misunderstood limits of folk science: an illusion of explanatory depth, Cogn. Sci. 26 (5) (2002) 521–562, https://doi.org/10.1207/s15516709cog2605_1.

[156] D. Kuhn, How do people know?, Psychol. Sci. 12 (1) (2001) 1–8, https://doi.org/10.1111/1467-9280.00302.

[157] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, Too much, too little, or just right? Ways explanations impact end users' mental models, in: IEEE Symposium on Visual Languages and Human Centric Computing, IEEE, 2013, pp. 3–10.

[158] J. Tullio, A.K. Dey, J. Chalecki, J. Fogarty, How it works: a field study of non-technical users interacting with an intelligent system, in: Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI, Association for Computing Machinery, New York, NY, USA, 2007, pp. 31–40.

[159] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 2019, pp. 220–229.

[160] M. Langer, C.J. König, A. Fitili, Information as a double-edged sword: the role of computer experience and information on applicant reactions towards novel technologies for personnel selection, Comput. Hum. Behav. 81 (2018) 19–30, https://doi.org/10.1016/j.chb.2017.11.036.

[161] D.T. Newman, N.J. Fast, D.J. Harmon, When eliminating bias isn't fair: algorithmic reductionism and procedural justice in human resource decisions, Organ. Behav. Hum. Decis. Process. 160 (2020) 149–167, https://doi.org/10.1016/j.obhdp.2020.03.008.

[162] M. Bazire, P. Brézillon, Understanding context before using it, in: A. Dey, B. Kokinov, D. Leake, R. Turner (Eds.), Modeling and Using Context, Springer, 2005, pp. 29–40.

[163] P. Dourish, What we talk about when we talk about context, Pers. Ubiquitous Comput. 8 (1) (2004) 19–30, https://doi.org/10.1007/s00779-003-0253-8.

[164] D.R. Bobocel, A. Zdaniuk, How can explanations be used to foster organizational justice, in: Handbook of Organizational Justice, 2005, pp. 469–498.

[165] R. Folger, R. Cropanzano, Fairness theory: justice as accountability, in: Advances in Organizational Justice, vol. 1, 2001, pp. 1–55.

[166] J.C. Shaw, E. Wild, J.A. Colquitt, To justify or excuse?: A meta-analytic review of the effects of explanations, J. Appl. Psychol. 88 (3) (2003) 444–458, https://doi.org/10.1037/0021-9010.88.3.444.

[167] J. Brockner, B.M. Wiesenfeld, An integrative framework for explaining reactions to decisions: interactive effects of outcomes and procedures, Psychol. Bull. 120 (2) (1996) 189–208, https://doi.org/10.1037/0033-2909.120.2.189.

[168] R. Wang, F.M. Harper, H. Zhu, Factors influencing perceived fairness in algorithmic decision-making: algorithm outcomes, development procedures, and individual differences, in: Proceedings of the 2020 Conference on Human Factors in Computing Systems, CHI, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–14.

[169] E.A. Lind, K. van den Bos, When fairness works: toward a general theory of uncertainty management, Res. Organ. Behav. 24 (2002) 181–223, https://doi.org/10.1016/s0191-3085(02)24006-x.

[170] J.A. Colquitt, J.M. Chertkoff, Explaining injustice: the interactive effect of explanation and outcome on fairness perceptions and task motivation, J. Manag. 28 (5) (2002) 591–610, https://doi.org/10.1177/014920630202800502.

[171] P. Liu, Z. Li, Task complexity: a review and conceptualization framework, Int. J. Ind. Ergon. 42 (6) (2012) 553–568, https://doi.org/10.1016/j.ergon.2012.09.001.

[172] D.A. Wilkenfeld, Functional explaining: a new approach to the philosophy of explanation, Synthese 191 (2014) 3367–3391, https://doi.org/10.1007/s11229-014-0452-z.

[173] D.A. Wilkenfeld, D. Plunkett, T. Lombrozo, Depth and deference: when and why we attribute understanding, Philos. Stud. 173 (2016) 373–393, https://doi.org/10.1007/s11098-015-0497-y.

[174] T. Lombrozo, The instrumental value of explanations, Philos. Compass 6 (8) (2011) 539–551, https://doi.org/10.1111/j.1747-9991.2011.00413.x.

[175] J.J. Williams, T. Lombrozo, Explanation and prior knowledge interact to guide learning, Cogn. Psychol. 66 (1) (2013) 55–84, https://doi.org/10.1016/j.cogpsych.2012.09.002.

[176] T. Lombrozo, B. Rehder, Functions in biological kind classification, Cogn. Psychol. 65 (4) (2012) 457–485, https://doi.org/10.1016/j.cogpsych.2012.06.002.

[177] C.G. Hempel, Deductive-nomological explanation, in: Aspects of Scientific Explanation, Free Press, 1965, pp. 335–376.

[178] W.C. Salmon, Scientific Explanation and the Causal Structure of the World, Princeton University Press, 1984.

[179] P. Gärdenfors, Knowledge in Flux: Modeling the Dynamics of Epistemic States, MIT Press, 1988.

[180] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, Harv. J. Law Technol. 31 (2) (2018), https://doi.org/10.2139/ssrn.3063289.

[181] C.F. Craver, Explaining the Brain, Oxford University Press, Oxford, 2007.

[182] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd edition, Cambridge University Press, 2009.

[183] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction and Search, 2nd edition, MIT Press, 2001.

[184] D. Borsboom, A. Cramer, A. Kalis, Brain disorders? Not really … why network structures block reductionism in psychopathology research, Behav. Brain Sci. 42 (2018) 1–54, https://doi.org/10.1017/S0140525X17002266.

[185] T. Lombrozo, Simplicity and probability in causal explanation, Cogn. Psychol. 55 (3) (2007) 232–257, https://doi.org/10.1016/j.cogpsych.2006.09.006.

[186] N. Vasilyeva, D. Wilkenfeld, T. Lombrozo, Contextual utility affects the perceived quality of explanations, Psychon. Bull. Rev. 24 (2017) 1436–1450, https://doi.org/10.3758/s13423-017-1275-y.

[187] V. Bellotti, K. Edwards, Intelligibility and accountability: human considerations in context-aware systems, Hum.-Comput. Interact. 16 (2–4) (2001) 193–212, https://doi.org/10.1207/s15327051hci16234_05.

[188] K. Hartley, L.D. Bendixen, Educational research in the internet age: examining the role of individual characteristics, Educ. Res. 30 (9) (2001) 22–26, https://doi.org/10.3102/0013189X030009022.

[189] H. Kauffman, A review of predictive factors of student success in and satisfaction with online learning, Res. Learn. Technol. 23 (2015), https://doi.org/10.3402/rlt.v23.26507.

[190] D.S. McNamara, E. Kintsch, N.B. Songer, W. Kintsch, Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text, Cogn. Instr. 14 (1) (1996) 1–43, https://doi.org/10.1207/s1532690xci1401_1.

[191] L. Goldberg, Language and individual differences: the search for universals in personality lexicons, in: L. Wheeler (Ed.), Review of Personality and Social Psychology, vol. 2, Edition, SAGE Publications, 1981, pp. 141–166.

[192] J.T. Cacioppo, R.E. Petty, The need for cognition, J. Pers. Soc. Psychol. 42 (1) (1982) 116–131, https://doi.org/10.1037/0022-3514.42.1.116.

[193] C.P. Haugtvedt, R.E. Petty, Personality and persuasion: need for cognition moderates the persistence and resistance of attitude changes, J. Pers. Soc. Psychol. 63 (2) (1992) 308–319, https://doi.org/10.1037/0022-3514.63.2.308.

[194] T.K. DeBacker, H.M. Crowson, The influence of need for closure on learning and teaching, Educ. Psychol. Rev. 21 (2009) 303–323, https://doi.org/10.1007/s10648-009-9111-1.

[195] D.M. Webster, A.W. Kruglanski, Individual differences in need for cognitive closure, J. Pers. Soc. Psychol. 67 (6) (1994) 1049–1062, https://doi.org/10.1037/0022-3514.67.6.1049.

[196] P.M. Fernbach, S.A. Sloman, R.S. Louis, J.N. Shube, Explanation fiends and foes: how mechanistic detail determines understanding and preference, J. Consum. Res. 39 (5) (2012) 1115–1131, https://doi.org/10.1086/667782.

[197] L. Hasher, R.T. Zacks, Working memory, comprehension, and aging: a review and a new view, in: Psychology of Learning and Motivation, Elsevier, 1988, pp. 193–225.

[198] R. Ackerman, T. Lauterman, Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure, Comput. Hum. Behav. 28 (5) (2012) 1816–1828, https://doi.org/10.1016/j.chb.2012.04.023.

[199] M.S. Prewett, R.C. Johnson, K.N. Saboe, L.R. Elliott, M.D. Coovert, Managing workload in human–robot interaction: a review of empirical studies, Comput. Hum. Behav. 26 (5) (2010) 840–856, https://doi.org/10.1016/j.chb.2010.03.010.

[200] K. Starcke, O.T. Wolf, H.J. Markowitsch, M. Brand, Anticipatory stress influences decision making under explicit risk conditions, Behav. Neurosci. 122 (6) (2008) 1352–1360, https://doi.org/10.1037/a0013281.

[201] S. Lupien, F. Maheu, M. Tu, A. Fiocco, T. Schramek, The effects of stress and stress hormones on human cognition: implications for the field of brain and cognition, Brain Cogn. 65 (3) (2007) 209–237, https://doi.org/10.1016/j.bandc.2007.02.007.

[202] P.A. Hancock, On the process of automation transition in multitask human–machine systems, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 37 (4) (2007) 586–598, https://doi.org/10.1109/tsmca.2007.897610.

[203] L. Chazette, O. Karras, K. Schneider, Do end-users want explanations? Analyzing the role of explainability as an emerging aspect of non-functional requirements, in: IEEE 27th International Requirements Engineering Conference, RE, 2019, pp. 223–233.

[204] L. Chazette, K. Schneider, Explainability as a non-functional requirement: challenges and recommendations, Requir. Eng. 25 (4) (2020) 493–514, https://doi.org/10.1007/s00766-020-00333-1.

[205] V. Arya, R.K.E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S.C. Hoffman, S. Houde, Q.V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J.T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K.R. Varshney, D. Wei, Y. Zhang, One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques, CoRR, arXiv:1909.03012 [abs], 2019.

[206] J. Woodward, Scientific explanation, in: E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, winter 2019 edition, Metaphysics Research Lab., Stanford University, 2019, pp. 1–101.

[207] M. Hall, D. Harborne, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle, A. Preece, A systematic method to understand requirements for explainable AI (XAI) systems, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 21–27.

[208] T. Miller, P. Howe, L. Sonenberg, Explainable AI: beware of inmates running the asylum, or: how I learnt to stop worrying and love the social and behavioural sciences, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 36–42.

[209] B. Kim, C. Rudin, J.A. Shah, The Bayesian case model: a generative approach for case-based reasoning and prototype classification, in: Advances in Neural Information Processing Systems, 2014, pp. 1952–1960.

[210] B. Kim, R. Khanna, S. Koyejo, Examples are not enough, learn to criticize! Criticism for interpretability, in: Advances in Neural Information Processing Systems, 2016, pp. 2280–2288.

[211] J.M. Carroll, M.B. Rosson, Paradox of the active user, in: Interfacing Thought: Cognitive Aspects of Human-Computer Interaction, MIT Press, Cambridge, MA, USA, 1987, pp. 80–111.