Auditing Practitioner Judgment for Algorithmic Fairness Implications

Ike Obi
Purdue University
West Lafayette, Indiana, United States
obii@purdue.edu

Colin M. Gray
Purdue University
West Lafayette, Indiana, United States
gray42@purdue.edu

2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS) | 978-1-6654-5713-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ETHICS57328.2023.10154992

***Abstract*—The development of Artificial Intelligence (AI) systems involves a significant level of judgment and decision making on the part of engineers and designers to ensure the safety, robustness, and ethical design of such systems. However, the kinds of judgments that practitioners employ while developing AI platforms are rarely foregrounded or examined to explore areas practitioners might need ethical support. In this short paper, we employ the concept of design judgment to foreground and examine the kinds of sensemaking software engineers use to inform their decisionmaking while developing AI systems. Relying on data generated from two exploratory observation studies of student software engineers, we connect the concept of fairness to the foregrounded judgments to implicate their potential algorithmic fairness impacts. Our findings surface some ways in which the design judgment of software engineers could adversely impact the downstream goal of ensuring fairness in AI systems. We discuss the implications of these findings in fostering positive innovation and enhancing fairness in AI systems, drawing attention to the need to provide ethical guidance, support, or intervention to practitioners as they engage in situated and contextual judgments while developing AI systems.**

***Index Terms*—Artificial Intelligence, Design Judgment, Fairness, Algorithms, Ethics**

I. Introduction

Artificial Intelligence (AI) systems are increasingly being employed for the distribution of consequential goods like access to healthcare [7], education [4], and criminal justice [34]. Scholars have increasingly raised concerns about the fairness implications of those systems, critiquing the increased automation of difficult-to-model sociopolitical services [27]. Although various researchers have proposed techniques for supporting practitioners to develop ethical AI systems and to ameliorate concerns and instances of bias and unfairness, much of the support has typically been focused on developing checklists [21], toolkits [1], explainable models [35], [36], and auditing mechanisms [28]. In comparison, very limited research and attention has been devoted to describing the complex decisions and trade-offs engineers engage in as they select data sets, train models, and create algorithms for the development of data-driven platforms—a type of complexity that we address in this short paper through the lens of *design judgment*.

This work is funded in part by the National Science Foundation under Grant No. 1909714.

978-1-6654-5713-2/23/$31.00 ©2023 IEEE

In this paper, we report findings from the observation of student software engineers as they engage in conversation while developing an AI image and speech recognition system. Through a discourse analysis [3], we examine the relationship between the discourse of the engineers, the implied design judgments within the discourse, and the potential fairness implication of those judgments. This analysis involved employing an existing framework of design judgment types [11], [26] and different concepts of fairness [14], [19], [30], [33] to foreground and deconstruct the different manifestations of design judgment the software engineers exhibited as they explored ways of building an AI system. By design judgment, we refer to the invocation of practical knowledge and sensemaking that is continuously surfaced through engagement by a technologist or designer with their work. Design judgments are subjective, contextual, and involve engagement with a repertoire of knowledge, or "knowing based on knowledge that is inseparable from the knower" [25], thus providing a sensemaking frame which informs specific and observable design and engineering decisions. Considering that AI platforms are socio-technical systems that are embedded within cultural and social contexts and have the potential to alter norms and cause harm, we use these judgments as a way to evaluate manifestations and instances of the encoding of different forms of fairness as justice in the development of AI systems and algorithms.

We approach this analysis not as a pretext for nudging software developers to learn about design judgment or to explicitly integrate design judgment into the software development process. Rather, we employ design judgment as an analytic lens to surface areas where engineers may need support and guidance towards the goal of developing fair and ethical data-driven systems that empowers their host communities. We further employ design judgment in conjunction with the concepts of fairness as justice to imply that the judgments engineers rely upon are micro enactments of value-and justice inscription which has downstream implications of affecting the fairness of the data-driven systems that emerge from such a process.

We aim to make two contributions in this short paper. First, we introduce design judgment as an additional framework for auditing practitioner judgment for fairness implications during the development process of AI systems. Second, we connect practitioner judgments to related fairness and justice implications, thereby foregrounding the potential material effect of the

Authorized licensed use limited to: Indiana University. Downloaded on July 07,2023 at 14:11:21 UTC from IEEE Xplore. Restrictions apply.

judgement of engineers. In the following sections, we draw on existing literature to explore the role of practitioners in developing ethical systems at the intersection of fairness and justice. We then describe our research methods and analysis strategies and present findings generated from those methods. We conclude with a discussion of the implications of our findings and draw attention to the need to provide ethical support to practitioners while developing AI and data-driven systems.

II. Background

A. Role of the Practitioner in Developing Ethical AI Systems

Technology practitioners, including engineers, data scientists, and designers, play a crucial in ensuring the design of ethical AI systems. Martin [22] highlighted that making mistakes while developing software systems and algorithms is unethical and recommended that software engineers should bear responsibility for any harms that arise from algorithms or software systems developed by them. Bostrom and Yudkowsky [2] examined the implications of value-driven algorithms recommending that such algorithms should be open to scrutiny to ensure that they do not cause harms. Kraemer et al. [18] argue that algorithm design is a value inscription activity and that two different practitioners with different values will design the same system and algorithm differently. Verbeek [32] also described how technology is increasingly playing an invisible role in mediating human actions through values that are deployed and inscribed in designed artifacts, where socio-technical practitioners do "ethics by other means." Mittelstadt and Floridi [24] explored ethical issues relating to big data, highlighting five main issues, including concerns about privacy and informed consent, among others. Mittelstadt [23] further highlighted that most of the existing principles of ethical AI might not succeed in curtailing harms caused by algorithms due to lack of coherence among practitioners on a standard normative principles of designing such systems.

Although most of these studies have targeted the improvement of ethics in data-driven and algorithmic systems, there are as of yet, very few studies that intimately engage with software engineers—the main actors responsible for potential harms—to describe their decision-making process and investigate how their complex and layered judgments might lead to unintended' harmful consequences such as bias and unfairness. Gray and Chivukula [10] have highlighted the ethical complexities that socio-technical practitioners face in their everyday work, highlighting that practitioner judgment is mediated by multiple factors, including organization structures, individual beliefs, and applied ethics. This dearth of studies that has engaged with practitioners in combination with the continued proliferation of issues of bias and algorithmic harm suggests that there is a need to engage closely with software engineers to better understand their challenges and the kinds of decisions they make while developing artificial intelligence platforms in order to explore ecologically appropriate ways of supporting their work practices.

B. Fairness in Data-Driven systems and Theories of Justice

Fairness is a manifestation of theories of justice [30]. Numerous scholars have explored ways of promoting fairness in the design of algorithmic systems and data-driven platforms. Cobbe et al. [5] introduced the reviewability framework to support the explainability of algorithms. Their goal was to ensure that algorithms are easy to review for areas of improvement or correction. Hampton [13] employed a critical lens to critique oppressive algorithmic systems, highlighting how algorithmic oppression is often invisible and ubiquitous. Herrington [15] suggested a distributive framework for measuring fairness. Implicit in his work is that the harms and benefits of algorithmic systems should be shared more equally between its users. He also highlighted how current distributive models does not take into account historical injustice that has been perpetuated on the protected communities. Numerous other scholars (e.g, [6], [16], [17]) have explored other approaches to improving fairness in data-driven systems.

Furthermore, various scholars have proposed computer ethics education as a way of improving fairness and ethical awareness in technology practice. Grosz et al. [12] explored new ways of integrating ethics into computing education. Despite current efforts, Raji et al. [29] revealed that ethics pedagogy for computer scientists is in a state of crisis and needs to be reviewed and updated. These findings highlight the need to explore new approaches for supporting engineers as they develop data-driven platforms. Building on this prior work in an academic context, in this paper we engaged student software engineers in an observation session to explore the manifestations of judgment in their practice. Next, we analyzed connections between the manifested judgments and different theories of fairness as justice to explore their intersections. Findings from this study will help us to better understand ways of supporting engineers as they develop data-driven platforms.

III. Method

In this research we employed an observation approach [31] to collect data on the kinds of complex and situated judgment software engineers and data scientists make while developing data-driven applications and systems. The session was held with a group of five graduate students at a large, public, research-intensive university in Midwestern USA.

A. Participants

Participants for this exploratory study were recruited via the researcher's professional network using a convenience sampling that met specific inclusion criteria. The inclusion criteria required any of the participants in the session to have a data science or software development experience or background. Particularly, they had to be conversant with either fields of machine learning, computer vision, or natural language processing, or all the listed fields. Hence, anyone that was not experienced in any of the fields was excluded from participating in the study due to the knowledge and experience requirement of the study. In all, we recruited five participants

Authorized licensed use limited to: Indiana University. Downloaded on July 07,2023 at 14:11:21 UTC from IEEE Xplore. Restrictions apply.

for this study. The participants were split into two sessions, with one participant participating in both sessions. Table 1 provides more information about the participants that were recruited for the study. Participants voluntarily participated and were not remunerated for their participation in this study.

TABLE I
STUDY PARTICIPANT INFORMATION

Participant ID	*Degree of Study*	*Years of Experience*
B1	Masters	More than 2 years
B2	Masters	One year four months
B3	PhD	More than 3 years
B4	PhD	2 years and above
B5	PhD	2 years and above

B. Data Collection

We used a lab protocol approach [8], [37] for data collection. Both sessions were framed by a design task that encouraged the participants to converse like they would in a normal team project meetings. The seating arrangement for both sessions was a round table with all participants sitting close to each other as they normally would during their project meetings with their laptop and project materials. The participants were briefed on the goal of the session before the session commenced and were notified the session would be recorded for analysis. All participants consented to continue their participation in the study under an approved study protocol from the Institutional Review Board. Each sessions was 60 minutes in duration.

During the first session, the participants were asked to develop AI-generated content that imitated a popular personality. During the second session, participants were asked to develop a speech recognition system that detected how long individual participants spoke during meeting and how often they were interrupted. Audio recordings were the primary source of data, and the primary researcher also generated field notes and memos during the session.

C. Data Analysis

Data for this study were analyzed using discourse analysis [9]. First, we generated a transcript from the observation using an online transcription service. Next, we loaded the transcript data into a qualitative analysis tool for data analysis and thoroughly cleaned the transcript. We then commenced coding the dataset for instances of design judgment using the framework provided by Nelson and Stolterman [26]. Next, we commenced a second round of open coding of the highlighted design judgments manifestations to connect to the concept of fairness. Table II provides information the design judgment framework. In the next section, we interpret findings from the observations section and its implications to goal of improving fairness and ethics in data-driven systems.

IV. FINDINGS

Findings from our analysis revealed that the participants applied different types of *design judgment* in support of reaching decisions regarding three key project objectives, including decisions about technical requirements and tools, decisions about the nature and essence of the product, and decisions around user-product alignment. We present relevant judgment expressed in relation to these decisions below and highlight opportunities for fairness intervention where salient.

A. Decisions about technical requirements and tools

Our analysis revealed that the participants applied a range of judgments that informed the kinds of tools and technical requirements they needed to execute the project.

1) Navigational, Framing, and Connective Judgments: The participants heavily relied on a clustering of navigational and framing judgements to determine how and what to bring into the development of the deepfake AI content. They engaged in these conversations to frame the project and set themselves on the right project path and towards producing a real system that would work. For example, participant B3 questioned B2 about potential implementation approaches, asking: *"so how do you train the tone and voice of the person that is to be imitated?"* B2 responded by suggesting that they should consider *"train[ing] it with transcript from existing videos of the person being targeted."* B3 then responded with another question using a connective judgment to link another requirement to the same goal, asking: *"what about the voice?"* B2 followed up and responded that *"voice generation is a different kind of thing,"* then after a quick back and forth B2 conceded that maybe that they can use *"GANS to generate audio,"* mentioning that *"they can learn the wavelength of somebody's voice data by synthesizing it with GANS."* Notably, at no point did any of the participants question the ethicality of the goals they were assigned to complete revealing opportunities for fairness intervention in real-life contexts.

2) Instrumental Judgments: After the initial navigational, framing, and connective judgments, the participants transitioned to primarily making instrumental judgments that enabled their exploration of how they would implement the project goals. B2 started by implicitly deciding that their first step would be select a model, noting: *"so the LSTM's would generate like a script for the GAN's and the GAN's says it out in that voice and then we need the image to sync with that."* B3 responded by stating *"the first step definitely would be training with the GAN."* B1 agreed and mentioned that they would need to get *"data of the target presenting their speech, and I don't know if there is a library or data set for that or if there is a dataset available."* B3 followed up, questioning whether *"get[ting] the data is the hardest part; do we get text, or do we get videos or video transcripts?"*. Participant B1 also mentioned the need for an additional tool highlighting that *"the next part would be the hardware to train it. You really need like GPUs to train it and GPUs are very hard to find and I do not have access to train it."* Altogether, these discussions focused on the limitations and opportunities afforded by potential tools or procedures typical of instrumental judgments. These exchanges highlight an opportunity for intervention, because if the practitioners do not follow best

Authorized licensed use limited to: Indiana University. Downloaded on July 07,2023 at 14:11:21 UTC from IEEE Xplore. Restrictions apply.

TABLE II
OPERATIONALIZED DEFINITIONS OF DESIGN JUDGMENT TYPES, QUOTED FROM GRAY ET AL. [11] AND NELSON & STOLTERMAN [26].

	Design Judgment Type	Operationalized Definition
1	Core Judgment	Statement about one's value thinking, usually revealed when pushed by "why" questions concerning one's judgment.
2	Appreciative Judgment	Placing high value and emphasis on certain aspects of a design situation while backgrounding, or lessening focus on others.
3	Instrumental Judgment	The selection utilization, or influence of a tool, concept, or method in reaching an established design goal.
4	Navigational Judgment	Considering a path, plan, or certain manner (of individual, disciplined preference) in approaching a task or a challenge to get a desired state
5	Quality Judgment	Making design decisions about the effectiveness of visual and other forms of style, or to demonstrate due diligence, often in accordance with company standards, in relation to a concrete design artifact.
6	Framing Judgment	Creating a working area for design activity to occur, often by introducing constraints (client or tool) or ways of assessing outcomes. This occurs dynamically across multiple levels.
7	Appearance Judgment	Assessment of overall quality, relating to an entire product or experience, rather than just a portion. This often includes part/whole relations within a frame of aesthetic experience or measurement against heuristic(s).
8	Connective Judgment	Making connections or bringing various design objects that are central to the design process and activity. The connections made in this context are not generalized but specific to a particular design situation .
9	Offhand Judgment	Recalling to consciousness previous judgments that have led to successful practices and opening them up the possibility of adaptation or use.
10	Compositional Judgment	Making connections or bringing various design objects together that are central to the design process and activity. The connections made in this context are generalized and not specific to a particular design situation but to the overall process.
11	Default Judgment	Giving an automatic response to a situation without deliberation.

practices and processes using known techniques, it might lead to outcomes that cause harm to users down the line.

B. Decisions about the essence and nature of the product

1) Appreciative Judgments: After connecting different resources and determining appropriate approaches to complete the task, the participants transitioned to making primarily appreciative judgments to ensure that they selected the right dataset(s) that would help them to develop a system that matched real-world expectations and their goals. Participant B2 asked: *"[how] do we get specific dataset to make it more accurate or a variety of data sets to make it generalized so that we can make it usable for different types of faces?"* Implicit in this position is B2's belief that datasets play a role in the final appearance of the product, thus, shaping the kinds of outcomes that can result—an appreciative stance with implied values. B3 supported this position and added that *"[the] dataset is the most important thing because that's what you need to train, and training is the most important part and to train you need data set. So, the more you train, the better the stuff gets."* The participants' discussions relating to appreciative judgments focused on the challenges of getting the necessary data set and the problems that could be inscribed into certain datasets revealing potential for fairness interventions.

2) Core Judgments: After deliberations on appearance, the participants intuitively leveraged core judgments as a means of reflecting and questioning some of their approaches seeking to interrogate value alignment of the project with their beliefs. For example, with respect to the question of collecting datasets, B2 highlighted the issue of ethics as follows: *"there's also like ethicality in whose face can you use? Like, do you ask them to use their face? Cos airports would definitely have data on faces, that they scan people, they have surveillance all the time. But can they just use that surveillance without asking?"* B1 interjected and recommended that they can use an AI-generated face to sidestep these concerns. However, B2 responded by asking *like, is it ethical to generate faces that might exist? We do not know if that face belongs to someone, but we did not get it from them. We generated it, you know. Yeah."* This interaction between the participants revealed an opportunity for fairness intervention to mitigate situations that could lead to outcomes during development.

C. Decisions about User-Product Alignment

1) Quality Judgments: Towards the end of the session, the participants started focusing on evaluating the potential quality of the product to be designed based on the judgments and decisions they had made so far. B2 mentioned that *"there will be a graph that tells a loss, which is like the difference between what you expect and what our model is, We're trying to minimize the loss because lower loss means higher accuracy."* Altogether, as a group, they described their success metric as developing a system that is widely valuable to its users, while also very accurate, pointing towards an overarching quality judgment.

V. DISCUSSION

Developing an AI system currently requires a high-level of knowledge about statistical approaches, compute resources, and datasets. This significant level of knowledge entails that engineers and designers ought to exert a considerable amount of judgment on how best to appropriate these resources and navigate the space towards developing systems that empower users and not expose them to harms. Despite this need to guide practitioners, most of the current mechanisms and resources designed to support engineers often focus on the knowledge, tools, and compute requirement—perhaps avoiding or marginalizing the kinds of situated and contextual judgments practitioners engage in while they develop data-driven systems. These judgments, as a crucial form of sensemaking that occurs

Authorized licensed use limited to: Indiana University. Downloaded on July 07,2023 at 14:11:21 UTC from IEEE Xplore. Restrictions apply.

prior to decision making, are infrequently studied and hence there is little understanding on how to provide support in an ecologically meaningful way.

Consequently, in this study, we employed the theoretical framework of design judgment to foreground and deconstruct the kinds of judgment that the participants applied in their everyday practice when developing data-driven platforms. Our findings revealed that practitioners engaged in a range of judgment types, including: instrumental, appreciative, navigational, quality, and core judgments. Viewing the interplay of these judgments allowed us to see the goals of the participants from a conceptual level and the potential fairness implications of those judgments if no intervention is provided to support them. Findings from this study also sensitize engineering educators to nudge their students to interrogate their design judgments during product development as a means of ensuring ethical outcomes.

VI. Conclusion

In this short paper, we observed student software engineers to explore the different kinds of judgment they employed as they developed an AI product. Findings from our analysis revealed that participants engaged in a range of different types of judgment that have implications for considering the fairness of the resulting designed system. We connected these judgment types to fairness implications to highlight the potential role supporting practitioner judgment might have in increasing the fairness of AI systems. Our findings contribute to an initial understanding of the utility of design judgment in deconstructing and making accessible the decision-making processes of software engineers towards developing fair and robust AI systems.

References

[1] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," 2018. https://arxiv.org/abs/1810.01943.

[2] N. Bostrom and E. Yudkowsky, "The Ethics of Artificial Intelligence ," Cambridge Handbook of Artificial Intelligence, Jul-2018. [Online]. Available: https://csiflabs.cs.ucdavis.edu/ ssdavis/188/EthicsofAI.pdf.

[3] G. Brown and G. Yule, Discourse analysis. Cambridge University Press, 2012.

[4] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in education," IEEE Access. 2020. https://ieeexplore.ieee.org/abstract/document/9069875.

[5] J. Cobbe, M. S. Ah Lee, and J. Singh, "Reviewable automated decision-making: Proceedings of the 2021 ACM Conference on Fairness, accountability, and transparency," 2021. https://dl.acm.org/doi/10.1145/3442188.3445921.

[6] K. Creel and D. Hellman, "The Algorithmic Leviathan," 2021. Available: https://dl.acm.org/doi/abs/10.1145/3442188.3445942.

[7] T. Davenport and Ravi Kalakota, "The potential for artificial intelligence in Healthcare," Future healthcare journal, 2019. https://pubmed.ncbi.nlm.nih.gov/31363513/.

[8] C. S. Gray, S. S. Chivukula, K. Melkey, and R. Manocha, "Understanding 'Dark' Design Roles in Computing Education," International Computing Education Research Workshop, Aug. 2021, doi: 10.1145/3446871.3469754.

[9] J. P. Gee. How to Do Discourse Analysis: A Toolkit. Routledge. 2014..

[10] C. M. Gray and S. S. Chivukula. Ethical Mediation in UX Practice. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–11. 2019.

[11] C. M. Gray, C. Dagli, M. Demiral-Uzan, F. Ergulec, V. Tan, A. A. Altuwaijri, K. Gyabak, M. Hilligoss, R. Kizilboga, K. Tomita, E. Boling. Judgment and Instructional Design: How ID Practitioners Work in Practice. Performance Improvement Quarterly 28, 3. pp. 25–49. 2015.

[12] B. J. Grosz, D. G. Grant, K. Vredenburgh, J. Behrends, L. Hu, A. Simmons, and J. Waldo. Embedded EthiCS: Integrating Ethics Across CS Education. Commun. ACM 62, 8. pp. 54–61. 2019.

[13] L. M. Hampton. Black Feminist Musings on Algorithmic Oppression. arXiv preprint. arXiv:2101.09869. 2021.

[14] J. W Harris. Property and Justice. OUP Oxford. 1996.

[15] J. Herington. Measuring Fairness in An Unfair World. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 286–292. 2020.

[16] S. Kacianka and A. Pretschner. Designing Accountable Systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 424–437. 2021.

[17] M. Kasy and R. Abebe. Fairness, Equality, and Power in Algorithmic Decision-making. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 576–586. 2021.

[18] F. Kraemer, K. V. Overveld, and M. Peterson. Is There An Ethics of Algorithms? Ethics and Information Technology 13, 3. pp. 251–260. 2011.

[19] J. Llewellyn and R. L Howse. Restorative Justice: A Conceptual Framework. Prepared for the Law Commission of Canada. 1999.

[20] S. L. Piano. Ethical Principles in Machine learning and Artificial Intelligence: Cases from the Field and Possible Ways Forward. Humanities and Social Sciences Communications 7, 1. pp. 1–7. 2020.

[21] M. A Madaio, L. Stark, J. W. Vaughan, and H. Wallach. Co-designing Checklists to Understand Organizational Challenges and Opportunities Around Fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–14. 2020.

[22] K. E. Martin. Designing Ethical Algorithms. MIS Quarterly Executive June. 2019.

[23] B. Mittelstadt. Principles Alone Cannot Guarantee Ethical AI. Nature Machine Intelligence 1, 11. pp. 501–507. 2019.

[24] B. D. Mittelstadt and L. Floridi. 2016. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. The Ethics of Biomedical Big Data. pp. 445–480. 2016.

[25] H. G. Nelson and E. Stolterman. The Design Way : Intentional Change in an Unpredictable World (2nd ed.). MIT Press, Cambridge, MA. 2020.

[26] H. G Nelson and E. Stolterman. The Design Way: Intentional Change in An Unpredictable World. MIT press. 2014.

[27] L. Nordling. 2019. A Fairer Way Forward for AI in Health Care. Nature 573, 7775, S103–S103. 2019.

[28] I. D. Raji and J. Buolamwini. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.pp. 429–435. 2019.

[29] I. D. Raji, M. K. Scheuerman, and R. Amironesei. You Can't Sit with Us: Exclusionary Pedagogy in AI Ethics Education. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 515–525. 2021

[30] J. Rawls. Justice as Fairness. The philosophical review 67, 2. p. 164–194. 1958.

[31] A. Mulhall, "In the field: notes on observation in qualitative research," Journal of Advanced Nursing, Feb. 01, 2003.

[32] P. Verbeek. Materializing morality: Design Ethics and Technological Mediation. Science, Technology, & Human Values 31, 3. pp. 361–380. 2006.

[33] M. U. Walker. Restorative Justice and Reparations. Journal of social philosophy. 2006.

[34] A. Završnik. Criminal Justice, Artificial Intelligence Systems, and Human Rights. In ERA Forum, Vol. 20. Springer. pp. 567–583. 2020.

[35] S. M. Lundberg and S. I. Lee. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30. 2017

[36] M. T. Ribeiro, S. Singh, and C. Guestrin. " Why Should i Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 1135-1144. 2016.

[37] K. Dorst, "Analysing design activity: New Directions in protocol analysis," Design Studies, vol. 16, no. 2, pp. 139–142, 1995.

Authorized licensed use limited to: Indiana University. Downloaded on July 07,2023 at 14:11:21 UTC from IEEE Xplore. Restrictions apply.