How Can AI Earn Trust of System Administrators in the IT-Security Domain?

DARIA SOROKO, University of Bremen, Germany
GIAN-LUCA SAVINO, University of St. Gallen, Switzerland
NICHOLAS GRAY, Julius Maximilian University of Würzburg, Germany

As networks grow in size and complexity and the number of cyberattacks increases, securing the network becomes challenging. Partial workflow automation with the help of Artificial Intelligence (AI) has become a recent remedy to address these concerns and help system administrators to handle large volumes of data and keep the system safeguarded. However, although beneficial, AI-aided software solutions in the network security context often come at the cost of transparency and control, suffering from the opacity of their algorithms. The subsequent lack of understanding about the system, its decisions and outcomes cultivates mistrust on the system administrator's behalf, which leads to the overall reluctance to use the new software solutions, slower response time in case of a security breach, and inability to resolve conflicts in the system. A recent study by Ehsan et al. focused on the concept of social transparency (ST). It showed that when applied to the IT-security context, peer support could help alleviate the issue of mistrust in the system and provide the missing knowledge to the sysadmin when faced with a novel situation. In this position paper, we outline our current in-progress work, where we investigate the goals and needs of our target group. We also test the applicability of the ST framework proposed by Ehsan et al. to the IT-security domain and its potential for facilitating greater trust in human-AI teams in system administration. Our first results show that ST can indeed yield benefits for sysadmins but only when coupled with other contextual information available in the system and only when it adheres to specific quality standards in terms of content it provides. As we continue our analysis, we aim to compile a set of desirable user requirements for a successful AI system adoption and support of human-AI teams in the network security domain.

CCS Concepts: • **Networks → Network security**; • **Human-centered computing → Collaborative and social computing theory, concepts and paradigms**; Scenario-based design; User centered design; **HCI theory, concepts and models**; Empirical studies in HCI; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: network monitoring, artificial intelligence, automation, explainable AI, social transparency, system administrators, trust

ACM Reference Format:
Daria Soroko, Gian-Luca Savino, and Nicholas Gray. 2018. How Can AI Earn Trust of System Administrators in the IT-Security Domain?. In . ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION & MOTIVATION

With an increasing number of services moving online, computer networks are growing in size and complexity. A larger variety of devices and applications have to be supported and the bandwidth demands rise [6, 27]. As a result, operating, monitoring and securing such large networks becomes a challenge for system administrators (sysadmins) and network operators (treated here as a subcategory of sysadmins). A central and often difficult task for sysadmins is to have a full understanding of the network. For this, the required information is commonly gathered and visualised by a Network

Management System (NMS). Yet, usability issues in existing software solutions and growing demands for broader knowledge as well as skills in the domain add another layer of complexity that is challenging to contend with even for seasoned professionals [7]. In addition, the passing of knowledge and experience from competent experts to novices is not commonplace and depends on the enterprise culture of an individual company. Naturally, there comes the need for a system solution that could address these user needs: 1) Assist in securing large networks efficiently, 2) provide greater usability and user experience to sysadmins, 3) support knowledge acquisition by the admins to resolve issues in the network (both from external resources and from experienced peers).

Intelligent user interfaces help to fulfill some of the stated needs, by automating parts of the sysadmins' workflow. The adoption of such systems has been gaining traction in the field [4, 5, 17, 21, 28], with more companies choosing to invest into advanced AI-based software solutions [9]. One of the common examples of this type of automation in the network monitoring and security domain is the so-called next-generation firewall (NGFW) [1, 8, 13]. Such products focus on automatic threat detection through the analysis of devices, users and their traffic, and prioritise security alerts to allow sysadmins to efficiently resolve conflicts and breaches within the system.

However, at a closer inspection such AI-based solutions often fall short in addressing the user needs identified earlier. This is mainly due to the lack of transparency of the system's decision-making processes that is sacrificed for the sake of convenience and efficiency. Consequently, the overall user experience of such systems becomes compromised. A great number of AI algorithms, especially those based on neural networks [5, 11, 25], suffer from opacity of their processes, making it difficult to understand how a system achieved its results [3, 14, 19, 22, 29]. When applied to computer networks, this lack of transparency and understanding may lead to a decreased trust in the system's output [15, 16, 18, 20, 23, 30], especially in high-stakes scenarios when the security of the system may be compromised, and the sysadmin in charge of the network is likely to bear the brunt of responsibility for any losses associated with the breach. The system's opacity also takes away the admin's control over their network, as they can no longer understand and navigate it. As a result, the users might fall into one of the two extremes: Either distrust AI completely, choose not to be responsible for its decisions and refuse to adopt the system entirely; Or choose to over-rely on the system, especially when overwhelmed by the number of tasks at hand. Both extremes come to the detriment of network security. The lack of understanding of the system contributes to poor usability. The lack of transparency about the system and its output hinders knowledge acquisition by sysadmins, and consequently its distribution.

2 RELATED WORK

An emerging discipline of Explainable AI (XAI) aims to address some of the transparency and trust concerns. And while XAI helps to investigate the more algorithm-centred side of the human-AI interaction, it falls short of including contextual information such as the socio-organisational context that is usually not included in AI models [12]. Social context in the form of peer support is an important resource for system administrators even outside of the AI-mediated scenarios as it aids the admins in decision-making when their own knowledge and experience are not sufficient [7]. However, AI automation in its current form adds a new layer of uncertainty and opacity. It makes the availability of social context imperative to their ability to make decisions in novel situations, to provide opportunities for sharing knowledge among peers and a sense of shared responsibility.

The way the system administrators usually engage with social context in daily work is through what is known as social navigation. Social navigation is a topic in social computing that takes the natural human tendency to follow other people's cues when feeling lost [7, 10]. One could think of it as a form of collective intelligence. The concept was originally inspired by human behaviour in the physical world and applied to information spaces on the web. However, the term has evolved to not exclusively refer to the act of actual navigation but to also signify a type of decision-support mechanism [10]. In practice, for sysadmins this often means consulting colleagues and online forums, such as Stack Overflow, to find remedies and solutions to their technical issues when facing these for the first time.

We see social transparency (ST) as the latest iteration of social navigation that sets its focus specifically on AI-aided decision-making. It is a framework proposed by Ehsan et al. [12] that focuses on using the external contextual knowledge about one's peers' interactions with AI to increase the explainability of an AI-based system and to aid the user in making their own decisions. The approach is meant to address, among other things, the issues of trust and lack of understanding in the AI-supported decision-making processes, provide means to incorporate and consolidate external human knowledge otherwise not included in the AI model and share it among the peers. The latter could provide great benefits for system administrators in particular. In the context of network monitoring and security, ST also has the potential to facilitate the process of triangulation when the admin is unsure about the system's outcome by providing peer information. This way the admin will be able to assess the trustworthiness of the system based on how his colleagues interacted with it in the past. This makes it a promising strategy that has the potential to solve the issues associated with the incorporation of AI into the workflow of system administrators in automated and semi-automated systems and to fulfill most of the user needs introduced earlier.

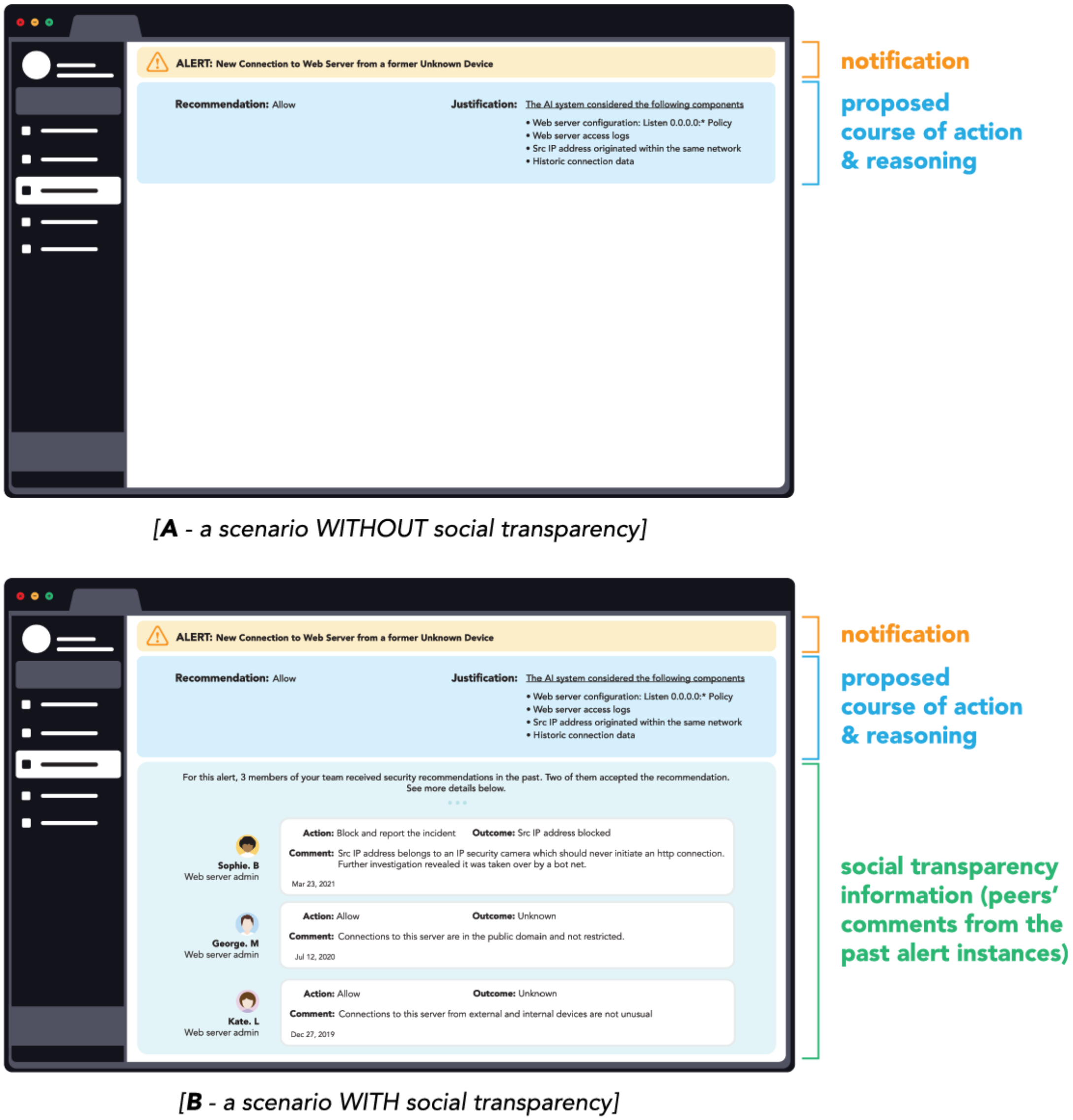
In this paper, we present a brief overview of our research work where we applied the social transparency framework to the domain of system administration. Our research questions and goals focused on:

- Testing the applicability of the concept of social transparency to the IT-security domain.
- Investigating the tools that sysadmins currently use to gather contextual information when solving new problems.
- Developing initial guidelines for incorporating social transparency into existing workflows of system administrators to facilitate usability of AI-based systems and ameliorating some of the issues associated with this type of automation.
- Expanding the original study from a methodological point of view by testing and validating an existing approach to quantitatively measuring trust in the context of speculative design studies and AI-mediated decision-making processes.

3 INITIAL STUDY

To answer our research questions, we conducted a mixed-methods study. Our approach was inspired by Ehsan et al. [12] and adopted their speculative design walk-through to the domain of network security and monitoring. However, we extended the method and incorporated a questionnaire developed by Ashoori and Weisz [2] to quantitatively measure trust. The original study by Ehsan et al. excluded the trust variable and only measured confidence. We wanted to address this omission. Our choice of the questionnaire was dictated by the fact that it was designed to evaluate trust in AI-infused decision-making scenarios, with a focus on "people's attitudes toward hypothetical situations", rather than specific behaviours. And since we were interested in investigating the sysadmins' attitudes towards ST as applied to IT-security, the intended purpose of the questionnaire suited our specific use case. It must be noted, however, that we excluded two questions in our study design as they did not apply to our context.

We recruited 12 IT-security experts with varying levels of expertise and experience in the domain. Experiments were conducted via the Zoom video conferencing software, lasted about 45 minutes and were video-recorded with the participants' consent for further analysis. Using a fictive task, sysadmins were confronted with two scenarios where an



ALERT: New Connection to Web Server from a former Unknown Device

Recommendation: Allow

Justification: The AI system considered the following components

- Web server configuration: Listen 0.0.0.0:* Policy
- Web server access logs
- Src IP address originated within the same network
- Historic connection data

[**A** - a scenario WITHOUT social transparency]

ALERT: New Connection to Web Server from a former Unknown Device

Recommendation: Allow

Justification: The AI system considered the following components

- Web server configuration: Listen 0.0.0.0:* Policy
- Web server access logs
- Src IP address originated within the same network
- Historic connection data

For this alert, 3 members of your team received security recommendations in the past. Two of them accepted the recommendation. See more details below.

Sophie. B
Web server admin

Action: Block and report the incident Outcome: Src IP address blocked
Comment: Src IP address belongs to an IP security camera which should never initiate an http connection. Further investigation revealed it was taken over by a bot net.
Mar 23, 2021

George. M
Web server admin

Action: Allow Outcome: Unknown
Comment: Connections to this server are in the public domain and not restricted.
Jul 12, 2020

Kate. L
Web server admin

Action: Allow Outcome: Unknown
Comment: Connections to this server from external and internal devices are not unusual
Dec 27, 2019

notification
proposed course of action & reasoning
social transparency information (peers' comments from the past alert instances)

[**B** - a scenario WITH social transparency]

Fig. 1. The UI mock-ups of the two scenarios used in the study.

AI-aided system detected an anomaly in the network and recommended a course of action. Through screen share the experimenter presented the slides that depicted a mock-up interface as shown on Figure 1.

In both scenarios, the system displayed a "new connection" alert and proposed to continue allowing access from the web server to an unknown device. It also provided justification for its decision, i.e. information it took into account. However, in one of the scenarios, the system also included the information about the way the participant's fictional peers, other system administrators, in the company reacted in a similar context in the past. The peers' comments in this case are an example of social transparency.

Having familiarized themselves with each situation, participants were asked whether they agreed with the systems' decision or if they would override it and block the access to the unknown device. Then, the participants were asked to fill out the questionnaire by Ashoori and Weisz [2] to assess their trust in the overall decision-making process in each scenario. At the end of study, we conducted short semi-structured interviews to ask participants about their thoughts on the framework and presented scenarios. We also inquired about the way the interviewees went about finding solutions to the problems they struggled with at work and the resources they tended to reach to the most.

Although the analysis is still ongoing, below we present some of the first impressions and emerging topics in our research.

4 FIRST RESULTS

4.1 General scepticism towards AI

As we move on with our analysis, it becomes apparent that sysadmins tend to be suspicious and cautious when it comes to using AI in the daily work and choose to not "blindly trust AI". This in part can be explained by the nature of training they receive. In particular, concepts such as Principle of Least Privilege [26] and Zero-Trust [24] contribute to the initial hesitation in adopting AI-based solutions in network monitoring and security. Consequently, admins prefer to observe the system's performance over time to develop their own opinion on it. Consistency of the system's decision and their quality influence the admin's assessment.

4.2 The more context the better

Another common thread we observe in the qualitative interview data is the admins' need for as much contextual information as possible in any given case. This may include:

- Technical information about devices in the network,
- System specifications,
- Information about the AI algorithm and data it was trained on,
- Data parameters that influenced the system's outcome (e.g. confidence score),
- Historical information about the system's performance.

Access to such information is meant to provide enough knowledge to the admin to be able to make a decision in a challenging situation, especially when they are unsure about the course of action suggested by the system.

4.3 Social transparency can be useful and dangerous

The participants' response to the concept of social transparency although largely positive has not be uniform. Some found it helpful, especially in situations where the information provided by the system was limited. Others thought it unnecessary or even dangerous, as they expected some of their colleagues to blindly follow the recommendation of another person without double checking its sensibility. In many cases, the study participants brought our attention to the need for certain quality standards for the peer information to adhere to. And although deemed useful in most cases, the participants stressed the importance of having access to technical information about the system mentioned above in addition to ST peer information.

5 CONCLUSION AND FUTURE WORK

This workshop paper presents our ongoing investigation of system administrators as target users in the context of (semi-)automated systems based on AI. We examine the admins' goals, needs, and pain points using qualitative analysis methods. We also test the applicability of the social transparency framework to the domain of network monitoring and security as a possible way of addressing some of the users' concerns, especially those pertaining to trust in AI-based systems. Our initial results show that while social transparency has a great potential for assisting sysadmins in their daily tasks, this type of intervention alone cannot solve all trust and usability issues in the IT-security domain associated with automation. Trust is a multifaceted issue and has many contributing aspects. The common sentiment we gather from our respondents is that AI has to earn the user's trust. Furthermore, although not exhaustive on its own, social transparency can yield benefits in combination with other supporting factors to provide greater control over the system to the users and facilitate their decision-making processes.

When it comes to the speculative format of our study, it provided an opportunity for us to collect a broader swath of information about our target audience and their needs. The less-technical UI mockup sparked fruitful discussions with the participants and prompted them to explicitly mention the type of information they would expect and would wish to see included in similar systems in real-life. At the same time, we recognize the limitations of our approach and plan to conduct a follow-up study with existing software (if available) or a high-fidelity prototype of such a system. Recruiting participants that already use AI-based software in their daily work is another challenge we would like to address in the future.

Naturally, we constantly test and revise our initial hypotheses and assumptions as we continue our analysis. We are looking forward to sharing the final results with the broader research community. In the meantime, we hope that this brief introduction to our work will spark conversation and facilitate knowledge exchange with our colleagues.

REFERENCES

[1] 2020. Vectra AI. https://www.vectra.ai/

[2] M. Ashoori and Justin D. Weisz. 2019. In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. *ArXiv* abs/1912.02675 (2019).

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[4] Kirk Bresniker, Ada Gavrilovska, James Holt, Dejan Milojicic, and Trung Tran. 2019. Grand Challenge: Applying Artificial Intelligence and Machine Learning to Cybersecurity. *Computer* 52, 12 (2019), 45–52. https://doi.org/10.1109/MC.2019.2942584

[5] Leong Chan, Ian Morgan, Hayden Simon, Fares Alshabanat, Devin Ober, James Gentry, David Min, and Renzhi Cao. 2019. Survey of AI in Cybersecurity for Information Technology Management. In *2019 IEEE Technology Engineering Management Conference (TEMSCON).* 1–8. https://doi.org/10.1109/TEMSCON.2019.8813605

[6] Checkpoint. 2020. *Security Report 2020: Checkpoint.* https://www.checkpoint.com/downloads/resources/cyber-security-report-2020.pdf

[7] S. Chiasson, R. Biddle, and Anil Somayaji. 2007. Even Experts Deserve Usable Security: Design guidelines for security management systems.

[8] Cybereason. 2021. empow. https://empow.co/

[9] Webroot Smarter Cybersecurity. 2019. Knowledge Gaps: AI and machine learning in cybersecurity. Perspectives from U.S. and Japanese IT Professionals. (2019). https://www-cdn.webroot.com/6015/4999/4566/Webroot_AI_ML_Survey_US-2019.pdf

[10] A. Dieberger, P. Dourish, K. Höök, P. Resnick, and A. Wexelblat. 2000. Social Navigation: Techniques for Building More Usable Systems. *Interactions* 7, 6 (Nov. 2000), 36–45. https://doi.org/10.1145/352580.352587

[11] Selma Dilek, Hüseyin Cakır, and Mustafa Aydın. 2015. Applications of Artificial Intelligence Techniques to Combating Cyber Crimes: A Review. *International Journal of Artificial Intelligence & Applications* 6, 1 (Jan 2015), 21–39. https://doi.org/10.5121/ijaia.2015.6102

[12] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. https://doi.org/10.1145/3411764.3445188

[13] Genua. 2021. cognitix Threat Defender. https://www.genua.de/en/it-security-solutions/cognitix-threat-defender

[14] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 42 pages. https://doi.org/10.1145/3236009

[15] Kevin Anthony Hoff and Bashir Masooda. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570 arXiv:https://doi.org/10.1177/0018720814547570 PMID: 25875432.

[16] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[17] Faouzi Kamoun, Farkhund Iqbal, Mohamed Amir Esseghir, and Thar Baker. 2020. AI and machine learning: A mixed blessing for cybersecurity. In *2020 International Symposium on Networks, Computers and Communications (ISNCC)*. 1–7. https://doi.org/10.1109/ISNCC49221.2020.9297323

[18] René F. Kizilcec. 2016. *How Much Information? Effects of Transparency on Trust in an Algorithmic Interface.* Association for Computing Machinery, New York, NY, USA, 2390–2395. https://doi.org/10.1145/2858036.2858402

[19] Jiwei Li, Xinlei Chen, E. Hovy, and Dan Jurafsky. 2016. Visualizing and Understanding Neural Models in NLP. In *HLT-NAACL*.

[20] Tim Miller. 2019. "But Why?" Understanding Explainable Artificial Intelligence. *XRDS* 25, 3 (April 2019), 20–25. https://doi.org/10.1145/3313107

[21] Mauro José Pappaterra and Francesco Flammini. 2019. A Review of Intelligent Cybersecurity with Bayesian Networks. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 445–452. https://doi.org/10.1109/SMC.2019.8913864

[22] Zhuwei Qin, Fuxun Yu, Chenchen Liu, and Xiang Chen. 2018. How convolutional neural network see the world - A survey of convolutional neural network visualization methods. arXiv:1804.11191 [cs.CV]

[23] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 97–101. https://doi.org/10.18653/v1/N16-3020

[24] Scott W Rose, Oliver Borchert, Stuart Mitchell, and Sean Connelly. 2020. Zero trust architecture. (2020).

[25] I. H. Sarker, Md. Hasan Furhad, and Raza Nowrozy. 2021. AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *SN Comput. Sci.* 2 (2021), 173.

[26] Fred B Schneider. 2003. Least privilege and more [computer security]. *IEEE Security & Privacy* 1, 5 (2003), 55–59.

[27] Sophos. 2021. *Sophos 2021 Threat Report.* https://www.sophos.com/en-us/medialibrary/PDFs/technical-papers/sophos-2021-threat-report.pdf

[28] Tim Stevens. 2020. Knowledge in the grey zone: AI and cybersecurity. *Digital War* 1 (2020), 164–170. https://doi.org/10.1057/s42984-020-00007-w

[29] Francesco Ventura, Tania Cerquitelli, and Francesco Giacalone. 2018. Black-Box Model Explained Through an Assessment of Its Interpretable Features. In *New Trends in Databases and Information Systems*, András Benczúr, Bernhard Thalheim, Tomáš Horváth, Silvia Chiusano, Tania Cerquitelli, Csaba Sidló, and Peter Z. Revesz (Eds.). Springer International Publishing, 138–149.

[30] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI.* 408–416.