International Journal of Communication 10(2016), 4991–5002 1932–8036/20160005

Auditing for Transparency in Content Personalization Systems

BRENT MITTELSTADT[1]
Oxford Internet Institute, University of Oxford, UK

Do we have a right to transparency when we use content personalization systems? Building on prior work in discrimination detection in data mining, I propose algorithm auditing as a compatible ethical duty for providers of content personalization systems to maintain the transparency of political discourse. I explore barriers to auditing that reveal the practical limitations on the ethical duties of service providers. Content personalization systems can function opaquely and resist auditing. However, the belief that highly complex algorithms, such as bots using machine learning, are incomprehensible to human users should not be an excuse to surrender high quality political discourse. Auditing is recommended as a way to map and redress algorithmic political exclusion in practice. However, the opacity of algorithmic decision making poses a significant challenge to the implementation of auditing.

Keywords: algorithms, transparency, information ethics, automation, politics, ethics, personalization, recommendation systems, bots

A central ideal of democracy is that political discourse should allow a fair and critical exchange of ideas and values. Political discourse is unavoidably mediated by the mechanisms and technologies citizens use to communicate and to receive information. Personalized content is now delivered to users through modern search engines, social media news feed features, and targeted advertisements (Goldman, 2006; Pariser, 2011). Traditional media channels still broadcast homogenous information to a heterogeneous audience.

Content personalization systems display information tailored to individual users, often based on perceived preferences or past behaviors. Content is filtered to fit the user's profile, meaning "the system can predict what will be relevant for the user, filtering out the irrelevant information, increasing relevance and importance to an individual user" (Bozdag, 2013, p. 211). Personalization is accomplished through interactions of (a) prioritization algorithms that decide which topics are (and are not) trending (Bozdag,

Brent Mittelstadt: brent.mittelstadt@oii.ow.ac.uk
Date submitted: 2016–08–29

[1] I gratefully acknowledge the feedback from the day-long workshop *Algorithms, Automation and Politics*, organized by the European Research Council–funded Computational Propaganda project of the Oxford Internet Institute and held as a preconference to the International Communication Association Meeting in Fukuoka, Japan, in June 2016. Any opinions, findings, conclusions, or recommendations expressed in this material are mine and do not necessarily reflect the views of the European Research Council.

2013); (b) profiling algorithms that infer user preferences and attributes from small patterns or correlations, by which individuals are clustered into meaningful groups according to their behavior, preferences, and other characteristics (Hildebrandt, 2008; Mittelstadt & Floridi, 2016; Schermer, 2011); and (c) automated bots that post and interact directly with users to promote certain content or viewpoints, seen, for instance, in widespread use of bots on Twitter, Facebook, and Reddit during the 2016 U.S. presidential election (Woolley, 2016) and 2016 UK European Union referendum (Howard & Kollanyi, 2016).

Content personalization systems, and the algorithms they rely upon, create a new type of curated media that can undermine the fairness and quality of political discourse. Automated systems can have inadvertent and unexpected effects, seen, for instance, in the potential misogynistic effects of YouTube's "reply girl" phenomenon in 2012 (Sandvig, Hamilton, Karahalios, & Langbort, 2014) and the formation of echo chambers of like-minded individuals on social networking sites that tend to be self-reinforcing (Leese, 2014; Macnish, 2012). Public assessment of the extent and source of these problems is often difficult, owing to the use of complex and opaque mechanisms that decide which content should be displayed next based upon unseen categorical computational judgments (Ananny, 2016; Sandvig et al., 2014).

At a minimum, personalization systems can undermine political discourse by curbing the diversity of ideas that participants encounter (Leese, 2014; Macnish, 2012) and by obscuring the external interests and mechanisms that influence their beliefs (Sandvig et al., 2014). As information gatekeepers, these systems threaten the fairness and openness of political discourse with subtle, often secretive, but sustained mediation of participants' views. Populations can be segmented so that only some groups are worthy of receiving certain opportunities or information, the fairness of which has been questioned (Cohen, Amarasingham, Shah, Xie, & Lo, 2014; Danna & Gandy, 2002; Rubel & Jones, 2014). Citizens may require a political right to transparency to limit the power of opaque content personalization systems on political discourse.

This article explores the challenges of enforcing a political right to transparency in content personalization systems. First, it explains the value of transparency to political discourse and suggests how content personalization systems undermine open exchange of ideas and evidence among participants. Second, it explores work on the detection of discrimination in algorithmic decision making, including techniques of algorithmic auditing that service providers can employ to detect political bias. Third, it identifies several factors that inhibit auditing and thus indicate reasonable limitations on the ethical duties service providers incur. Content personalization systems can function opaquely and be resistant to auditing because of poor accessibility and interpretability of decision-making frameworks. Finally, the article concludes with reflections on the need for regulation of content personalization systems.

A Political Right to Transparency

Transparency is often assumed to be an ideal for political discourse in democracies. Transparency is generally defined with respect to "the availability of information, the conditions of accessibility and how the information . . . may pragmatically or epistemically support the user's decision-making process"

(Turilli & Floridi, 2009, p. 106). A minimal requirement for democratic political discourse is a willingness to see things from one another's point of view, which cannot be achieved without encountering and understanding alternative views.

Political discourse can be a type of communicative action necessary for social living. According to Jürgen Habermas's (1984, 1985) theory of communicative action, humans have a range of ways to communicate, with communicative action being the best or most highly developed. Whenever human beings communicate, a set of validity claims arise: truth (*wahrheit*), rightness (*richtigkeit*), and authenticity (*wahrhaftigkeit*). Communicative action requires the speaker to engage in a discourse whenever any of these validity claims are questioned. Discourse requires participants to be willing to engage with the speaker, to take the speaker seriously, and to be willing to change their positions based upon the argument (Habermas, 1997). In the context of political discourse, a prerequisite for these conditions is transparency—each participant must be willing to openly share validity claims and explain them when questioned. The claims, evidence, beliefs, or values must be accessible to participants in the discourse for the questioning to be possible.

Habermas (1990) eventually developed *discourse ethics* as an explicit ethical theory incorporating the principles of his theory of communicative action. Discourse ethics goes some way to explaining the political value of transparency by providing two principles to evaluate the quality of any discourse: the discourse principle and the universality principle. The discourse principle states that norms can be valid if they meet with the approval of all participants in a practical discourse. The universality principle goes beyond acceptability to affected participants and states that the consequences and side effects arising from the general adherence to a norm have to be acceptable for all involved stakeholders and, ideally, for everybody.

The algorithmic personalization of content inevitably involves normative choices, despite the frequent portrayal of algorithmic decision making as neutral or objective (Bozdag, 2013; Naik & Bhide, 2014). Yet, work on the normativity of information technologies in general and algorithms in particular suggests that algorithms do not operate neutrally (Friedman & Nissenbaum, 1996; Newell & Marabelli, 2015). Research on information filtering (Bozdag, 2013), surveillance (Macnish, 2012), analytics (Tene & Polonetsky, 2013), and clinical decision-support systems (Kraemer, van Overveld, & Peterson, 2011) also confirm the bias behind code.

Algorithms are created in value-laden development processes and inevitably make biased decisions. An algorithm's design and functionality reflects the values of its designer and intended uses, if only to the extent that a particular design is preferred as the best or most efficient option. Coded automation will also interact with users in unanticipated ways. Emergent bias can be the result (Johnson, 2006). Personalization algorithms can also be purposefully or inadvertently discriminatory (Romei & Ruggieri, 2014), what amounts to political sabotage by obscuring opposition messages.

Personalization systems prioritize particular actors and types of information and knowledge according to the explicit or inferred desires of users (Hinman, 2008). Both the consumption of information and interactions between human actors are core components of modern political discourse. Systems that

determine which information users encounter and that encourage communication between particular users inevitably influence political discourse. The quality of discourse is undermined when personalization systems hinder political actors' capacity to assess the veracity and history of their beliefs.

This type of hindrance is typical of content personalization systems. Explanations of how they function are not normally publicly available. Users therefore cannot assess the ways in which information encountered online is limited in scope or representations of political viewpoints. Without this explanation, validity claims are not fully open to discursive questioning; hence, the person advancing the claim cannot explain these limits or how his or her opinion may be biased as a result of prejudiced or incomplete information about the topic at hand. This is arguably a characteristic of normal self-reflection, insofar as humans are never aware of all their biases. However, the difference is that personalization systems purposefully and systematically limit queries for new information according to a prevailing set of values. Content personalization thus introduces new gaps in self-reflection as a matter of routine.

A political right to transparency in content personalization would thus be a right for citizens to be kept informed about the methods and extent to which personalization systems advance political agendas and influence actors in political discourse. The right would not necessarily prevent this influence, but rather inform actors of its existence and the informational blind spots personalization sustains by default.

Algorithm Auditing

Assuming political actors possess a right to transparency in content personalization systems, what might the right require of service providers? At a minimum, information about the influence of personalization systems handling political information must be accessible and comprehensible for people to be aware of how their political views are being externally shaped (Turilli & Floridi, 2009). Poorly comprehensible and opaque information cannot be transparent, regardless of its accessibility.

Auditing is one possible mechanism for achieving transparency. For all types of algorithms, auditing is a necessary precondition to verify correct functioning. For platforms that mediate political discourse, auditing can create a procedural record to demonstrate bias against a particular group. Auditing can help to explain how citizens are profiled and the values prioritized in content displayed to them.

Auditing is a process of investigating the functionality and impact of decision-making algorithms. Functionality auditing allows for prediction of results from new inputs and explanation of the rationale behind decisions, such as why a new input was assigned a particular classification. If the logic behind decisions made by an algorithm must be understood, reporting only the features of data relevant to the classification may be sufficient (Burrell, 2016). Code audits are also feasible in some cases, although algorithms designed by large teams over time are too difficult to audit because of the human effort and expertise required to untangle the logic of the code (Burrell, 2016). Additionally, bias and discrimination may emerge only when an algorithm processes particular data, meaning a code audit alone would detect only a limited range of problems (Sandvig et al., 2014).

In principle, functionality auditing could occur during an algorithm's development. Errors and bias can emerge at each stage of development, from an algorithm's initial definition as a mathematical object, through implementation into a software system, to configuration for a specific task. Functioning can be verified at each stage:

(1) Definition: It can be proven that an algorithm is mathematically correct.
(2) Implementation: It can be proven that a piece of code correctly implements a well-defined algorithm into a technological system according to its specifications.
(3) Configuration: It can be proven that the system has been appropriately configured to provide accurate results.

When evaluating an algorithm's functionality, the correctness of its definition and implementation should be taken for granted. This is important because certain unethical outcomes should already be ruled out at this stage (Turilli, 2007). Verification of functionality is possible through impact auditing. Inaccuracies and prejudices in training data sets can produce discrimination against entire classes of data subjects. Such systematic mistakes should not be treated as mere collateral damage if they are preventable or solvable through systematic auditing.

Many kinds of technical systems break down, but users are not always able to detect technical faults. Automobiles can be operated without understanding how an internal combustion engine works. A driver with basic functional knowledge can determine when the car he or she is driving has failed. If the engine does not start, the car has failed and requires repair. For the driver, identifying when a problem has occurred is trivial compared to identifying why. To understand the nature of the failure or the steps that led to it and how to fix it, specialist knowledge and skills that most drivers do not possess are required. Compare this with user profiling for the sake of content personalization. To unpack the nature of a failure, such as the steps that led to an inaccurate profile, specialist knowledge and access that the lay person is unlikely to possess are also required. However, in profiling, it remains highly unlikely that the failure will be evident to the data subject; unlike cars, profiles and personalization systems do not fail in obvious ways or in ways that users necessarily find problematic (Sandvig et al., 2014). Profiles implicitly structure the information and opportunities offered to the data subject, but it is difficult to identify when better information could have been offered or how the user would have perceived its value.

Impact auditing investigates the types, severity, and prevalence of effects of an algorithm's outputs (Barocas & Selbst, 2015; Hajian, Domingo-Ferrer, & Martinez-Balleste, 2011). Owing to the poor accessibility and interpretability of many decision-making algorithms, much of the work in auditing focuses on impact, for instance, to detect discrimination in decisions already made (Barocas & Selbst, 2015; Kroll et al., 2016; Sandvig et al., 2014). Several approaches comparable to audit studies in the social sciences and A/B testing are also feasible, including noninvasive user, scraping, sock-puppet, and crowdsourced audits (Sandvig et al., 2014). In each type of auditing, human users, bots, or scraped data identify problematic effects in decisions made by the algorithm. Audit studies can identify disproportionate impact across a population. For example, personalized pricing on Amazon.com can be detected by users sharing time-stamped prices for a product. However, this type of testing does not reveal anything about the

process and culture that created the system (Ananny, 2016), why prices are different, or why only certain groups of users are affected.

Auditing is therefore useful for revealing when content personalization has biased political discourse. However, auditing is infeasible without strong regulatory compulsion or cooperation from service providers (Sandvig et al., 2014). For a political right to transparency to have any meaning, the providers of social networking services must cooperate in audits.

Feasibility of Auditing

Auditing mechanisms may be able to render algorithmic influence in political discourse transparent. However, many epistemic, technical, and practical challenges must first be overcome. Personalization systems in general "are opaque in the sense that if one is a recipient of the output of the algorithm, rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs" (Burrell, 2016, p. 1).

Thus, an algorithm is opaque when a user can access but not interpret information about the algorithm (Burrell, 2016). Personalization systems are typically proprietary and inaccessible to users, and information about the functionality of algorithms is often kept secret for competitive advantage (Glenn & Monteith, 2014; Stark & Fins, 2013), national security (Leese, 2014), or privacy. A political right to transparency might undermine the privacy of data subjects and autonomy of service providers.

Algorithmic personalization contrasts with traditional media, which provides identical content to all consumers. Reporters and editors can normally articulate the rationale for producing a particular news item. In contrast, the rationale of an algorithm can be epistemically inaccessible, rendering the legitimacy of decisions difficult to challenge. Interpretability is the degree to which the decision-making logic of an algorithm is comprehensible to humans (Lisboa, 2013).

Specialist knowledge is often required to comprehend the decision-making logic of an algorithm. Simplified reporting of classification and decision structures would be ideal where feasible (Tene & Polonetsky, 2013). Even when information is released in good faith, the algorithm's decision-making rules may be so complex as to exceed human capacities for comprehension or practical resources for oversight, especially when algorithmic output can be modeled only in complex, probabilistic ways (Van Otterlo, 2013).

Is it even possible to map algorithmic decision-making frameworks to explain and prevent problematic effects? A decision can be poorly interpretable because of the number of decision-making rules involved, which can be difficult to visualize (Matthias, 2004). The computational resources, human effort and time required to reverse engineer a classification scheme by omitting individual data points can be prohibitive (Sandvig et al., 2014). Complex decision-making structures can easily exceed auditors' resources.

Many personalization systems rely upon machine learning classifier algorithms, which tend to be poorly interpretable (Bozdag, 2013). Machine learning algorithms process new inputs and construct models or classification structures; image recognition technologies, for example, can decide what types of objects appear in a picture. The algorithm learns by defining rules to determine how new inputs will be classified (Schermer, 2011; Van Otterlo, 2013). Normal operation does not require the human operator to understand the rationale of the decision-making rules the algorithm produces (Matthias, 2004).

The rationale for algorithmic decisions can be particularly difficult to interpret when decision-making rules are defined in situ. Training produces a structure of rules and weights to classify new inputs and predict unknown variables. Once trained, an algorithm can process and categorize new data automatically without operator intervention (Burrell, 2016). The rationale of the algorithm is "hidden . . . and cannot be used as evidence" for findings (Leese, 2014, p. 503), leading to the portrayal of machine learning algorithms as "black boxes" (Floridi, 2010). This is a problem not only of the system being too complex for a single expert to decipher; the logic of the algorithm's decisions may be fundamentally incomprehensible to humans. Decision-making rules need not be created with human comprehension in mind (Burrell, 2016).

A simple example will demonstrate this problem. Consider an image-recognition neural network algorithm being trained to identify the breed of a dog from a picture. The algorithm is given a set of hand-labeled images and expert research on defining traits of different breeds, such as visual models of skull shape, as training data. The algorithm learns in the sense that it builds classes for each breed based on characteristics it perceives in the pictures. Algorithmic and human learning differ in the mechanisms underlying perception. The mechanisms that allow algorithms to perceive inputs as something differ from the subconscious mechanisms of human perception but produce similar outputs—an observed classification. In both cases, these processes are subjectively opaque, or at least imperfectly describable to an external observer. One cannot assume that these characteristics will be the same as those perceived by humans when identifying a breed of dog (Burrell, 2016). Unless explicitly designed to do so, the algorithm learns without concern for the human interpretability of its work.

Predicting how the system will handle new inputs remains highly uncertain because of the gap between machine and human comprehension (Burrell, 2016). To reiterate a passage from Allen, Wallach, and Smit's (2006) seminal work on machine learning: "No single person or group can fully grasp the manner in which the system *will* interact or respond to a complex flow of new inputs" (p. 14, emphasis added). The rationale for decisions already made can be similarly obscured for audits (Barocas & Selbst, 2015). Some forms of machine learning prohibit any reconstruction of the rationale of a decision using replication. For instance, approaches to machine learning based on trial and error or estimates with random number generation will produce similar but not identical results when given identical training data. A machine learning classifier does not need to correctly classify all new inputs to function as intended. Errors, or false positive and negatives, are an inevitable aspect of learning. The impact of proposed changes to a learning algorithm's design or training data set is thus highly unpredictable. Observing impact can fail to provide actionable insight into the system's functionality, much less help an auditor adapt a system for socially and ethically desirable ends.

Conclusion

Despite the many barriers, algorithm auditing may be quickly approaching (Tutt, 2016). Expectations that algorithmic decisions will be minimally comprehensible and accessible to questioning can already be found in data protection legislation. The forthcoming EU General Data Protection Regulation, for example, will require data processors to maintain a relationship with data subjects and explain the logic of automated decision making when questioned[2] (European Commission, 2012). The regulation may prove a much needed impetus for algorithmic auditing. However, with opacity, implementing this right in a practically useful form for data subjects will be extremely difficult.

Several challenges remain before personalization systems can be audited, but they must not be taken for granted. It is a mistake to take a lack of access, expertise, or reporting mechanisms as a symptom of overwhelming complexity (Leese, 2014; Matthias, 2004; Schermer, 2011). The belief that highly complex algorithms, particularly those involved in machine learning, are incomprehensible to human observers should not be used as an excuse to surrender high quality political discourse. The theoretical and practical feasibility of predicting and auditing different types of classifiers, particularly those in machine learning, must also be explored.

Another outstanding challenge is how to trigger and manage auditing. Users have few opportunities to identify bias in personalization. Their perspectives will normally be limited to the content displayed. Errors in content can be identified, and hypotheses drawn about the cause of the error, such as why an irrelevant advertisement was displayed. The user's perspective is limited to observable returned content. Auditing must ideally go beyond this to include the information or content that has not been displayed; personalized content cannot be considered biased without some knowledge of the available alternatives.

One possibility for managing algorithmic auditing would be a regulatory body to oversee service providers whose work has a foreseeable impact on political discourse by detecting biased outcomes as indicated by the distribution of content types across political groups (Barocas & Selbst, 2015). Tutt (2016) suggests a regulatory agency for algorithms may be required, and this agency can "classify algorithms into types based on their predictability, explainability, and general intelligence" (p. 15) to determine what must be regulated. Certain types of machine learning could, for example, be banned or severely restrained for content personalization systems to prevent emergent bias or discrimination against protected political classes via proxy features. A regulator or other trusted third party could also require companies to file qualitative disclosures that "provide meaningful notice about how the algorithm functions, how effective it is, and what errors it is most likely to make" without revealing proprietary design details (Tutt, 2016, p. 17).

---

[2] It is worth noting that this right has existed in some form as the Data Protection Directive 95/46/EC. To my knowledge, no member state has yet implemented a workable mechanism for individuals to question automated decision making.

No matter how auditing is pursued, standards to detect evidence of political bias in personalized content are urgently required. Methods are needed to routinely and consistently assign political value labels to content delivered by personalization systems. This is perhaps the most pressing area for future work—to develop practical methods for algorithmic auditing.

The right to transparency in political discourse may seem unusual and farfetched. However, standards already set by the U.S. Federal Communication Commission's fairness doctrine—no longer in force—and the British Broadcasting Corporation's fairness principle both demonstrate the importance of the idealized version of political discourse described here. Both precedents promote balance in public political discourse by setting standards for delivery of politically relevant content. Whether it is appropriate to hold service providers that use content personalization systems to a similar standard remains a crucial contemporary question.

References

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *Intelligent Systems, IEEE*, *21*(4). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1667947

Ananny, M. (2016). Toward an ethics of algorithms convening, observation, probability, and timeliness. *Science, Technology & Human Values*, *41*(1), 93–117. http://doi.org/10.1177/0162243915606523

Barocas, S., & Selbst, A. D. (2015). Big data's disparate impact. Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2477899

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227. http://doi.org/10.1007/s10676-013-9321-6

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, *(3)*1. Retrieved from http://bds.sagepub.com/content/3/1/2053951715622512

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, *33*(7), 1139–1147. http://doi.org/10.1377/hlthaff.2014.0048

Danna, A., & Gandy, O. H., Jr. (2002). All that glitters is not gold: Digging beneath the surface of data mining. *Journal of Business Ethics*, *40*(4), 373–386.

European Commission. (2012). *Regulation of the European parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data: General data protection regulation* (No. COM[2012] 11). Retrieved from http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf

Floridi, L. (2010). *Information: A very short introduction*. Oxford, UK: Oxford University Press.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347.

Glenn, T., & Monteith, S. (2014). New measures of mental state and behavior based on data collected from sensors, smartphones, and the Internet. *Current Psychiatry Reports*, *16*(12), 1–10. http://doi.org/10.1007/s11920-014-0523-3

Goldman, E. (2006). Search engine bias and the demise of search engine utopianism. *Yale Journal of Law & Technology*, 2005–2006, 6–8.

Habermas, J. (1984). *The theory of communicative action, vol. 1: Reason and the rationalization of society*. Boston, MA: Beacon Press.

Habermas, J. (1985). *The theory of communicative action, vol. 2: Lifeworld and system: A critique of functionalist reason*. Boston, MA: Beacon Press.

Habermas, J. (1990). *Moral consciousness and communicative action*. Cambridge, MA: The MIT Press.

Habermas, J. (1997). *Between facts and norms: Contributions to a discourse theory of law and democracy*. London, UK: Polity Press.

Hajian, S., Domingo-Ferrer, J., & Martinez-Balleste, A. (2011). Discrimination prevention in data mining for intrusion and crime detection. In *IEEE Symposium on Computational Intelligence in Cyber Security* (pp. 47–54). Paris, France: IEEE. http://doi.org/10.1109/CICYBS.2011.5949405

Hildebrandt, M. (2008). Defining profiling: A new type of knowledge? In M. Hildebrandt & S. Gutwirth (Eds.), *Profiling the European citizen* (pp. 17–45). Rotterdam, Netherlands: Springer.

Hinman, L. M. (2008). *Searching ethics: The role of search engines in the construction and distribution of knowledge*. Berlin, Germany: Springer.

Howard, P. N., & Kollanyi, B. (2016). Bots, #strongerin, and #brexit: Computational propaganda during the UK-EU referendum. Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2798311

Johnson, J. A. (2006). Technology and pragmatism: From value neutrality to value criticality. Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2154654

Kraemer, F., van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology*, *13*(3), 251–260. http://doi.org/10.1007/s10676-010-9233-7

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2765268

Leese, M. (2014). The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue*, *45*(5), 494–511. http://doi.org/10.1177/0967010614544204

Lisboa, P. J. (2013). Interpretability in machine learning: Principles and practice. In *10th Annual Conference on Fuzzy Logic and Applications* (pp. 15–21). Genoa, Italy: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-03200-9_2

Macnish, K. (2012). Unblinking eyes: The ethics of automating surveillance. *Ethics and Information Technology*, *14*(2), 151–167. http://doi.org/10.1007/s10676-012-9291-0

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. http://doi.org/10.1007/s10676-004-3422-1

Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics*, *22*(2), 303–341. http://doi.org/10.1007/s11948-015-9652-2

Naik, G., & Bhide, S. S. (2014). Will the future of knowledge work automation transform personalized medicine? *Applied & Translational Genomics*, *3*(3), 50–53. http://doi.org/10.1016/j.atg.2014.05.003

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of datification. *The Journal of Strategic Information Systems*, *24*(1), 3–14. http://doi.org/10.1016/j.jsis.2015.02.001

Pariser, E. (2011). *The filter bubble : What the Internet is hiding from you*. London, UK: Viking.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582–638. http://doi.org/10.1017/S0269888913000039

Rubel, A., & Jones, K. M. L. (2014). Student privacy in learning analytics: An information ethics perspective. Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2533704

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Auditing algorithms: Research methods for detecting discrimination on Internet platforms.* Paper presented at the 2014 International Communication Association Preconference on Data and Discrimination: Converting Critical

Concerns into Productive Inquiry, Seattle, WA. Retrieved from http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf

Schermer, B. W. (2011). The limits of privacy in automated profiling and data mining. *Computer Law & Security Review*, *27*(1), 45–52. http://doi.org/10.1016/j.clsr.2010.11.009

Stark, M., & Fins, J. J. (2013). Engineering medical decisions. *Cambridge Quarterly of Healthcare Ethics*, *22*(4), 373–381. http://doi.org/10.1017/S0963180113000224

Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*, *239*. Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20

Turilli, M. (2007). Ethical protocols design. *Ethics and Information Technology*, *9*(1), 49–62. http://doi.org/10.1007/s10676-006-9128-9

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105–112. http://doi.org/10.1007/s10676-009-9187-9

Tutt, A. (2016). An FDA for algorithms. Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2747994

Van Otterlo, M. (2013). A machine learning view on profiling. In M. Hildebrant & K. de Vries (Eds.), *Privacy, due process and the computational turn: Philosophers of law meet philosophers of technology* (pp. 46–64). London, UK: Routledge.

Woolley, S. (2016). Automating power: Social bot interference in global politics. *First Monday*, *21*(4). http://doi.org/10.5210/fm.v21i4.6161