# AI Opaqueness: What Makes AI Systems More Transparent? (CAIS/ACSI 2020 Panel)

**Victoria L. Rubin, Jacquie Burkell, Sarah E. Cornwell, Toluwase Asubiaro, Yimin Chen, Danica Potts,** and **Chris Brogly**[*]
Faculty of Information and Media Studies (FIMS),
University of Western Ontario, London, Canada

**Suggested Citation**: Rubin, V.L., Burkell, J., Cornwell, S.E., Asubiaro, T., Chen, Y., Potts, D., and Brogly, C. (2020) AI Opaqueness: What Makes AI Systems More Transparent? (Panel), in the CAIS/ACSI2020 An Open Virtual Conference: Diverging Trajectories in Information Science, the 48th Annual Conference of the Canadian Association for Information Science Conference/l'Association canadienne des sciences de l'information. Accessible at: https://www.cais2020.ca/. The video content is also accessible at https://www.cais2020.ca/talk/ai-opaqueness/]

**Abstract:**

Artificially Intelligent (AI) systems are pervasive, but poorly understood by their users and, at times, developers. It is often unclear how and why certain algorithms make choices, predictions, or conclusions. What does AI transparency mean? What explanations do AI system users desire? This panel discusses AI opaqueness with examples in applied context such as natural language processing, people categorization, judicial decision explanations, and system recommendations. We offer insights from interviews with AI system users about their perceptions and developers' lessons learned. What steps should be taken towards AI transparency and accountability for its decisions?

## 1. Problem Statement (by Dr. Victoria L. Rubin).

**Artificially Intelligent (AI) algorithms are pervasive.** AI routinely mediates our interactions with online information. Social media platforms and other online services constantly advise their users on what actions to take: what to watch, read, purchase, who to contact, and which routes to take. Personal data are often used as the basis of personalization and profiling, as AI extrapolates certain characteristics to find patterns. However, transparency in AI-enabled computing is minimal.

**AI is poorly understood by users** due to a lack of awareness or understanding of the AI system internal workings. Computing mechanisms are often intentionally withheld or obscured due to proprietary nature of algorithmic 'know-hows.' Users are left wondering about how and why AI systems make certain choices, predictions, or conclusions.

**Machine Learning (ML) is often a 'black box' to the developers** who may not be able to scrutinize the myriads of statistical calculations that are either inaccessible or no longer humanly traceable.

There are issues with AI input, mechanisms, output and associated explanations that this panel discusses, as various important aspects of AI opaqueness:

A. **AI input data transparency** addresses inherent data biases, stereotypes and language transparencies.

[*] See the panelists' brief biographies at the end.

B. **AI output transparency** is gained through interpretability of the automated results. How AI verdicts are rendered **– through what mechanisms -** should be explained to users, to the best of developers' abilities.
C. **AI decision accountability** should be based on what system users expect of AI in a particular application context, and how particular verdicts are justified.

**Why is this important?** If we are to co-exist with algorithmic mediation of information in online environments and accept automated verdicts, we need ways to assess AI technologies, and hold developers responsible for the decisions offered. In turn, assessments can increase trust in the systems or reveal flaws.

## 2. **Panel Structure**

The panel joins efforts of two research groups to discuss various aspects of AI opaqueness as an emerging trend in Library and Information Science and Technology (LIS&T) in 5 brief talks. The talks are preceded by a brief primer layout out the problems. After a brief summary, the audience is invited to participate in a live Q&A session (on Zoom).

## 3. **Panel Contributions**

The following five talks explain how external requirements may impact system design, what users expect to see in output, what opaqueness means in a specific context, how ambiguous language-based input can be, and what solutions should be sought to rectify the associated problems:

### **1. What's in an Explanation? Judicial Reasons as Models for Algorithmic Decision-Making Explanations (by Dr. Jacquie Burkell).**

What constitutes an 'explanation' for an algorithmic decision? The answer depends on a number of factors, including who is asking seeking the explanation and for what purpose. Recently, particularly under new provisions of the European General Data Protection Regulation (GDPR), one specific audience has come into sharp focus. That legislation, which has prompted widespread discussion, states that the *subjects of algorithmic decisions* have the right to an explanation of the decision reached. Simple transparency regarding code and the data that go into the decision will rarely if ever constitute an acceptable explanation for those affected by algorithmic decisions.

Judicial reasons – the written explanations for judicial decisions – are one form of public-facing explanation for decisions. These explanations must satisfy the criterion of public accountability, and therefore provide one potential model for algorithmic explanations that must accomplish the same outcome.  The discussion integrates three bodies of literature: (i) the purpose and function of 'explainable AI'; (ii) the relevant case law, judicial commentary and legal literature focused on the form and function of reasons for judicial decisions; and (iii) the psychological and sociological functions of these reasons for judicial decisions from the perspective of the public. Together, this literature suggests that while judicial reasons, instead of being accurate reflections of the decision *process*, are essentially decision *justifications* that situate and justify the decision within the rule of law. This form of explanation – one that explains *why*, rather than *how* a decision was reached – meets the needs of that affected, who are also looking for justification of the decision that affects them. One model for algorithmic explanations, therefore, is a post-hoc articulation of principles and precedents that support the decision or action — an explanation that places the result in dialogue with established and external standards and practices.

**2. What does the public desire in an explanation when they are subjected to algorithmic decision-making? (by Danica Potts)** The use of algorithms to make decisions about people has become commonplace in both the public and private sectors. As these algorithms become more advanced, there is a concern for the ethical implications of subjecting people to these 'black boxes.' What do people want to know about this process, when they are subjected to these algorithmic decisions? Policy and regulation are starting to require some transparency and communication of explanations to users, e.g., the 'right to an explanation' embedded in the European Union's General Data Protection Regulations (Goodman and Flaxman, 2016). However, there is limited guidance on what such explanations look like.

Explanations can include various levels of detail and complexity such as about the system itself—even providing the code—or about the factors that led to a decision. The timing, format, and delivery can also vary. Context is key: an explanation is successful only if it serves the recipient's interests (Van Fraasen, 1980).

We interviewed 20 members of the public to gain insight into what AI's users desire in an explanation. Scenarios explored included hiring decisions, credit applications, and passport applications, among others. As expected, there is not a clear consensus. Most, however, care less about the description of the algorithmic system itself, and more about *why* that particular decision was made—but often only in the case of a negative outcome.

**3. Opaqueness in Recommender Systems (by Toluwase Asubiaro).** Recommender systems suggest products and services to users of online systems based on data collected about the users. They also persuade, and emphasize relevance and usefulness (Gretzel & Fesenmaier, 2006). Due to their potential negative social effects, there is a growing concern about popular recommender systems' operations. Some of them amplify conspiracy theories, promote gamified news, infiltrate mainstream discourse with extreme and nonsensical contents, and promote misinformation (DiResta, 2018). For example, popular but questionable YouTube recommender system, through its "algorithmic selective exposure," is characterized by increased "likelihood for users to come across videos with a contrary message, just because of thematic congruence" (Schmitt et al., 2018), and it systematically leads users to extreme and radical content (Ribeiro et al., 2019; Tufekci, 2018). Tangential topics or groups are recommended to Facebook users based on the content that similar profiles have searched in its bid to present content from friends or groups with homophilic profile. "YouTube leads viewers down a rabbit hole of extremism" (Tufekci, 2018), on the other hand, Facebook, "rather than pulling a user out of the rabbit hole, the recommendation engine pushes them further in" (DiResta, 2018)

Explanations, a mechanism for achieving transparency (Abdollahi & Nasraoui, 2018), are lacking, which leads to many 'whys' and 'hows' in the results presented to users. One of the sources of bias in ML algorithms is the data source (Abdollahi & Nasraoui, 2018). Are recommendations still based on user data? And if so, which types of data are collected? Or, is there some pre-conceived content that is being 'fed' to the users? If recommendations are more about the businesses they advertise and not the users, one begins to ask if recommendation systems still work to help reduce information overload, as originally intended. Do ‘big tech’ corporations need to be asked to abandon their rhetoric of ‘helping users’ and acknowledge user exploitation?

**4. How do algorithms reproduce biases?: Hidden variables in language data (by Sarah E. Cornwell).** ML (‘intelligent’) algorithms manipulate input variables with the goal of determining which combination(s) will produce the best results on a given task. The ‘black boxed’ nature of these programs makes it difficult for users to know which variables are included and why one combination might be better than another. In addition, the complexity of language data means that developers are often unaware of the variables which have been provided for the program to identify and manipulate. For example, is a construction like “she be running” an indication of a second language speaker? a typo? or a speaker using a Black English variety in which ‘be’ communicates that the behaviour is habitual? One’s age, gender, place of origin, native language, ethnicity, social class, and religion can be determined from relatively small amounts of linguistic data, but these factors are rarely – if ever – controlled for in AI studies using NLP. This means that algorithms can reproduce biases and stereotypes about these variables: biased training data cannot be used to produce an unbiased AI. This is especially concerning when developers are unaware of the types of biases that may be introduced in NLP datasets.

**5. Towards a More Transparent NLP System Design: Clickbait Detector as an Example (by Yimin Chen and Chris Brogly).** We demonstrate a recently developed system that uses supervised Machine Learning methodology to identify clickbait in digital news (Rubin et al., 2019). "Clickbait" is a hyperlinked headline that primarily attracts readers’ attention but leads to uninformative content, and it can be contrasted with traditional more informative ‘headlinese’ (Chen and Rubin, 2017). The Clickbait Detector automatically distinguishes clickbait from non-clickbait with 94% accuracy, when tested on 11,000 hyperlinks (Brogly and Rubin, 2019). The user feeds in a website URL and the system decides how clickbaity the news website is (slightly, moderately or extremely clickbaity, or not at all). The user interface shows real-time colour-coded analysis of any news website and a label for each individual hyperlink.

For the purposes of this panel, the Clickbait Detector, serves as a counter example to secretive ‘know-hows.’ The 38 features of news headlines, used for its development, are detailed in a journal publication accompanying the system (Brogly and Rubin, 2019). Visualizing intermediate steps in the UI demystifies the verdict to the users. The system’s code is open access published via GitHub and is accessible to public for download, and ML/NLP experts for improvements.

## 4. Conclusions (by Dr. Victoria L. Rubin)

Each panelist discusses AI opaqueness in an applied context: judicial decision explanations, automated people categorization, recommender systems, perceptions of language data and detection of various manipulative language in natural language processing (NLP). The unifying themes are: What does AI opaqueness mean? How do we experience non-transparency of AI-based system in our interactions with them? What needs to change, and how, in each case?

This panel positions the problem of AI opaqueness in LIS&T from the users' perspective but acknowledges developers' responsibilities. There is a need for greater clarity in explanation on how AI systems make their decisions in classification, prediction, detection or selection of results. From the developers' perspective, we call for more user-centered approaches in AI system design. More critical research on transparent algorithmic practices is needed in LIS&T. The introduction of regulatory interventions should also be considered.

We look forward to the audiences input on the outlined problem. We welcome further reflections and suggestions on best policies for information professionals in schools, libraries, and other institutions, as well as educational campaigns to raise awareness of AI systems opaque practices.

## References

Abdollahi, B., & Nasraoui, O. (2018). Transparency in Fair Machine Learning: The Case of Explainable Recommender Systems. In J. Zhou & F. Chen (Eds.), *Human and Machine Learning* (pp. 21–35). Springer International Publishing. https://doi.org/10.1007/978-3-319-90403-0_2

Brogly, C. and Rubin, V. L. 2019. Detecting Clickbait: Here's How to Do It / Comment détecter les pièges à clic. Canadian Journal of Information and Library Science, 423-4:154-175

Chen, Y. And Rubin V. L. (2017). Perceptions of Clickbait: A Q-Methodology Approach. In The Proceedings of The 45th Annual Conference Of The Canadian Association For Information Science/L'Association Canadienne des Sciences de L'information (CAIS/ACSI2017), Ryerson University, Toronto, May 31 - June 2, 2017.

DiResta, R. (2018). The Web's Recommendation Engines Are Broken. Can We Fix Them? *Wired*. https://www.wired.com/story/creating-ethical-recommendation-engines/

Goldstein, D. G., Johnson, E. J., Herrmann, A., & Heitmann, M. (2008, December 1). Nudge Your Customers Toward Better Choices. *Harvard Business Review, December 2008*. https://hbr.org/2008/12/nudge-your-customers-toward-better-choices

Goodman, B., & Flaxman, S. (2016, June). EU regulations on algorithmic decision-making and a "right to explanation". In *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY*. http://arxiv.org/abs/1606.08813 v1.

Gretzel, U., & Fesenmaier, D. R. (2006). Persuasion in Recommender Systems. *International Journal of Electronic Commerce*, *11*(2), 81–100. https://doi.org/10.2753/JEC1086-4415110204

Lewis, P., & McCormick, E. (2018, February 2). How an ex-YouTube insider investigated its secret algorithm. *The Guardian*. https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot

Nicas, J. (2018, February 7). How YouTube Drives People to the Internet's Darkest Corners. *Wall Street Journal*. https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2019). Auditing Radicalization Pathways on YouTube. *ArXiv:1908.08313 [Cs]*. http://arxiv.org/abs/1908.08313

Rubin, V. L, Brogly, C., Conroy, N., Chen, Y., Cornwell, S., & Asubiaro, T. (2019). A News Verification Browser for the Detection of Clickbait, Satire, and Falsified News. Journal of Open Source Software, 4(35), 1208. https://doi.org/10.21105/joss.01208

Schmitt, J. B., Rieger, D., Rutkowski, O., & Ernst, J. (2018). Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube. *Journal of Communication*, *68*(4), 780–808. https://doi.org/10.1093/joc/jqy029

Tufekci, Z. (2018, March 10). YouTube, the Great Radicalizer. *The New York Times*. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.

## PANEL PARTICIPANTS' BRIEF BIOGRAPHIES

**Victoria L. Rubin** is Associate Professor at the Faculty of Information and Media Studies (FIMS) at the University of Western Ontario, the Director of the Language and Information Research Lab (LiT.RL). She broadly specializes in information retrieval and natural language processing techniques, currently working on automated detection of misinformation and disinformation in news.

**Jacquie Burkell** is (Acting) Associate Vice-President (Research) and is an associate professor in the Faculty of Information & Media Studies. She holds a PhD in Psychology (Cognitive Science) from Western and Jacquelyn served as the faculty's Assistant Dean of Research for seven years and chaired the Associate Deans (Research) group from 2016-2018. Throughout her career, Jacquelyn has served on a wide variety of academic committees, including the 2016 URB Task Force Steering Committee – Support for Research in Social Sciences, Arts, and Humanities at Western. A highly collaborative scholar, Jacquelyn is a co-investigator on two SSHRC partnership grants – one examining artificial intelligence in the context of justice, the other focused on youth equality and privacy online. More broadly, her research focuses on the social impact of technology and examines how technological mediation changes social interaction and information behaviour.

**Sarah E. Cornwell** is a doctoral candidate in the LIS program at FIMS. Building on previous degrees in linguistics (MA), and Anthropology & Cognitive Psychology (BAS), her research interests include multilingualism, natural language processing, and everyday information seeking. In essence, she focuses on the interaction of information technologies and human linguistic diversity.

**Toluwase Asubiaro** is PhD candidate at the Western University's LIS program. He holds a B. Sc. in Mathematics and Master of Information of Science. He started his career in Language Technology in 2012 as a volunteer research assistant in African Languages Technology Initiative (ALT-I), Ibadan, Nigeria. His major research interest is informetrics, natural language processing, information retrieval and automatic language identification. Prior to his present position in LiT.RL, he had contributed to studies on language technology for Yoruba, a Nigerian language.

**Yimin Chen** is a PhD candidate in Library and Information Science in the Faculty of Information and Media Studies at the University of Western Ontario. His research examines the communicative practices of online communities and cultures, with a focus on Internet trolling behaviors and the controversy surrounding them. His previous projects range from fake news and deception detection, to library automation, to the impact of political memes on social media.

**Danica Potts** is a PhD student in LIS at Western University interested in the social and ethical issues of artificial intelligence. Her PhD research focuses on the nature of trust between humans and artificial intelligence and the implications of trust on users' information behaviour and decision-making processes. Other projects Danica is a part of look at algorithmic decision making and explanations, and algorithmic literacy.

**Chris Brogly** is a Doctoral Student in the Health Information Science program at the University of Western Ontario. He completed his MSc in Computer Science from Western in 2017 and a BCS Co-op from the University of Windsor in 2015. His research interests are in health informatics, predictive modelling, and the internet.