Requirements for Explainable Smart Systems in the Enterprises from Users and Society Based on FAT

Yuri Nakao
FUJITSU LABORATORIES LTD.
Kawasaki, Japan
nakao.yuri@jp.fujitsu.com

Junichi Shigezumi
FUJITSU LABORATORIES LTD.
Kawasaki, Japan
j.shigezumi@jp.fujitsu.com

Hikaru Yokono
FUJITSU LABORATORIES LTD.
Kawasaki, Japan
yokono.hikaru@jp.fujitsu.com

Takuya Takagi
FUJITSU LABORATORIES LTD.
Kawasaki, Japan
takagi.takuya@jp.fujitsu.com

ABSTRACT

As statistical methods for smart systems prevail, requirements related to explainability based on fairness, accountability, and transparency (FAT) become stronger. While end users need to confirm the fairness of the output of smart systems at a glance, society needs thorough explanations based on FAT. In this paper, we offer a conceptual framework for considering the explainability of smart systems in enterprises practically. A conceptual model that has two layers, a core layer and interface layer, is provided, and we discuss the ideal environment in which there exist explainable smart systems that can meet the demands of both users and society.

CCS CONCEPTS

• **Applied computing → Enterprise architecture frameworks**;

KEYWORDS

Explainable Smart Systems; FAT; Explainability; Enterprise

ACM Reference Format:
Yuri Nakao, Junichi Shigezumi, Hikaru Yokono, and Takuya Takagi. 2019. Requirements for Explainable Smart Systems in the Enterprises from Users and Society Based on FAT. In *Joint Proceedings of the ACM IUI 2019 Workshops.* Los Angeles, USA, March 20, 2019, 7 pages.

1 INTRODUCTION

Explainability for algorithmic systems has been needed since the initial stage of research on intelligent systems [1]. This is because those who use the analysis results obtained with intelligent systems are human decision makers, and they need to judge the validity of the results. For conventional rule-based intelligent systems, it is not difficult to keep the process of decision-making explainable. However, recent statistical methods for smart systems, such as machine learning or especially deep learning, have difficulty with direct extraction of explanations that users or other human stakeholders can accept [16]. Therefore, as machine learning becomes popular, the social necessity for explainable smart systems also becomes stronger.

This necessity for explainability for smart systems is categorized into two sides: the user side and social side (Figure 1). From the social side, it can be pointed out that demand has increased for explainability based on fairness, accountability, and transparency (FAT). For example, in the current situation of data regulation, it is mentioned that meaningful information is needed related to decisions made with algorithmic systems [27]. Such a necessity for explainability that exists

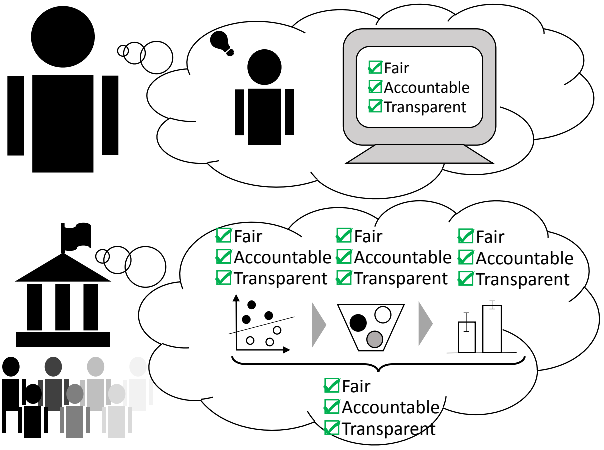

Figure 1: FAT needed from end users and from society. End user (upper) wants to know fairness, accountability, and transparency at glance. Society (lower) needs to know that all phases of data processing are FAT.

IUI Workshops'19, March 20, 2019, Los Angeles, USA

on the basis of discussion on FAT because the entire process of decision-making involving the use of support systems needs to be explained to confirm FAT. From the user side, the explainability of smart systems is necessary basically because users should understand the results of systems used for the decisions made around themselves or should operate the systems and adjust the results as they like [23, 29].

Explainable smart systems in enterprises ideally need to meet demands from both users and society at the same time regarding the explainability described above. These demands exist regardless of whether the enterprises are public organizations or private companies because, while private companies have trade secrets or intellectual property to protect, enterprises that are responsible for the needs of society also have their own citizens or customers who are recipients of their services. To meet the demands of both users and society, enterprises have to overcome difficulties that occur because of a trade-off between the two. When they try to meet the social demand for explainability, complete accountability and transparency are ideally needed. This means that the complex process of decision-making has to be traceable, and the reason for each decision has to be able to be elaborated on when requested. This evokes the problem of information overload from the user side because the processes of decision-making in enterprises are multi-tiered, various, and difficult for users to understand at first glance. Some research on machine learning considering fairness takes how to avoid information overload into account [12, 24, 31]. However, the necessity of avoiding information overload limits the transparency or traceability of decision-making processes.

In this paper, we offer a conceptual framework for considering the explainability of smart systems in enterprises. We suggest considering systems in organizations as integrated smart systems and splitting the discussion on the systems in regard to two layers: a core layer and interface layer. The core layer is the part that has functions for guaranteeing that the FAT required by society is met. The interface layer is the part for discussing how to filter information from the core layer to make it understandable to users. We focus mainly on the core layer because the interface layer is discussed well in existing papers. Discussion on the core layer is described not only from a technological perspective but from an organizational one. Our contribution to the community is that we clarify the differences in the necessity for explainability from the perspective of users and of society, and we suggest a conceptual framework of the explainability needed in enterprises.

2 RELATED WORK

There have been several discussions on the explainability of smart systems in enterprises. As statical approaches for data analysis are popularized, several articles summarize statistical methods for industry from the perspective of explainability [7] or appropriate industrial domains to apply statistical methods to [9]. Several papers discuss explainability from the perspective of users [10, 11]. For example, Chander et al. focus on how explanations extracted by artificial intelligence are accepted by different types of users and discuss the appropriateness of the results from state-of-the-art explanation methods for each type of users [10].In these papers, a system is considered to be a single agent that extracts explanations that were decided on by a single department in charge. However, in reality, smart systems have to be considered as the integrated systems because insider the systems, there are some heterogeneous methods or criteria for data processing.

Besides the explainability from the user side that is discussed in existing papers, there is a need for explainability from the perspective of the social side because there are several stakeholders, such as governments or public organizations, that require FAT for both public and private enterprises. While there is discussion on the transparency of information in heterogeneous organizations in the field of ethics [30], there are no concrete requirements for achieving FAT with explainable smart systems in enterprises. In this paper, our focus is to make a conceptual framework for breaking issues on explainable smart systems down into concrete requirements for the environments in which the systems exist.

As practical measures, there have been several technological approaches to FAT in the industry community. For example, IBM launched new Trust and Transparency capabilities on their cloud service[1]. To keep AI systems fair, their system provides functions that automatically detect and produce alerts for biases in decisions made by the system. The functions visualize the confidence scores of data recommended to be added to the model used for the system. Related to this kind of explanation in the meaning of visualization, several companies, such as simMachines[2], and Fujitsu [21] have developed machine learning technologies for extracting results in ways that are understandable to users. However, these approaches focus only on how explanations of fairness are intelligible to their users. While such approaches can confirm the understandability of explanations, they cannot cover all FAT issues. In this paper, by focusing on the social side, explainability can be discussed on the basis of FAT. This will help in developing new technologies that can be used to thoroughly confirm explainability in our ideal environment in which systems that are explainable from the perspective of FAT exist, and we offer a list of ideal conditions.

[1] https://newsroom.ibm.com/2018-09-19-IBM-Takes-Major-Step-in-Breaking-Open-the-Black-Box-of-AI

[2] https://simmachines.com/

3 DETAILS OF EXPLAINABILITY

To develop our conceptual framework for discussing the explainability of smart systems in enterprises, we need to elaborate on the explainability required from users and from society in terms of the necessity of FAT in smart systems.

FAT as a Reason for Explainability

Nowadays, requirements for the explainability of algorithmic systems from domestic governments, academic communities[3], or international organizations are based on discussion about FAT [18]. As web services or decision support systems with complex algorithms prevail, many people are making decisions under the influence of algorithms without knowing it [14]. This problematic situation is pointed out within the concepts named "algorithmic awareness" [2, 19], "filter bubble" [26], or "social echo chamber" [20]. Discussions on these have lead to requirements for displaying existence of algorithms explicitly, clarifying the effects of algorithms used for daily decisions, and making algorithmic systems transparent.

The social necessity for explainability based on FAT is especially seen in the context of data regulations explicitly. There are requirements in the GDPR, a data regulation from the EU taking effect globally, for algorithmic systems to extract meaningful information related to decisions made by the systems. The necessity for meaningful information is mainly due to the necessity for users to obtain enough information on algorithmic decision-making to have an actionable discrimination claim [27]. Thus, the necessity for explainability to show fairness in smart systems results in the necessity for accountability and transparency to confirm fairness in the entire process of algorithmic decision-making. Therefore, it is important to consider explainability from the perspective of FAT.

To discuss the details of this kind of explainability, we elaborate on fairness, accountability, and transparency as the first step. We define fairness, accountability, and transparency in our context in the following part as summarized in Table 1.

Fairness. Fairness is a multifaceted concept that varies according to the domain focused on. For example, in the field of machine learning, there are two definitions of fairness: group fairness and individual fairness [17, 32]. According to Zemel et al., group fairness is considered as the concept meaning that the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole, and the definition of individual fairness is that similar individuals should be treated similarly [32]. Moreover, there are diverse ways of considering the definition of fairness outside of the machine learning community [6]. The definition varies depending on what kind of justice is considered [6], or, more simply, on who should be protected, to what extent protected people should be protected, and so on.

Therefore, the concept of fairness that we discuss here is that there is no bias regarding the data of individuals that are thought to be sensitive by stakeholders related to discrimination in the results of analyses or decisions supported with smart systems. This definition of fairness means that the content of fairness itself changes flexibly depending on stakeholders including customers, government, and other social groups, but fairness in one domain should be confirmed among the stakeholders.

Accountability. In conventional discussions, accountability is a concept that is achieved not only with algorithms but by all parts of an enterprise [15]. The accountability of algorithms should be considered together with that of groups of people in enterprises. As a foundation for accountability in the field of computer science, Nissenbaum stated that accountability can be considered something related to moral blameworthiness [25]. In brief, blameworthiness is defined as the conditions that someone's actions caused a harm and her or his actions were faulty.In her article, while it is mentioned that blameworthiness is not identical to accountability, accountability is grasped by "*the nature of an action and the relationship of the agent (or several agents) to the action's outcome.*" Moreover, Diakopoulos describes the demand toward accountable algorithms as what exists simultaneously with the demand toward accountability of the people behind the algorithms [15].

Following these discussions, we consider the concept of accountability as a concept that indicates situations in which an organization possesses a structure that confirms the validity of the output of each phase of data processing in decision-making supported by smart systems, and, at the same time, of the output of the entire process. For example, to confirm the validity of the output of a process, there has to be identical department responsible for the phase. The reason that the responsibilities for each phase of data processing and the entire process of it are described respectively is that confirming the validity of each result from each phase of data processing is different from confirming the validity of the results from the entire process. For example, even if the outputs of each phase are fair, the outputs of the entire process of analysis are sometimes unfair because training data extracted from a previous phase of data processing that seem to be fair can cause unfair result in a later phase [5]. Moreover, this definition of accountability implies that there is one requirement for the methods embedded in smart systems for which departments have a responsibility, that is, that the methods should be

[3] IEEE ethically aligned design https://ethicsinaction.ieee.org/

Table 1: The Summary of Our Definition of Fairness, Accountability, and Transparency

	Definition
Fairness	• There are no biases regarding sensitive features related to discrimination. • Sensitive features are decided by stakeholders and, if needed, changed flexibly.
Accountability	• Methods employed in all phases of data processing can provide reasons for the outputs. • The structure of an organization confirms responsibilities for each and entire data analysis results.
Transparency	• Outsiders can trace the entire process of data analysis or algorithmic decision-making.

equipped with mechanisms that can provide reasons for the outputs for those who have responsibility.

Transparency. The concept of transparency is considered both in the field of computer science and in that of ethics. In the former, transparency means to show the internal functions of a system. For example, in the field of recommender systems or of data analysis, when a system explains the reason results were extracted, the system becomes more transparent [13, 22, 28]. In the latter, transparency is considered to be a macro concept that focuses on the entire process of operation related to data in organizations. The concept covers what occurs among people, data, or algorithmic systems, including practices, norms, and other factors [3, 4] or all phases of data processing [30]. It is useful to introduce transparency as a concept for clarifying the details of data processing to develop smart systems in enterprises.

According to the discussions above, we define the concept of transparency in the context of smart systems in enterprises to mean that the entirety of data processing can be traced by outsiders if needed. Of course, there are difficulties in achieving thorough transparency because there are matters of privacy, trade secrets, or other sensitive factors. However, whether full information is really disclosed or not, it is necessary to confirm that it is possible for third parties to trace the detailed process of algorithmic decision-making. Additionally, the definition is consistent with transparency considered in the field of computer science in that traceability of data processing is achieved by expressing the results of processing in interpretable ways.

Two Types of Explainability

Explainability based on FAT that is needed as per the discussions above should ideally always be provided to users both in and outside of an enterprise. However, realistically speaking, this is difficult because of the complex structure of the data analysis process. Therefore, the necessity for explanation from the end users' perspective should be separated from that of society (Figure 2).

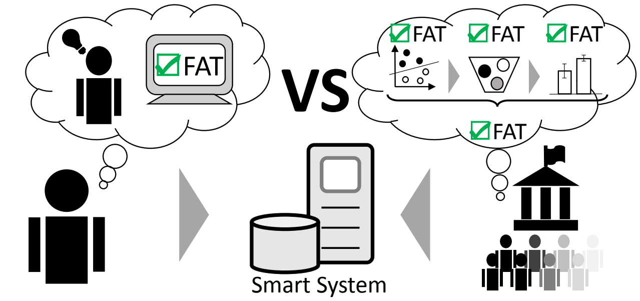

Figure 2: End user and society need different types of FAT for same system. There is trade-off between those two needs.

Needs from End Users. The explainability needed from the user side has been considered in research on how users recognize, understand, or learn about explanations from smart systems [23, 29]. Additionally, several pieces of existing research on machine learning focus on how easily end users read the results of analyses [24, 31]. We name the need to be able to read results as assumed in the field of computer science "readability."

It is difficult to confirm explainability for accountability and transparency when there has to be readability as well. This is because, while fairness can be guaranteed in each single process of analysis, accountability and transparency need to be considered among multiple processes or departments, which is hard for end users to understand at first glance. Therefore, when explainability based on readability for users is discussed, entire FAT can not be guaranteed.

Needs from Society. The social need for FAT in enterprises has become stronger as data driven analysis technologies become more popular. Current regulations [18, 27] and visions[4] mention the importance or necessity of FAT in algorithmic systems. Additionally, there are also important concepts such as filter bubble or algorithm awareness [2, 19, 26]. Among

[4] Ibid.

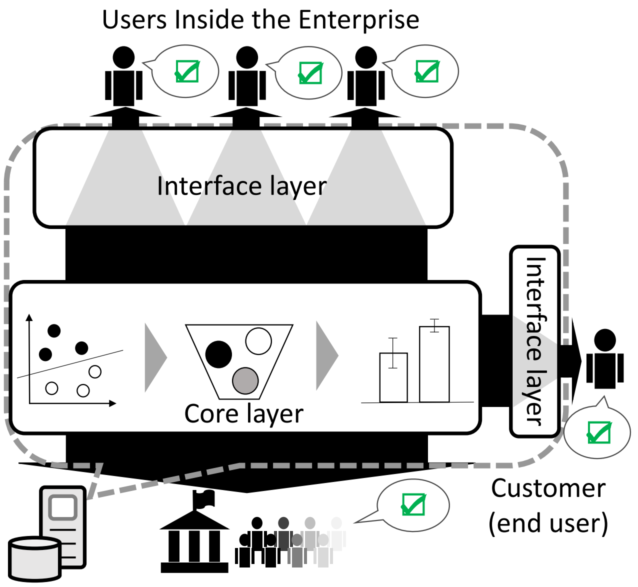

Figure 3: Two layers of smart system: core layer and interface layer. Core layer is to meet social demands, and interface layer is for meeting users' demands. Interface layer filters information from core layer not only for end users but for users inside enterprises. Core layer has potential ability to show the full information on data processing to society.

these discussions, both public and private enterprises are usually faced with a demand for thorough confirmation of FAT. We use the word "thorough" in the sense that confirmation is not limited because of readability. Of course, protecting privacy or trade secrets has to be considered, but the social requirements for FAT are that enterprises be fair, accountable, and transparent as long as possible.

4 CONCEPTUAL FRAMEWORK

According to the discussions above, the functions for meeting demands from individual users and those from society are different and should be discussed separately. However, this is difficult because these two types of demands are easily jumbled, for they have to be considered as requirements for the same system in an enterprise. Existing research has not suggested frameworks for discussing explainability based on FAT separately for the purpose of meeting these two different demands. Therefore, we need to visualize a framework for considering these separate discussions first.

Core Layer and Interface Layer

We suggest that the structure of a smart system be considered in two separate parts: a core layer and interface layer. We show the structure of the two layers in Figure 3. Simply speaking, the interface layer is set to meet demands from end users, and the core layer is for meeting those from society. The interface layer covers FAT aware technologies that take readability into consideration [24] and discussions on what kind of information should be displayed to end users [10]. The core layer covers discussions on confirming FAT for the entire structure of an enterprise [30] and technologies that focus on making results fair without considering readability. Technologies for showing the potential results of analyses or decision-making, such as technologies related to data structure, can be helpful for considering FAT for the core layer [8]. Compared with the technologies discussed for the core layer, technologies for the interface layer can be interpreted as technologies that play the role of filtering information from the core layer and showing readable results to users. Discussions related to the core layer include the structural issues of enterprises. However, there are few studies in the field of computer science on which environments explainable smart systems can exist with respect to FAT. Therefore, the ideal conditions under which explainable smart systems can exist have to be discussed.

Explainability and FAT in Smart Systems

As a first step toward thinking of the ideal environment in which FAT-aware explainable smart systems exist in enterprises, we offer a fivefold list (Table 2) based on our definition of FAT described in the previous section. In the list, there are technological factors and structural ones for composing such an environment. We use the word "structural" to mean what is related to the organizational structure of enterprises. While the list in Table 2 is an example of the requirements for an environment in which there are explainable smart systems, it is helpful when taking different definitions of FAT in account.

Detecting discriminative features. To guarantee the fairness of results, smart systems should detect features related to discriminative outputs. According to our definition of fairness, results from FAT-aware explainable smart systems must not extract outputs that have biases regarding sensitive features. Therefore, technologies for detecting sensitive factors need to be implemented in the systems. Moreover, not only sensitive factors such as gender or nationality but features that are correlated with the features (i.e. one's present address) should ideally be detected. Some conventional works in the field of machine learning approach this problem [32].

Adjustability of factors in the model. After detecting the features related to unfair outputs, users should be able to adjust, that is, delete or add features. This is because of our definition of fairness, that is, that sensitive features related to discrimination are decided by stakeholders and change flexibly in accordance with the context. In technological manners, this means that users including decision makers can change what

Table 2: The Summary of Requirements of FAT for Smart Systems

type of factor	requirements	F	A	T
technological	1. Detection of discriminative features	✓	-	-
technological	2. Adjustability of features in the model	✓	(✓)	(✓)
technological	3. Interpretability of all models used in a system	-	✓	✓
structural	4. Departments in charge of each process	-	✓	✓
structural	5. Departments in charge of entire process	-	✓	-

features should be included or excluded in terms of fairness. This function of smart systems helps people in enterprises account for their decisions in consideration of fairness. This means the function helps to keep the environment accountable and, moreover, transparent indirectly.

Interpretability in all models used in system. All models in smart systems need to be interpretable to confirm the accountability and transparency of the systems. According to our definition of accountability, details on the mechanisms of methods in explainable smart systems have to be composed in the way that the method can provide results for its output. Therefore, the outputs of the technologies in systems including statistical methods have to be able to be interpreted by human decision makers. This condition would be helpful for confirming transparency because, in order for outsiders to trace the process of data analysis, it is necessary that the output of each phase of analyses be shown in a form that outsiders can understand.

Departments in charge of each process. There have to be departments that are responsible for each phase of data processing in enterprises to confirm accountability and transparency. This requirement is not related to technology but to the structure of an enterprise. While this condition corresponds to a part of our definition of accountability, what is important is that this requirement is useful when considering it with respect to transparency. When outsiders trace the phases of data processing, they have to ask for the release of detailed information in some way or other. For this request to be effective, each public or private organization has to have an individual department or group of people responsible for each phase of data processing internally. This kind of requirement for enterprises shows that departments in charge of each process are needed in terms of transparency.

Departments in charge of the entire process. To guarantee accountability, the ideal environment has to have a department, a group of people, or an individual that is responsible for the final output of a system. This requirement is based on our definition of accountability that the structure of an enterprise in which an explainable smart system exists has to have a function for confirming the validity or rationality of the final results of decision-making. Precisely speaking, making a process for decision-making transparent does not directly mean confirming accountability [4]. Therefore, departments that are responsible for the final results of data processing must guarantee accountability.

5 CONCLUSION AND FUTURE WORK

We discussed requirements related to explainability based on FAT from two sides, the user side and the social side. Describing these two sides and giving detailed definitions on fairness, accountability, and transparency, we clarified the differences and trade-offs between the two sides. After this, we suggested a conceptual framework that splits the structure of smart systems in enterprises into two layers: a core layer and interface layer. On the basis of this framework, we focused on the core layer and set the ideal environment in which explainable smart systems based on FAT exist with a fivefold table. We suggested a first step toward considering the environment including not only the systems themselves but also the structure of the organization of enterprises. As future work, technologies or the principles for meeting both social and user demands will be built on the basis of our framework.

In this paper, we proposed an abstract conceptual framework to recognize the situation around the discussion of FAT from two perspectives. To show the usefulness of our framework, it is preferred that there are case studies, such as analysis of specific FAT-aware technology or of structure of enterprises. With the case studies, there can be chances to obtain the guidance to implementation in technological manners and methods of organizational analysis of various enterprises with our framework.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*

(CHI '18). ACM, New York, NY, USA, Article 582, 18 pages.

[2] Oscar Alvarado and Annika Waern. 2018. Towards Algorithmic Experience: Initial Efforts for Social Media Contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 286, 12 pages.

[3] Mike Ananny. 2016. Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology, & Human Values* 41, 1 (Sep. 2016), 93–117.

[4] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (Dec. 2018), 973–989.

[5] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (Jun. 2016), 671–732.

[6] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*'18)*. PMLR, New York, NY, USA, 149–159.

[7] Indranil Bose and Radha K. Mahapatra. 2001. Business data mining - a machine learning perspective. *Information & Management* 39, 3 (Dec. 2001), 211 – 225.

[8] Randal E. Bryant. 1992. Symbolic Boolean Manipulation with Ordered Binary-decision Diagrams. *ACM Comput. Surv.* 24, 3 (Sep. 1992), 293–318.

[9] Erik Brynjolfsson and Tom Mitchell. 2017. What can machine learning do? Workforce implications. *Science* 358, 6370 (Dec. 2017), 1530–1534.

[10] Ajay Chander and Ramya Srinivasan. 2018. Evaluating Explanations by Cognitive Value. In *Machine Learning and Knowledge Extraction*. Springer International Publishing, Cham, Switzerland, 314–328.

[11] Ajay Chander, Ramya Srinivasan, Suhas Chelian, Jun Wang, and Kanji Uchino. 2018. Working with Beliefs: AI Transparency in the Enterprise. In *Joint Proceedings of the ACM IUI 2018 Workshops*.

[12] Nicole Cruz, Jean Baratgin, Mike Oaksford, and David E. Over. 2015. Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology* 6, Article 192 (Feb. 2015), 9 pages.

[13] Anupam Datta, Shayak Sen, and Yair Zick. 2017. *Algorithmic Transparency via Quantitative Input Influence*. Springer International Publishing, Cham, Switzerland, 71–94.

[14] Nicholas Diakopoulos. 2014. Algorithmic Accountability Reporting: On the Investigation of Black Boxes. *Tow Center for Digital Journalism A Tow/Knight Brief* (Dec. 2014), 1–32.

[15] Nicholas Diakopoulos. 2016. Accountability in Algorithmic Decision Making. *Commun. ACM* 59, 2 (Jan. 2016), 56–62.

[16] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning As a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 278–288.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. ACM, New York, NY, USA, 214–226.

[18] Lilian Edwards and Michael Veale. 2018. Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'? *IEEE Security & Privacy* 16, 3 (Jul. 2018), 46–54.

[19] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn'T Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 153–162.

[20] Seth Flaxman, Sharad Goel, and Justin M. Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly* 80, S1 (Mar. 2016), 298–320.

[21] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. 2018. Explainable AI: The New 42?. In *Machine Learning and Knowledge Extraction*. Springer International Publishing, Cham, Switzerland, 295–303.

[22] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, New York, NY, USA, 241–250.

[23] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 126–137.

[24] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1675–1684.

[25] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and Engineering Ethics* 2, 1 (Mar. 1996), 25–42.

[26] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. The Penguin Group, London, UK.

[27] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (Dec. 2017), 233–242.

[28] Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. ACM, New York, NY, USA, 830–831.

[29] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (Aug. 2009), 639 – 662.

[30] Matteo Turilli and Luciano Floridi. 2009. The ethics of information transparency. *Ethics and Information Technology* 11, 2 (Jun. 2009), 105–112.

[31] Fulton Wang and Cynthia Rudin. 2015. Falling Rule Lists. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. PMLR, San Diego, California, USA, 1013–1022.

[32] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML'13)*. JMLR.org, III–325–III–333.