Operationalizing Human-Centered Perspectives in Explainable AI

Upol Ehsan*
Georgia Institute of Technology
Atlanta, GA, USA
ehsanu@gatech.edu

Philipp Wintersberger*
CARISSMA, Technische Hochschule Ingolstadt (THI)
Ingolstadt, Bavaria, GERMANY
philipp.wintersberger@carissma.eu

Q. Vera Liao
IBM Research AI
Yorktown Heights, NY, USA
vera.liao@ibm.com

Martina Mara
Johannes Kepler University Linz
Linz, Upper Austria, AUSTRIA
martina.mara@jku.at

Marc Streit
Johannes Kepler University Linz
Linz, Upper Austria, AUSTRIA
marc.streit@jku.at

Sandra Wachter
Oxford Internet Institute, University of Oxford
Oxford, England, UK
sandra.wachter@oii.ox.ac.uk

Andreas Riener
Technische Hochschule Ingolstadt (THI)
Ingolstadt, Bavaria, GERMANY
andreas.riener@thi.de

Mark O. Riedl
Georgia Institute of Technology
Atlanta, GA, USA
riedl@cc.gatech.edu

ABSTRACT

The realm of Artificial Intelligence (AI)'s impact on our lives is far reaching – with AI systems proliferating high-stakes domains such as healthcare, finance, mobility, law, etc., these systems must be able to explain their decision to diverse end-users comprehensibly. Yet the discourse of Explainable AI (XAI) has been predominantly focused on algorithm-centered approaches, suffering from gaps in meeting user needs and exacerbating issues of algorithmic opacity. To address these issues, researchers have called for human-centered approaches to XAI. There is a need to chart the domain and shape the discourse of XAI with reflective discussions from diverse stakeholders. The goal of this workshop is to examine how human-centered perspectives in XAI can be operationalized at the conceptual, methodological, and technical levels. Encouraging holistic (historical, sociological, and technical) approaches, we put an emphasis on "operationalizing", aiming to produce actionable frameworks, transferable evaluation methods, concrete design guidelines, and articulate a coordinated research agenda for XAI.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; *Visualization theory, concepts and paradigms*; *Visualization design and evaluation methods*; • **Computing methodologies** → Philosophical/theoretical foundations of artificial intelligence.

*Both authors contributed equally to this proposal.

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05.
https://doi.org/10.1145/3411763.3441342

KEYWORDS

Explainable Artificial Intelligence, Interpretable Machine Learning, Interpretability, Artificial Intelligence, Critical Technical Practice, Human-centered Computing, Trust in Automation, Algorithmic Fairness

ACM Reference Format:
Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3411763.3441342

1 INTRODUCTION

Our lives are increasingly algorithmically mediated by Artificial Intelligence (AI) systems. The purview of these systems have reached consequential and safety-critical domains such as healthcare, finance, automated driving, etc. Despite their continuously improving capabilities, these AI systems suffer from opacity issues where the mechanics underlying their decisions often remain invisible or incomprehensible to end-users. Crucial to trustworthy and accountable Human-AI collaboration thus is the *explainability* of AI systems—these systems need to be able to make its decisions explainable and comprehensible to affected humans [18, 23, 27].

Explainability has been sought as primary means, even fundamental rights, for people to understand AI in order to contest and improve AI, guarantee fair and ethical AI, as well as to foster human-AI cooperation. For example, the European parliament states that AI systems should be *"understandable to non-technical audiences and providing them with meaningful information, which is necessary to evaluate fairness and gain trust"*, since an entrusted expert group concluded, *"whenever an AI system has a significant impact on people's lives, it should be possible to demand*

a suitable explanation of the decision making process." Despite initial regulatory steps towards algorithmic accountability, how to achieve this goal in different usage contexts and accommodate different groups of stakeholders remains poorly understood. Consequently, "explainable artificial intelligence" (XAI) has become a prominent interdisciplinary domain in the past years [3], including researchers from fields such as machine learning, data science and visualization, human-computer interaction/human factors, design, psychology, or law. Although XAI has been a fast growing field, there is no agreed-upon definition of it, let alone methods to evaluate it, nor guidelines to create XAI technology. Discussions to chart the domain and shape these important topics call for human-centered and socio-technical perspectives, input from diverse stakeholders, as well as participation of the broader HCI community.

In this workshop, we offer a junction point of cross-disciplinary stakeholders of the XAI landscape—from designers to engineers, from researchers to end-users. Speaking to CHI21' main theme (*"making waves, combining strengths"*), the goal of this workshop is to examine how human-centered perspectives in XAI can be operationalized at the conceptual, methodological, and technical levels towards a Human-Centered Explainable AI (HCXAI). We put the emphasis on *"operationalizing"*: aiming to produce actionable frameworks, contextually transferable evaluation methods, concrete design guidelines, etc. for explainable AI, and encourage a holistic approach (historical, sociological, and technical) when it comes to articulating operationalization of these human-centered perspectives.

2 TOWARDS HUMAN-CENTERED EXPLAINABLE ARTIFICIAL INTELLIGENCE

For a systematic approach for human-centered XAI, we first need to establish a common grounds for the discourse. Broadly speaking, from a Social Construction of Technology (SCOT) perspective [6], the current fluidity in XAI can be characterized as *relevant social groups* (e.g, different stakeholders like AI researchers, policy makers, practitioners, etc.) having *interpretive flexibility* (different interpretations) on the constructs in the field. In different communities, related terms like explainability, interpretability, intelligibility, and transparency have been used interchangeably [1, 3, 22]. Many define explainability as a property of an AI system's functioning or decisions being *easy to understand* by people [2, 3, 15]. Explainability is often viewed more broadly than model transparency, or directly interpretable models [14, 20, 25]. For example, a growing area of XAI is concerned with generating *post-hoc* explanations [11], which do not necessarily describe how a model works, but *justify* an opaque model's decision in an accessible manner [20].

Proper operationalization seeks to contextually situate ambiguities amongst research communities when it comes to definitions, concepts, methodology, and evaluation. The interpretive flexibility requires precision in articulating the issues, which can be achieved by proper operationalization of the concepts, methods, and techniques. Operationalization does not carry normativity – it does not attempt to force a *closure* where there is no need for alternative

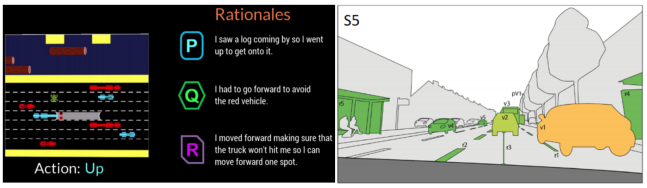

Figure 1: Recent works of the authors on human-centered real-time capable explanations. In [10, 11], the AI agent generated natural language rationales real-time while playing the computer game *Frogger* (left). This approach may be extended to real environments - in [28], potential passengers of automated vehicles selected objects in the driving environment that should be included in explanations to make decisions of driving algorithms more comprehensible.

design. In fact, the recent growth spurt in XAI indicates that we are far from a closure of interpretations or methods.

The algorithm-centered XAI community, mostly consisting of machine learning researchers, has been criticized for *"inmates running the asylum"* [21], as explanations techniques are often developed based on researchers' own intuition rather than the needs of their intended audience. Bringing user-centered perspectives to the forefront of XAI is necessary as the field work towards supporting diverse types of users and stakeholders, such as data scientists developing models, decision-makers using AI systems, people who might be affected by AI, regulatory agencies, etc [4]. How they interpret and react to explanations might vary significantly depending on the motivational, social, cognitive as well as professional and educational profiles [10, 12].

When adopting a human-centered XAI approach, the question is not just about the *"who"* (the explanation for is essential), but also about the *why* [10]. Reasons for explanations include *"trustworthiness, causality, transferability, informativeness, fairness, accessibility, interactivity, or privacy awareness"* [3]. Consequently, understanding the *who* and the *why* is crucial because it governs what the explanation requires for a given problem. It also scopes how the data is collected, what data can be collected, and the most effective way of describing the why behind an action. For instance: with automated vehicles, the engineer may have different requirements of explainability than the rider in that car. By studying why end-users seek explanations across various AI systems, Liao et al. [19] summarize the user goals that XAI should aim to support include to gain further insights for decision-making, to appropriately evaluate AI's capability, to adapt their usage or control of the AI, and to improve AI performance.

Another important aspect is *"where"*, the application domain or context for XAI. Many XAI contributions addressed case-based explanations, for example using local (specific to the decision/instance) or global (perspective of the overall model) methods [8]. Recent work has introduced XAI features in model development tools [12, 17], AI-assisted decision-support tools [29], for model fairness evaluation [8], etc. With a goal for accessible explanations for end-users (who might not be AI experts), there has been progress in natural language based explanations generated real-time. For

instance, Ehsan et al. [11] used a neural machine translation approach to generate rationales in plain English to justify the actions of an AI agent completing a sequential decision-making task. Generating accessible explanations real-time has critical implications in domains such as automated driving [28] (see Figure 1).

In potentially safety-critical domains, skepticism towards AI decisions may be more relevant than unlimited trust to prevent misuse [24] (i.e., over-reliance causing harm). Here, the XAI community may learn from other research areas like human factors, where over-trust (for example in aviation or driving automation) has been discussed for quite some time.

Besides definitions and methodologies, another relevant topic is the evaluation of XAI systems, as without proper metrics, *"any claim in this respect dilutes among the literature, not providing a solid ground on which to stand"* [3]. The AI community has approached evaluation by defining normative, quantifiable metrics, such as fidelity, completeness, stability, compactness, etc. [7, 26]. Arriata et al. [3] emphasize that defining metrics should *"be approached by the community as a whole"* to allow evaluating and comparing XAI approaches under different application contexts, models and purpose. Mohseni et al. [22] performed an in-depth review of 69 papers and identified diverse types of evaluation methods used in the literature, including objective (task performance, user prediction of model output, compliance/reliance, etc.) and subjective measurements (interviews, surveys, self-reports, etc.). The authors categorized evaluation measures in the five themes *Mental Models* (how the AI works), *Usefulness and Satisfaction, User Trust and Reliance, Human-AI Task Performance*, and *Computational Measures* (correctness and completeness in terms of explaining what the model has learned). Hoffman et al. [16] introduced the "Explanation Satisfaction Scale" consisting of 8 items, which address factors such as understanding, satisfaction, completeness, accuracy, or trust. Given that the effectiveness of XAI depends on the *who, where*, and *why*, it is necessary to evaluate XAI by involving targeted users and with the targeted context [9], while using evaluation methods that fully capture impact but also user experience of a given explanation.

Despite the progress in XAI, there are gaps in critically situating the interpretive flexibility of our perspectives. We are yet to systematically understand how we can transfer the technologies we build to real-world AI systems, ones that are socially-situated and culturally-embedded – we need a critically constructive community-wide discourse around these issues to address our intellectual blind spots and develop the human-centered XAI lens in a systematic manner.

3 GOALS OF THE WORKSHOP

By facilitating as a junction point of diverse perspectives from relevant stakeholders in XAI, the goal is to achieve clarity in charting the future of XAI from historical, sociological, and technological perspectives. Bridging works from researchers, designers, and practitioners from the fields of XAI, HCI, psychology, machine learning, and social sciences, we want to re-center our focus on the human. We want to operationalize our understanding of the different facets of XAI at the conceptual, methodological, and technical levels. Operationalization can include aspects such as frameworks, transferable evaluation methods, actionable design guidelines, etc.

Thus, we are interested in a wide range of topics, from sociotechnical aspects of XAI to human-centered evaluation techniques to responsible use of XAI. We are especially interested in the discourse around one or more of the questions discussed above: *who* (e.g., clarifying *who* the human is in XAI, how different *who's* interpret explainability), *why* (e.g., how social and individual factors influence explainability goals), and *where* (e.g., contextual explainability differences in diverse application areas). With an effort towards equitable conversations, we particular welcome participation from the Global South and from stakeholders whose voices are underrepresented in the dominant XAI discourse. The following list of guiding questions, by no means, is an exhaustive one; rather, it is provided as sources of inspiration:

- Who are the consumers and relevant stakeholders of XAI? What are their needs for explainability? What values are reflected and tensions arise in these needs?
- Why is explainability sought? What user goals should XAI aim to support? How are these goals shaped by technological, individual and social factors?
- Where, or in what categories of AI applications, should we prioritize our XAI efforts on? What do we need to understand about the users as well as the socio-organizational contexts of these applications?
- What are we missing from a technocentric view of XAI? Which human-centered and socio-technical perspectives should we bring in to better understand the who, why, where, to move towards human-centered XAI?
- How can we develop transferable evaluation methods for XAI? What key constructs need to be considered?
- Given the contextual nature of explanations, what are the potential pitfalls of standardization of evaluation metrics? How to take into account the who, why, and where in the evaluation methods?
- What are the explainability challenges where we move beyond the dominant one-to-one Human-AI interaction paradigms? How might a human-centered perspective address these challenges?
- What are the important research questions to be answered when we move towards a human-centered explainable artificial intelligence? Why are they important to be addressed now?

4 AUDIENCE

We expect approximately 30 (max 40) participants excluding the organizers. A call for participation will be distributed through HCI-related mailing lists (such as CHI Announcements), as well as our own lists of potential participants from previous workshops (complying with EU GDPR). We will further reach out to the IUI, psychology, ML and pervasive computing communities as well as the official ACM SIGCHI community to disseminate the call. The organizers will also use their social media accounts to advertise the call for participation.

5 WORKSHOP STRUCTURE

To facilitate a successful virtual format, we build on existing experience in hosting virtual workshops [5]. Since student volunteers are unlikely to be available, beyond session chairs and breakout shepherds, we will have dedicated organizers to handle technical issues, manage Discord conversations, and handles other logistical issues. The main part of the workshop will be held in form of two 3-hour sessions (including breaks) on two subsequent days (May 8th/May 9th 2021, 9:00AM-12:00AM EST each day) to allow participation from a wide range of time zones (see Table 1). Those will be supported by asynchronous pre- and post-workshop activities. Before the workshop, we will prepare a collection of reading material on our workshop website[1] to allow newcomers getting informed about the theoretical foundations of the workshop topics. Further, we will set up a Discord channel and encourage participants to introduce themselves via messages or short videos. We will incentivize asynchronous activities with virtual ice-breakers and fun activities (e.g., virtual scavenger hunts that can be done asynchronously). The goal is to get as many introductions done as possible. Participants can also find group members should the alignment be there. Dedicated organizers will manage the workspaces . On workshop day 1 (Saturday), we start with introducing the topic, the overall workshop goals (as we cannot guarantee that all participants will have read the provided material), and reserve some time so that participants who have not provided personal information in the Discord channel to briefly introduce themselves.

To spark interest and motivation, we will include a short keynote by an invited speaker, either a researcher focusing on XAI, or a member of an institution that relies on algorithmic decision making to bring in the perspective of end-users. In the second session of day 1, position papers will be presented in form of pre-recorded videos (length depending on the number of accepted submissions). We will emphasize participants to note down questions regarding these submissions to be asked in the following panel discussion.

Then, chaired by organizers, we will have a virtual panel discussion including authors of all position papers and the keynote speaker (given a high number of accepted position papers, we will conduct two rounds of panel discussions, combining similar topics). We choose this format to allow attendees preparing questions for the respective authors (the chair will provide ice-breaker questions), and as we believe this could result in a more lively discussion than typical post-presentation Q&As. After the panel discussion, shepherded by dedicated organizers, participants will be split into breakout groups (ideally, some groups have been established already before the workshop in our discord channel).

Groups are intended to individually work on selected (or emerging) topics related to the workshop goals. Individual groups can consider the *"who, where, and why"*-questions (see above). For instance, they may elaborate on an experimental design to investigate the interplay of explanations and involved constructs (trust, fairness, etc.) or develop taxonomies/frameworks capturing around diverse user groups and application areas. To do so, we will set up Miro boards for each breakout session as a master board to regrouping later. Such a format can be useful to provide an experience similar to a physical workshop were groups work at different tables in the same room. Organizers will be shepherd each group facilitating discussion and managing time.

Prior experience organizing virtual workshops has shown us that allotting adequate time for in-depth group discussion is essential [5]. Thus, we want to give participants enough time to come up with interesting insights. Participants will use the first 90-minute session (with a 10-min break min-point) on day 2 (Sunday) to continue working. Finally, the break-out groups wrap up their findings in form of short presentations (contents of the Miro board) and present/discuss their findings with the whole audience. At the end, the authors will wrap up the workshop, including the most relevant results, an exploration of potential future work, and actively promote research cooperation between attendees.

Table 1: Draft of the workshop structure, suggesting two 3-hour sessions (including breaks) on two subsequent days, as well as asynchronous activities before and after the workshop.

Time	Duration	Session
Before the Workshop		
-	2 weeks	Participants introduce themselves in the Discord channel and have access to provided workshop-related materials
*Workshop **Day 1**, Saturday 5/8/2021 9:00AM EST*		
9:00	35min	Introduction of workshop organizers, (remaining) participants, topics, and goals
9:35	15min	Keynote by invited speaker
10 min break		
10:00	50min	Video presentations of position papers
10:50	30min	Position paper panel discussion(s)
10 min break		
11:30	30min	Breakout group building and beginning of group work
*Workshop **Day 2**, Sunday 5/9/2021 9:00AM EST*		
9:00	45min	Breakout group work
10 min break		
9:55	45min	Breakout group work (continued)
10 min break		
10:50	40min	Break-out group findings presentations
11:30	30min	Workshop wrap up
After the workshop		
t+2w	-	Results summary posted on workshop website Initiating follow-up activities

6 PLANNED OUTCOMES

We will devote time after the final session to discuss, as a community, best practices to share the workshop's contributions and continue the discourse. The workshop's website (https://hcxai.jimdosite.com/) will serve as an archival repository of all the position papers and recorded sessions. The discussions and position papers in the workshop, hopefully, will generate series of articles and/or expanded papers which can be published on peer-reviewed

[1] https://hcxai.jimdosite.com/

journals (special issues such as TOCHI; e.g, how [13] lead to [30]). Depending on stakeholder preferences we will set up virtual spaces for continued discussions (e.g., Discord Workspaces) and explore future workshops in diverse venues such as NeurIPS, FAccT, IUI, and AIES to raise awareness and bridge translational human-centered XAI work.

7 ORGANIZERS

Upol Ehsan is a doctoral student in the School of Interactive Computing at Georgia Tech. Adopting a sociotechnically informed human-centered lens, he focuses on explainability of AI systems. Bridging his background in philosophy and engineering, his work resides at the intersection of AI and HCI with a focus on designing *explainable, encultured,* and *ethical* technology. His work has pioneered the notion of Rationale Generation in XAI and also charted the vision for a Reflective Human-centered XAI. His current focus is on advancing a socially-sitated understanding of XAI that will expand our understanding of explainability beyond its current techno-centric roots.

Q. Vera Liao is a Research Staff Member in IBM Research AI. Her current research focuses are on explainable AI and conversational agents. Her research work received multiple awards at ACM CHI and IUI, and contributed to IBM's Watson Assistant and AI Explainability 360. She actively organizes events that connect the HCI and AI communities, including several workshops and a panel at CHI, IUI and CSCW. She serves on the Editorial Board of International Journal of Human-Computer Studies (IJHCS) and ACM Transactions on Interactive Intelligent Systems (TiiS) , the Organizing Committee for IUI 2019 and CSCW 2021. She received Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign, and Bachelor in Industrial Engineering from Tsinghua University.

Martina Mara is a Professor of Robopsychology at the Johannes Kepler University Linz since 2018. Prior to that, she has worked for non-university research institutions such as the Ars Electronica Futurelab for more than a decade. Her work focuses on human-centered technology development and psychological aspects of AI and Robotics. She teaches Tech & Media Psychology or Responsible AI. As a member of the Austrian Council for Robotics and AI (ACRAI) she advises political decision makers. Among others, Martina has been awarded the Vienna Women's Prize for Digitization.

Mark Riedl is an Associate Professor in Georgia Tech's College of Computing and Associate Director of the Machine Learning Center at Georgia Tech. His research focuses on making agents better at understanding humans and communicating with humans. His research includes commonsense reasoning, story telling and understanding, explainable AI, and safe AI systems. He is a recipient of an NSF CAREER Award and a DARPA Young Faculty Award.

Andreas Riener is professor for Human Machine Interface and Virtual Reality at Technische Hochschule Ingolstadt (THI) with co-appointment at the CARISSMA research center. He is program manger for User Experience Design and leads the usability research and driving simulator labs. In 2017, he founded the interdisciplinary Human-Computer Interaction Group. His reserach interests include ergonomics, driver state assessment, and trust/acceptance/ethics in automated driving.

Mark Streit is a Full Professor of Visual Data Science at the Johannes Kepler University Linz, Austria. He finished his PhD at Graz University of Technology in 2011. His scientific areas of interest include visualization, visual analytics, and explainable AI. He won multiple best paper and runner-up awards at InfoVis, EuroVis, BioVis, and CHI. Marc is also co-founder and CEO of datavisyn, a spin-off company that develops data visualization solutions for applications in pharmaceutical and biomedical R&D.

Sandra Wachter is an Associate Professor in Data Ethics, Artificial Intelligence, Robotics and Internet Regulation at the University of Oxford, a Fellow at The Alan Turing Institute, and a Visiting Professor at Harvard University. Her work covers legal and ethical issues associated with Big Data, AI and algorithms. Sandra is a member of the World Economic Forum Council on Values, Ethics and Innovation, an affiliate at the Bonavero Institute of Human Rights and a member of the European Commission Expert Group on Autonomous Cars.

Philipp Wintersberger is a researcher at the research center CARISSMA/THI. He obtained his doctorate in Engineering Science from Johannes Kepler University Linz specializing Human-Computer Interaction and Human-Machine Cooperation. He worked 10 years as a software engineer/architect before joining the Human-Computer Interaction Group at CARISSMA/THI to research in the area of Human Factors and Driving Ergonomics. His publications focus on trust in automation, attentive user interfaces, transparency of driving algorithms, as well as UX/acceptance of automated vehicles and have received several awards in the past years.

8 CALL FOR PARTICIPATION

AI-powered decisions increasingly pervade consequential domains of our lives in high-stakes domains (healthcare, finance, legal). Explainability has been sought as primary means, even fundamental rights, for people to understand, contest to foster equitable and just Human-AI collaboration. Although explainable AI (XAI) has been a fast-growing field, there is no agreed-up definition of, let alone methods to evaluate and guidelines to create XAI technologies. Discussions to chart the domain and shape these important topics call for human-centered and sociotechnical perspectives. In this workshop, we offer a junction point of cross-disciplinary stakeholders of the XAI landscape– from designers to engineers, from researchers to end-users. The goal is to examine how human-centered perspectives in XAI can be operationalized at the conceptual, methodological and technical levels. Consequently, we call for papers up to 6 pages excluding references that address topics involving the who (e.g., relevant diverse stakeholders), why (e.g., social/individual factors influencing explainability goals), or where (e.g., diverse application areas or evaluation). Papers should follow the CHI Extended Abstract format and be submitted through the workshop's submission site [2]. All accepted papers will be presented, provided at least one author attends the workshop and registers at least one day of the conference. Further, contributing authors are invited to provide their views in form of short panel discussions with the workshop audience. With an effort towards an equitable discourse, we particular welcome participation from the Global

[2] https://hcxai.jimdosite.com/

South and from stakeholders whose voices are under-represented in the dominant XAI discourse.

ACKNOWLEDGMENTS

This work is supported under the FH-Impuls program of the German Federal Ministry of Education and Research (BMBF) under Grant Number 13FH7I01IA (SAFIR). We are grateful to members of the Human-centered AI lab at Georgia Tech for their input during brainstorming of these ideas.

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*. 1–18.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[5] Zoe M. Becerra, Nadia Fereydooni, Andrew L. Kun, Angus McKerral, Andreas Riener, Clemens Schartmüller, Bruce N. Walker, and Philipp Wintersberger. 2020. Interactive Workshops in a Pandemic...The Real Benefits of Virtual Spaces. *submitted to IEEE Pervasive Computing* (2020).

[6] Wiebe E Bijker, Thomas P Hughes, Trevor Pinch, et al. 1987. The social construction of technological systems.

[7] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[8] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.

[9] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *stat* 1050 (2017), 2.

[10] Upol Ehsan and Mark O Riedl. 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. *arXiv preprint arXiv:2002.01092* (2020).

[11] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

[12] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2020. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2020).

[13] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. 2016. Human-centred machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 3558–3565.

[14] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[15] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[16] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[17] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[18] Andreas Holzinger. 2018. Explainable AI (ex-AI). *Informatik-Spektrum* 41, 2 (2018), 138–143.

[19] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[20] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[21] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[22] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv* (2018), arXiv–1811.

[23] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).

[24] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[25] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. 2018. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 19–36.

[26] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 56–67.

[27] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[28] Philipp Wintersberger, Hannah Nicklas, Thomas Martlbauer, Stephan Hammer, and Andreas Riener. 2020. Explainable Automation: Personalized and Adaptive UIs to Foster Trust and Understanding of Driving Automation Systems. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Virtual Event, DC, USA) *(AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 252–261. https://doi.org/10.1145/3409120.3410659

[29] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

[30] G. Zhou, J. Lu, C.-Y. Wan, M. D. Yarvis, and J. A. Stankovic. 2008. *Body Sensor Networks*. MIT Press, Cambridge, MA.