Making Transparency Clear

The Dual Importance of Explainability and Auditability

Aaron Springer
Computer Science
University of California Santa Cruz
Santa Cruz, CA, USA
alspring@ucsc.edu

Steve Whittaker
Psychology
University of California Santa Cruz
Santa Cruz, CA, USA
swhittak@ucsc.edu

ABSTRACT

Algorithmic transparency is currently invoked for two separate purposes: to improve trust in systems and to provide insight into problems like algorithmic bias. Although transparency can help both problems, recent results suggest these goals cannot be accomplished simultaneously by the same transparency implementation. Providing enough information to diagnose algorithmic bias will overwhelm users and lead to poor experiences. On the other hand, scaffolding user mental models with selective transparency will not provide enough information to audit these systems for fairness. This paper argues that if we want to address both problems we must separate two distinct aspects of transparency: explainability and auditability. Explainability improves user experience by facilitating mental model formation and building user trust. It provides users with sufficient information to form accurate mental models of system operation. Auditability is more exhaustive; providing third-parties with the ability to test algorithmic outputs and diagnose biases and unfairness. This conceptual separation provides a path forward for designers to make systems both usable and free from bias.

CCS CONCEPTS

• Human-centered computing~Human computer interaction (HCI)

KEYWORDS

Transparency, trust, explanation, bias, auditability, algorithms, intelligent systems.

ACM Reference format:

Aaron Springer and Steve Whittaker. 2019. Making Transparency Clear: The Dual Importance of Explainability and Auditability. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019, 4 pages.*

1 Introduction

We are at a pivotal time in the use of machine learning as intelligent systems increasingly impact our daily lives. Machine learning algorithms underlie the many intelligent systems we routinely use. These systems provide information ranging from routes to work to recommendations about criminal parole [2,4]. As humans with limited time and attention, we increasingly defer responsibility to these systems with little reflection or oversight. For example, as of February 2018, over 50% of adults in the United States report using a range of voice assistants on a daily basis to accomplish tasks such as navigating to work, answering queries, and automating actions [27]. Improvements to the increasing use of voice assistants are largely driven by improvements in underlying algorithms.

Compounding these advances in machine learning is the fact that many people have difficulty understanding current intelligent systems [38]. Here, we use 'intelligent systems' to mean systems that use machine learned models and/or data derived from user context to make predictions. The machine learning models that often power these intelligent systems are complex and trained upon massive troves of data, making it difficult for even experts to form accurate mental models. For example, many Facebook users did not know that the service curated their newsfeed using machine learning, they simple thought that they saw a feed of all their connections posts [15]. More recently, users of Facebook and other systems have been shown to generate simple "folk theories" that explain how such systems are working [14,38]. Although users cannot validate such folk theories that does not stop users from acting upon them. [14] demonstrated that users went so far as to modify how they interacted with Facebook to try to force the system to present a certain outcome consistent with their user folk theory. There is potential for danger in other contexts when users are willing to act upon their folk hypotheses when not given the ability to understand the system. Furthermore, there are many challenges regarding the best ways to effectively communicate underlying algorithms to users [35,39].

Another concern is the user experience of opaque algorithmic systems. Without any form of transparency, users may trust and understand these systems less [11,24]. Even in low-stakes systems like the Netflix recommender, users still struggle to understand how to control and influence internal algorithms [6]. These problems surrounding user experience, trust especially, become more pronounces in high stakes scenarios such as the medical field where elements of user experience like trust are essential to a program's use.

Furthermore, academics and industry practitioners are discovering other significant issues in deploying these systems. Intelligent systems powered by machine learning can learn and embody societal biases. Systems may therefore treat users differently based on characteristics of users' speech and writing [31,37] or even based upon characteristics that are protected under law [2]. In a particularly egregious example, an intelligent system used to help inform parole decisions was found to discriminate against people of color [2].

Despite these challenges of bias and user experience, many critics have coalesced around a concept they believe could address these challenges: transparency. The insight underlying transparency is that an algorithm should reveal itself to users. There are many important potential benefits for algorithmic transparency. Transparency enables important oversight by system designers. Without transparency it may be unclear whether an algorithm is optimizing the intended behavior [39], or whether an algorithm has negative, unintended consequences (e.g. filter bubbles in social media; [26]). These arguments have led some researchers to argue that machine learning must be 'interpretable by design' [1], and that transparency is even essential for the adoption of intelligent systems, such as in cases of medical diagnoses [40]. Transparency has taken on the role of a cure-all for machine learnings woes.

However, problems remain. Transparency is currently ill-defined [12]. Transparency is purported to address machine learning problems such as bias [25], while simultaneously improving the user experience [18,21]. This paper argues that achieving both goals may be impossible with a single implementation. An implementation of transparency that allows someone to infer system bias will likely overwhelm users and lead to less usage—which in turn will lead to developers refusing to implement transparency. Transparency should be disaggregated into two separate classes: explainability and auditability. Explainability is concerned with building interfaces that promote accurate mental models of system operation leading to a better user experience. Auditability is concerned with allowing users or third-party groups to audit a deployed algorithmic system for bias and other problems. Separating these aspects of transparency allows us to build systems with improved user experiences while maintaining high standards of fairness and unbiased outcomes.

2 Why Do We Need Transparency?

2.1 Poor User Experiences in Intelligent Systems

A wealth of prior work has explored issues surrounding algorithm transparency in the commercial deployments of systems for social media and news curation. Social media feeds are often curated by algorithms that may be invisible to users (e.g., Facebook. Twitter, LinkedIn). Work on algorithmic folk theories shows that making the designs more transparent or seamful, allowed users to better understand and work within the system [14].

Addressing the user experience in intelligent systems has now become a pressing concern for mainstream usability practitioners. The Nielsen Norman group recently completed a diary study examining the user experience of normal people with systems such as Facebook, Instagram, Netflix, and Google News [6]. Mirroring the work on Facebook folk theories, users found it unclear which aspects of their own behavior the intelligent systems used as inputs. Users were also frustrated by the lack of control over the output. Overall, users struggled to form correct mental models of system operation which led to poor user experiences.

Other work shows the importance of transparency for building trust in algorithmic systems, an important part of the user experience. Users who receive explanations better understand and trust complex algorithmic systems [24]. In the presence of disagreement between the system and the user, transparency can improve user perceptions of trust and system accuracy [11,23,34]. But in addition to improving user experience, advocates point to transparency as a counter to more pernicious problems such as algorithmic bias.

2.2 Revealing Bias

Intelligent systems and predictive analytics have been shown to learn and perpetuate societal biases. One clear example of this is COMPAS, an algorithm used widely within the United States to predict risk of recidivism. In 2016 ProPublica published an article noting that the COMPAS system was more likely to predict higher risk scores for people of color than other populations, even when the ground truth was similar [2]. The COMPAS system had been in use for over 5 years in some locations before these biases were publicized [13].

Other work shows how interfaces can discriminate based on ways of speaking and writing. YouTube captions have been shown to be less accurate for speakers with a variety of accents [37]. Common voice interfaces can struggle with specific ways of speaking [31]. These problems likely arise from how algorithms were trained on a non-diverse set of voices (i.e., 'distributional drift'), and then deployed broadly to all people. Even textual methods are not immune to embodying societal biases. Word embeddings have been shown to harbor biases related to gender. For example, one of the roles most closely related to 'she' within the learned word embeddings is "homemaker"; in contrast, an occupation closely related to "he" is "boss" [5].

The fear is that the embodiment of these societal biases within machine learning systems will perpetuate them. For example, biased recidivism algorithms will exacerbate existing inequalities, creating a cycle where those who are not currently privileged will have even less opportunity in the future. An example of this is shown in the posting of job ads online. Men saw significantly more job ads for senior positions compared to women, when searching online [10]. In other cases, African-American names in Google search are more likely to display ads for criminal records, which has been noted as a possible risk for job applicants [36].

It is not simple to fix these problems. Algorithmic bias problems are everywhere; but fixing them requires fitting complex research and auditing practices into iterative agile workflows [32]. This combination requires new tools and extensive organizational buy-in [9]. Even with these processes and tools, not all biases will be found and fixed before a system is deployed.

Transparency has been invoked as a solution to bias. Best-selling books such as Weapons of Math Destruction call for increased transparency as a counter to algorithmic bias [25]. Even the call for papers for this workshop notes that 'algorithmic processes are opaque' and that this can hide issues of algorithmic bias [20]. The idea is that transparency can expose the inner working of an algorithm, allowing users to see whether or not the system is biased. This allows third parties to have the ability to audit the algorithmic systems they are using. However, showing complete algorithmic transparency may have negative impacts on the user experience.

3 Transparency Troubles

Although transparency is an active research area in both machine learning and HCI communities, we believe that a major barrier to current conceptualizations of transparency is the potential negative effects on user experience. Even though a goal of much transparency research is to improve the user experience by building trust, studies are continually showing that transparency has mixed effects on the user experience with intelligent systems.

One system built by our research team clearly reveals problems with our current concept of transparency. The E-meter is an "intelligent" system with an algorithm that assesses the positivity and negativity of a users' writing emotional writing in real time [33]. Users were asked to write about personal emotional experiences and the system interpreted their writing to evaluate how each user felt about their experiences. The E-meter was transparent; it highlighted the words used by the machine learning model conveyed their corresponding emotional weights through a color gradient. The results were unexpected. Users of the transparent system actually felt the system was less accurate overall [34]. Why was this? In some cases, seeing inevitable system errors undermined user confidence, and in other cases, users overrode correct system models that conflicted with their own (inaccurate) beliefs.

Further tests on the E-meter system showed other problems with transparency. Users with a non-transparent version of the E-meter thought that the system performed more accurately [35]. On the other hand, users with transparency seemed to find it distracting. Users of the transparent system were also prone to focus errors exposed by the transparency, even when the overall mood prediction was correct. Clearly, distracting users and leading them to believe the system is more errorful does not create a positive user experience.

Furthermore, users may not want complete transparency for other reasons. Providing such information may be distracting due to the overhead in processing that transparency requires [7]. Transparency negatively affects the user experience in less accurate systems [23]. Short explanations of what a system is doing can improve trust but full transparency can result in less trust in intelligent systems [21].

Together these studies provide strong evidence that exhaustive transparency may undermine the user experience. It may distract users, provide them with too much information, and provoke unnecessary doubt in the system. Transparency is trying to do too much. We cannot exhaustively convey the inner workings of many algorithms, nor is that what users want. However, without making these complete inner-workings transparent, how can we audit these systems for unfairness and bias?

As we have shown in previous work, diagnosing and fixing algorithmic bias is not a simple task, even for the creators of a system [9]. These creators have access to the complete code, data, and inner workings of the system; even with this access, fixing algorithmic bias is a challenge. How much harder will it then be for third parties and users to diagnose algorithmic bias through a transparent interface which does not display all of this information? We cannot reasonably expect that our current operationalization of transparency by explanation will allow third parties to diagnose bias in deployed systems.

In summary, these two goals of transparency conflict. We cannot simultaneously improve the user experience while providing a mechanism for diagnosing algorithmic bias. Providing enough information to diagnose algorithmic bias will overwhelm users and lead to poor experiences. On the other hand, scaffolding user mental models with selective transparency will not provide enough information to audit these systems for fairness. In order for transparency to be successful, we need to clarify our aims. We must separate transparency into two related concepts: explainability and auditability.

4 Two Facets of Transparency

The first facet, *explainability*, has a single goal: to improve the user experience. Many problems with intelligent systems occur because users lack proper mental models of how the system operates [14] and helping users form an accurate mental model improves satisfaction [22]. Therefore, the goal of explainability is to facilitate an 'accurate enough' mental model formation to enable correct action within the system. Attempting to go beyond helping users form heuristics may lead to worse user experience [35]. We need to give users heuristics and approximate understandings so that they can feel that they are in control of the interface.

The key to explainability is to reveal only the information needed by users [12]. This separates it from many current conceptualizations of transparency that aim for completeness. Explanations that aim for completeness may induce poor user experiences because they are too complex [19] or conflict with users' mental models [30,35]. In addition, explaining only the needed elements conforms better to the extensive bodies of social science research that study explanation. Explanations should follow Grices's maxims [17], i.e. to only explain as much as is needed and no more. Explanation should be occasioned [16], it should present itself when needed and disappear when not. Exhaustive transparency does conform with HCI experimental results or these social science theories; which is why it is essential that we study explainability.

Explainability can happen through a variety of means. For example, we can use natural language to explain results. For example, Facebook has a feature labeled 'Why am I seeing this?'

on ads that provides a natural language explanation of the user profile factors that led to the targeted ad. These explanations can also involve data and visualization intended to fill in gaps in the user's mental models [12]. The range of explanation types is large, from simple natural language to explorable explanations. This is necessary given the many domains in which explanations are needed. Explanations must be tailored to the domain; doctors have very different needs than mobile fitness coach users. For example, doctors are making high-stakes decisions and are likely to be very invested in each decision; therefore, the explanations for doctors should be more complete and contain more information. Such lengthy explanations may not be successful in more casual settings such as an intelligent mobile fitness coach where users may be less motivated to process a lengthy explanation. Again, explanations are to improve the use of the system and the user experience, not to provide the user the ability to ensure the system is fair and free from bias.

But how can transparency satisfy its second goal of ensuring fair algorithms? Explainability is insufficient to meet this requirement. It is not possible to ensure that an intelligent system is fair on the basis of the natural language explanations it provides. How then, can we determine whether algorithms are fair and free from bias?

In addition to explainability, the second facet of transparency is *auditability* of deployed systems. We define auditable as the ability for users or third parties to validate and test the deployed system by providing their own data for the system to predict on. While some systems are currently auditable, it is mostly adversarial; auditors must use methods such as sock-puppet auditing to determine whether a system is biased [29]. For an example of auditability, in Facebook, users are beholden to seeing advertisements targeted to their profile information. An auditable version of Facebook advertisements would have the ability to supply any profile data and receive back what targeted advertisements the supplied data would generate. A current example of systems that are easily auditable is current facial recognition APIs created by cloud providers; these are programmable and thus supplying data and checking for bias can be done by independent researchers [28].

Other definitions of auditability rely on seeing the code itself [8], but this may not be necessary. Relying on seeing the code itself complicates the audit process considerably due to source code being highly valued intellectual property. Rather we should pursue audits that allow the user or a third party to generate their own conclusions about the fairness of the algorithm, rather than relying on the explanations it generates. We do not need to know how the underlying algorithm works to ensure that it is generating fair predictions for all possible subgroups. Under many criterions of fairness such as independence and separation, all we need to know are the predicted output and the data [3]. Knowledge about the inner-workings of the algorithm is not required to ensure fairness. The expectation is not that every user has the skill or desire to audit these algorithms but rather that auditability is possible, in case it should be needed.

Given space constraints, we do not attempt to prescribe here exactly how auditability should be implemented. According to our definition, it could be as simple as an exposed public API endpoint that takes parameters and returns a prediction. While an API endpoint is the simplest implementation for developers, there is no reason that a user interface to supply data and view predictions could not be created. For instance, the E-meter we talked of earlier exhaustively exposed its predictions and data to users allowing them to edit and explore what text results in different predictions. These both fit the definition of auditability by allowing the user to provide known data as input and receive a prediction. While an API endpoint is a simple solution, further research should explore what form auditability should take in interactive programs.

6 Conclusion

Algorithmic transparency is purported to improve the user experience and simultaneously help diagnose algorithmic bias. We argue that these goals cannot be accomplished simultaneously with the same implementation. Exposing enough information to diagnose algorithmic bias overwhelms users and leads to a poor user experience. We therefore distinguish two aspects of transparency: explainability and auditability. Explainability aims to improve the user experience through making users aware of inputs and reasons for the system predictions; this is necessarily incomplete, providing just enough information to allow users to form simple mental system models. Auditability ensures that third parties and users can test a system's predictions for fairness and bias by providing their own data for predictions. Distinguishing these two aspects of transparency provides a way forward for industry implementations of usable and safe algorithmic systems.

ACKNOWLEDGMENTS

We would like to thank Victoria Hollis, Ryan Compton, and Lee Taber for their feedback on this project. We would also like to thank the anonymous reviewers for their insightful comments that helped refine this work.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–18. https://doi.org/10.1145/3173574.3174156

[2] Julia Angwin and Jeff Larson. 2016. Machine Bias. *ProPublica*. Retrieved October 27, 2017 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. Fairness and machine learning. *Fairness and machine learning*. Retrieved January 10, 2019 from https://fairmlbook.org/

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*: 1–14. https://doi.org/10.1145/3173574.3173951

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4349–4357.

[6] Raluca Budiu. 2018. Can Users Control and Understand a UI Driven by Machine Learning? *Nielsen Norman Group*. Retrieved January 10, 2019 from https://www.nngroup.com/articles/machine-learning-ux/

[7] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important?: a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, 169–178. Retrieved April 25, 2017 from http://dl.acm.org/citation.cfm?id=2166996

[8] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1: 2053951715622512. https://doi.org/10.1177/2053951715622512

[9] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and Addressing Algorithmic Bias in Practice. *Interactions* 25, 6: 58–63. https://doi.org/10.1145/3278156

[10] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1: 92–112. https://doi.org/10.1515/popets-2015-0007

[11] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The Role of Trust in Automation Reliance. *Int. J. Hum.-Comput. Stud.* 58, 6: 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

[12] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces* (IUI '18), 211–223. https://doi.org/10.1145/3172944.3172961

[13] Electronic Privacy Information Center. 2018. EPIC - Algorithms in the Criminal Justice System. Retrieved November 5, 2018 from https://epic.org/algorithmic-transparency/crim-justice/

[14] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I like it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*, 2371–2382. Retrieved April 25, 2017 from http://dl.acm.org/citation.cfm?id=2858494

[15] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn'T Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 153–162. https://doi.org/10.1145/2702123.2702556

[16] Harold Garfinkel. 1991. *Studies in Ethnomethodology*. Wiley.

[17] H. P Grice. 1975. *Logic and conversation*.

[18] Chloe Gui and Victoria Chan. 2017. Machine learning in medicine. *University of Western Ontario Medical Journal* 86, 2: 76–78. https://doi.org/10.5206/uwomj.v86i2.2060

[19] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW '00*, 241–250. https://doi.org/10.1145/358916.358995

[20] IUI ATEC. 2018. IUI ATEC Call for Papers. Retrieved January 10, 2019 from https://iuiatec.files.wordpress.com/2018/09/iui-atec-2019-call-for-papers.pdf

[21] René F. Kizilcec. 2016. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. 2390–2395. https://doi.org/10.1145/2858036.2858402

[22] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1–10. https://doi.org/10.1145/2207676.2207678

[23] Brian Y. Lim and Anind K. Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing*, 415–424. Retrieved April 25, 2017 from http://dl.acm.org/citation.cfm?id=2030168

[24] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. *Why and why not* explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, 2119. https://doi.org/10.1145/1518701.1519023

[25] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[26] Eli Pariser. 2011. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Books Limited.

[27] PricewaterhouseCoopers. 2018. Consumer Intelligence Series: Prepare for the voice revolution. *PwC*. Retrieved October 30, 2018 from https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/voice-assistants.html

[28] Inioluwa Deborah Raji and Joy Buolamwini. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. 7.

[29] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. 23.

[30] James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek Abdelzaher, and John O'Donovan. 2015. Getting the Message?: A Study of Explanation Interfaces for Microblog Data Analysis. In *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, 345–356. https://doi.org/10.1145/2678025.2701406

[31] Aaron Springer and Henriette Cramer. 2018. "Play PRBLMS": Identifying and Correcting Less Accessible Content in Voice Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 296:1–296:13. https://doi.org/10.1145/3173574.3173870

[32] Aaron Springer, Jean Garcia-Gathright, and Henriette Cramer. 2018. Assessing and Addressing Algorithmic Bias—But Before We Get There... In *2018 AAAI Spring Symposium Series*.

[33] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2017. Dice in the Black Box: User Experiences with an Inscrutable Algorithm. Retrieved April 24, 2017 from https://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15372

[34] Aaron Springer and Steve Whittaker. 2018. What are You Hiding? Algorithmic Transparency and User Perceptions. In *2018 AAAI Spring Symposium Series*.

[35] Aaron Springer and Steve Whittaker. 2019. Progressive Disclosure: Designing for Effective Transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19*.

[36] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Queue* 11, 3: 10:10–10:29. https://doi.org/10.1145/2460276.2460278

[37] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. *EACL 2017*: 53.

[38] Jeffrey Warshaw, Nina Taft, and Allison Woodruff. 2016. Intuitions, Analytics, and Killing Ants: Inference Literacy of High School-educated Adults in the US. 16.

[39] Daniel S. Weld and Gagan Bansal. 2018. The Challenge of Crafting Intelligible Intelligence. *arXiv:1803.04263 [cs]*. Retrieved September 20, 2018 from http://arxiv.org/abs/1803.04263

[40] Jenna Wiens and Erica S. Shenoy. 2018. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases* 66, 1: 149–153. https://doi.org/10.1093/cid/cix731