Algorithmic Opacity: Making Algorithmic Processes Transparent through Abstraction Hierarchy

Pragya Paudyal, and B.L. William Wong
Middlesex University London
The Burroughs, Hendon, London NW4 4BT, England
{p.paudyal/w.wong} @mdx.ac.uk

ABSTRACT

In this paper we introduce the problem of algorithmic opacity and the challenges it presents to ethical decision-making in criminal intelligence analysis. Machine learning algorithms have played important roles in the decision-making process over the past decades. Intelligence analysts are increasingly being presented with smart black box automation that use machine learning algorithms to find patterns or interesting and unusual occurrences in big data sets. Algorithmic opacity is the lack visibility of computational processes such that humans are not able to inspect its inner workings to ascertain for themselves how the results and conclusions were computed. This is a problem that leads to several ethical issues. In the VALCRI project, we developed an abstraction hierarchy and abstraction decomposition space to identify important functional relationships and system invariants in relation to ethical goals. Such explanatory relationships can be valuable for making algorithmic process transparent during the criminal intelligence analysis process.

Keywords

Algorithmic transparency; Abstraction Hierarchy; opacity; transparency; ethical decision-making; Machine learning

INTRODUCTION

Criminal intelligence analysts have to deal with a large volume of often fragmentary pieces of information from which to understand a situation and to solve crime cases. Machine learning has helped by locating and extracting potentially relevant information through advanced data analytics. As many machine-learning techniques have been developed in criminal justice, medicine, finance, and other areas, to help in decision-making the general public demands transparency of the system so that they can ascertain the validity of the conclusions drawn from such black box computation. Algorithmic opacity is a condition where the internal workings of computational methods are hidden from the user. However, internal algorithmic processes are often so complex that it is also difficult for the designer to explain the techniques used to recommend or make decisions. We call this algorithmic decision-making, i.e. the process where we delegate decision making to an algorithm.

One of the widely used criminal risk assessment tools, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, has been widely used for predicting recidivism risk at court. Recidivism is defined as "the tendency of a convicted criminal to reoffend". This system predicts the defendant's risk of committing a felony within two years of assessment based on the individual's past criminal records, and 137 features concerning the individual. The features used by COMPAS do not include information such as race, ethnicity, or any other aspects of the data which may correlate to race; however, the scores produced appear to favor white defendants over black defendants by under predicting recidivism for white defendants and over predicting for black defendants. In addition, the predictions produced by the system are invariably inaccurate. Lack of transparency and little oversight of the inner workings of a system can erode the rule of law and diminish individual right.

Machine learning algorithms may be referred to as black boxes. From an ethical and justice viewpoint, the black box nature of machine learning algorithms can lead to the problem of automation surprise (Sarter, Woods, & Billings,1997). It typically refers to an action that is performed by technology where the outcome is unexpected by the users. The COMPAS issue raises some important questions, such as: How can longer sentencing of re-offenders be justified? How can black box automation and lack of transparency be avoided? And how can we be accountable for unethical legal decisions?

We have developed a system - VALCRI (Visual Analytics for Sense-making in Criminal Intelligence Analysis), designed as a next generation criminal intelligence analysis system based on a sense-making technology supported by advanced data processing and analytics software. VALCRI integrates machine-learning techniques for effective analysis of crime data. One of the goals VALCRI is to make the system transparent and visible to inspection in order to avoid the problem of black box automation surprises.

In this paper we use human factors principles to addresss the following issues: (1) How to avoid blackbox automation and lack of transparency; (2) Holding analysts accountable for unethical legal decisions. We hypothesise that human factors principles can be used to make the VALCRI system transparent and open to inspection in order to hold decision-makers accountable.

VALCRI: A COMPLEX SYSTEM

VALCRI facilitates human reasoning and analytic discourse by being tightly coupled with semi-automated human-mediated semantic knowledge extraction capabilities. VALCRI integrates machine learning to search for semantically similar data across structured and unstructured data in various use cases, such as comparative case analysis, associative search, maps and timeline analysis among others.

VALCRI operates in and exhibits characteristics of complex systems. Complex systems typically demonstrate high numbers of known and hidden interdependencies between components. Outputs from complex systems are often emergent, and is therefore difficult to know exactly which input contributes to an observed output (Ormand 2011). Complex systems exhibit several defining characteristics, such as feedback, strongly interdependent variables, and extreme sensitivity to initial conditions. VALCRI has many inter-related and inter-dependent components, such as automated knowledge extraction, text analytics, and self-evolving ontologies, based on crime profiles with many emergent outcomes, such as conclusions based on evidence assembled and constructed into explanatory narratives.

Sarter, Woods, & Billings (1997) explain that in complex domains, users have to deal with: (i) familiar events; (ii) unfamiliar but anticipated events; and (iii) unfamiliar and unanticipated events (Rasmussen, 1985). A major challenge for machine learning in VALCRI is dealing with unfamiliar and unanticipated situations.

PROCESSING STAGES IN MACHINE LEARNING ALGORITHMS

One important purpose of machine-learning algorithms is to enable analysis of massive quantities of data and enable humans to develop insights into making decisions and predictions. The data mining process consists of various steps; we will describe these stages briefly in this section.

(a) Data Pre-processing is an important initial step. When analyzing data it is necessary to make sure that the inputs are suitable for mining. The vast amount of data received by the police is collected from diverse and external sources. As a result, the initial quality of the data will be incomplete (missing values, lacking certain attributes, lacking features values, containing only aggregate data), noisy (duplication of the data, containing errors, outlier values), and consisting of inconsistent data. Data preparation involves data cleaning, data integration, data transformation, and data reduction.

(b) Data Mining is an automated analysis of data, using algorithms to find patterns and relations in data. Data mining is concerned with identifying patterns of characteristics and behavior based on historical data, which is often used for making predictive judgment. Clustering, classification, regression, and association are some of the common techniques used in data mining. Most of these techniques use numerical data for mining. If any data is in categorical form, it needs to be converted into numerical form; this transition impacts accuracy and outcome.

(c) Data Visualization is the process which allows the analyst to read and interpret data easily and quickly. Classic visualization techniques have been effective for small and intermediate size data. However, we face challenge when we apply classic visualization techniques to big data due several data points and dimensions (Tang, Liu, Zhang, & Mei, 2016). Projecting high-dimensional data into space with fewer dimensions is a challenging issue in data mining and machine learning. It is very important to preserve the intrinsic structure of high-dimensional data (Sacha et al., 2017).

Although different DR techniques have been developed, the problem of preserving the intrinsic structure of data is not yet fully resolved. Tang, Liu, Zhang, & Mei (2016) highlight some of issues where: (i) performance deteriorates when the dimensionality of the data grows; (ii) sensitivity to different data sets; and (iii) efficiency of the graph visualization step, which significantly declines when the size of the data increases. Moreover, a study conducted by Paudyal et al (2017) suggests that depending on the type of algorithm or the features you choose, the result varies. However, some analysts are not aware of these stages, or the undesirable consequences they may bring. These problems present many ethical issues, such as privacy, accuracy, integrity, and biased outcome.

ALGORITHMIC OPACITY

We define algorithmic opacity as a condition where algorithms lack visibility of computational processes, and where humans are not able to inspect its inner workings to ascertain for themselves how the results and conclusions were computed. Such computational modules are also referred to as "black boxes". Pasquale (2015) describes the black box as a system whose workings are mysterious. We know the input and output, but it is not possible to know how the results were processed and calculated.

The opacity of algorithm makes it difficult to scrutinize. As a consequence, there is a lack of clarity to the public in terms of how a certain decision was made (Diakopoulos, 2014), and potential incomprehensibility for human reasoning (Danaher, 2016). A wide range of ethical concerns, such as privacy, fairness, autonomy, bias, accountability, accuracy, discrimination has been discussed in the literature (Centre for Internet and Human Rights, 2015; Gillespie, 2012; O'Neil, 2016; Wagner, 2016; Ziewitz, 2015). The opacity of the machine learning algorithm inhibits over-sight (Burrell, 2016) of the automation surprises which have come about as a result of the obscured inner workings of the algorithm (Sarter, Woods, & Billings, 1997). The complexity and opacity of the algorithm make it difficult to understand if the decision made meets ethical requirements. In machine learning algorithms, it is not possible to assess the validity and the manner by which automation recommendations have come about.

The Palantir, Facebook and Cambridge Analytica scandal is a major example which highlights how companies are using people's data in unacceptable ways due to lack of transparency in the process. As Justice Louis Brandeis (cited in Pasquale, 2015) wrote - "sunlight is said to be the best of disinfectants (Brandeis, cited in Pasquale, 2015)." Likewise, transparency can be seen as a powerful solution for removing the opaqueness of the algorithm system.

Making algorithmic processes transparent is a challenging task. We are bound legally by Article 15 (4) and Recital 63 of the EU's General Data Protection Regulations (Information Commissioner Office, 2017) to respect the rights and freedom of others. Moreover, Burrell (2016) argues that explaining the internal logic of algorithmic workings to experts and non-experts alike is difficult because of the complexity of the computational system. Moreover, transparency allows special interest groups to act quickly and manipulate the code for dishonest reasons. As a result, algorithmic workings can bring about unfair outcomes to weaker population segments (Zarsky, 2013).

NEED FOR TRANSPARENCY IN VALCRI

During the design and development of the VALCRI system, the need for transparency was viewed from two aspects: (i) the lack of visibility arising from "black box" automation makes it difficult for end-users to be held accountable and to verify their decisions; and (ii) the need for analysts to show a paper trail leading to a particular conclusion. In addition, transparency in VALCRI is needed to comply with legislation; to build trust and accountability; to identify uncertainty and bias; and to make ethical decisions.

EXPLANATION: A POSSIBLE APPROACH?

Explanation is one approach for making computational processes transparent. Explanation facilitates insight in order to help the user make decisions and take action. Furthermore, analysts will be able to evaluate if the outcome had been reached by rational arguments, and does not conflict with ethical or legal norms. Explanation in machine learning is necessary to achieve trustworthiness, and for an evaluation of the ethical and moral standards of a decision. (Doshi-Velez & Kim, 2017) highlight the need for explanation: to understand why a system is not working as expected; to make sure the system is making sound decision; to provide explanation to make fair decisions. Keil (2006) reports that people require different levels of explanation depending on, among other factors, expertise, level of understanding of particular subject area, or cultural influence.

To make the VALCRI process transparent through explanation we investigated the use of the Abstraction Decomposition Space (ADS) and Abstraction Hierarchy (AH) to identify important functional relationships and system invariants in relation to ethical goals.

HOW TO IMPLEMENT EXPLANATION: ECOLOGICAL INTERFACE DESIGN (EID)

The main purpose of EID is to provide the user with a visual interface display that will allow the user to understand the complex relationship in an easy way. This allows the user to understand the constraints of the work environment, and how the action they take impacts reaching their objectives. The AH is a framework used to document analysis of complex socio-technical systems. According to Lintern (2013), the abstract dimension consists of an AH that is a diagram constructed through means-ends relations. This method shows how-why relations to each other (Naikar, 2013); use of "means-ends" relation in the VALCRI system will enable us to make visible the structural relationships according to different levels of constraints. When looking for a reason for why one decision was made over another, we tend to consider the holistic properties of a system at the higher level of the abstraction. However, the reason for a certain decision could be because of a different process within the system's component. As many components have an influence on certain outcomes, it is difficult to explain a particular property for an outcome. We conducted ADS and AH analysis of VALCRI based around the ML computational processes. The resulting ADS and AH models are presented below:

	System	Subsystem	Component

	VALCRI	**Data Extraction**
		Semi-Automated Semantic Knowledge Extraction
		Ontology, NLP
		Data Mining & Analytic
			Concept Classification Table;
			Similarity space selector; WOC;
			Concept graph;
			DOTS

Functional purpose	Aid LEAs to get insight of the data to solve the crime; Help find connection, LEAs often miss quickly and ethically	Data preparation for analysis process	Interactive visualizations that aid the analysts in their work
Values and priority measure	Conservation of information and information flow, information accuracy= information flow and techniques used	Quality, Accuracy	Fair, Accurate, Ethical
Purpose-related function		Provide quick overview of the concepts and underlying term that associates with each crime; Identify and group crime report according to their similarity ; Analyze the commonalities between crime cases in order to support reasoning and decision-making; Interactively explore and steer the computation to develop a task-driven similarity model; Help analyst to understand the characteristic of the data and cluster; Record analytic provenance with aim of capturing and evaluating user interaction; Graph representation for logic based on the semantic.	
Object-related purpose			Data collection, Data cleaning, Data transformation, Crime reports based ontologies, Measure the distance/ dissimilarities between crime cases, Weighted similarity metric, Use different DR algorithms to produce low-dimensional embedding of the data, Use of algorithms to cluster, feature selection
Physical object			Concept Classification Table, Similarity space selector, WOC, Concept graph, DOTS

Figure 1. Explanation of the VALCRI system using abstraction-decomposition space

The ADS organizes information in a systematic manner to provides a big picture of the system. The VALCRI system can be decomposed into three levels: the VALCRI system, subsystem, and components. At the VALCRI system level, the system is modeled as a single entity. The subsystem and component represents the detailed granularity of the system. The VALCRI system has five levels of abstraction. While the ADS describe the same system; the concepts at each level are comparatively distinct. Examining a system from a different level gives different conceptual viewpoints. Furthermore, the user will have a different understanding of the system based on their experience.

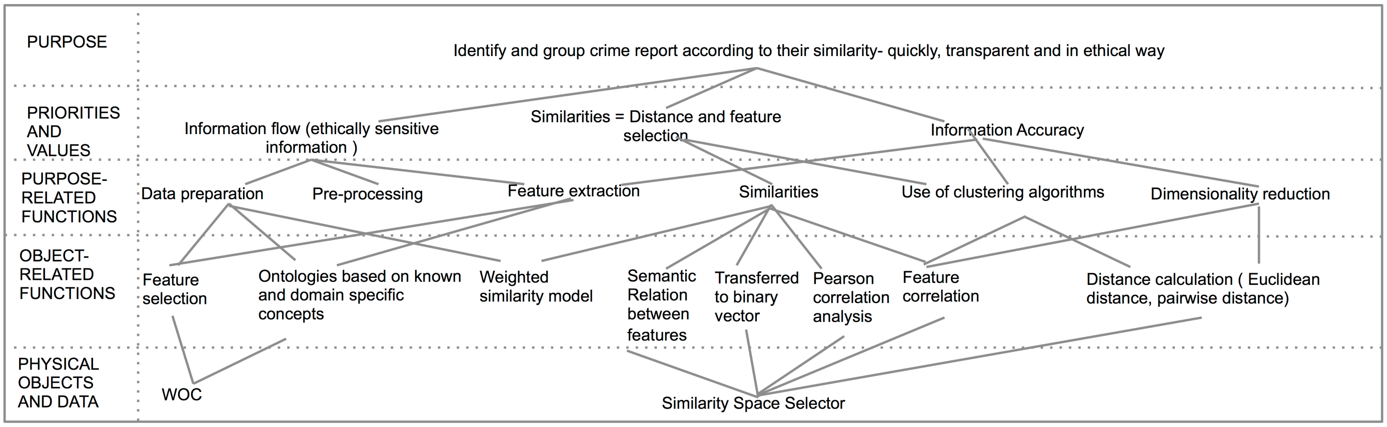
PURPOSE
Identify and group crime report according to their similarity- quickly, transparent and in ethical way
PRIORITIES AND VALUES
Information flow (ethically sensitive information)
Similarities = Distance and feature selection
Information Accuracy
PURPOSE-RELATED FUNCTIONS
Data preparation
Pre-processing
Feature extraction
Similarities
Use of clustering algorithms
Dimensionality reduction
OBJECT-RELATED FUNCTIONS
Feature selection
Ontologies based on known and domain specific concepts
Weighted similarity model
Semantic Relation between features
Transferred to binary vector
Pearson correlation analysis
Feature correlation
Distance calculation (Euclidean distance, pairwise distance)
PHYSICAL OBJECTS AND DATA
WOC
Similarity Space Selector

Figure 2. AH of part of VALCRI system

Explanation of the system from different levels shapes the user's mental model, which in turn influences the user's understanding of the system's limits and boundary constraints. We chose one purpose-related function of VALCRI - "Identify and group crime according to their similarities" - as a use case. Using AH methodology, we gained fine-grained explanation the different steps involved, in order to identify the relationship, and find a way of visualizing those relationships.

Abstraction Hierarchy

The AH for our specific use case is represented graphically in figure 2, and is outlined in the following sections:

Functional purpose: This level of abstraction corresponds to the rationale behind the design system. One FP of the VALCRI system can be described as to group crime reports according to their similarities. One of the important tasks during the investigation is to identify and group crime reports according to their similarity. The similarity of the report is based on the concepts or the features chosen. During the analysis process, analysts will receive millions of records and thousands of extracted features from each report. The goal is to identify similarities within the reports.

Abstract function: Usually, abstraction represents the criteria that must be respected for a system to achieve its functional purpose. Criteria are fundamental laws, principles, or values, which can serve as a basis for evaluation or judgment. The criteria that must be respected for the VALCRI to achieve "group crime report, according to similarities" includes ensure quality, few errors, effective data analysis, adhere to the ethical and legal values and finally, aid intelligence analyst in further investigation. Analyst can use the criteria at this level to evaluate how well the purpose-related functions are fulfilling its functional purpose. Abstraction functions allow analysts to reason from first principles. First principles are important when dealing with unanticipated situations. In the case of the VALCRI, may apply certain heuristics to ensure they are respecting ethical and social values when collecting, processing data.

Purpose- related function: This level represents the function that a system must be capable of supporting so it can satisfy the purpose function. Feature extraction, data preparation, and feature selection, dimensional reductions are some of the function that VALCRI must enable to achieve crime report according to their similarities. Miller and Vicente (cited in Naikar, 2013) argue that the purpose-related level can be viewed as describing the "uses" of the object related functions. Feature extraction points to the uses that selection of features, DR algorithms, and their distance calculation; feature correlation, semantic relation between features etc. serve in the VALCRI. In the VALCRI, purpose related function such as feature extraction; data preparation, and feature selection, dimensional reductions must be managed in a way that attains "crime report according to their similarities" within the bound of system's resources.

Object-related function: A system's object-related function serves to archive its purpose-related functions. In the VALCRI system, textual crime report enables the purpose-related function of feature extraction; in a similar way, the semantic relation between features, calculation, and visualization of feature characteristic, semantic relation between features, transferred to binary vector etc. enables the purpose related function of feature selection. Object-related functions are highly dependent on the properties of the physical objects.

Physical object: This level represents the physical objects of the system. In the VALCRI system, the representation includes information about each object. The physical object based on the specific use case is WOC and Similarity Space Selector. A system's physical object affords a system to achieve its purpose-related function. In the VALCRI system, selection of algorithm affordance to k-mean, PCA, MDS for dimensional reduction and visual clustering. These are the objects that analysts can change as a consequence the result received will vary. Reising (2000), argues that the physical objects represent the properties necessary for classification, identification and configuration for navigation in the system.

DISCUSSION AND CONCLUSION

In this paper we have briefly outline our approach of providing explanation to algorithmic opacity by using AH and ADS. When giving explanation of something we concentrate on the how and why question. The AH and ADS help in answering the how, what and why questions for the algorithmic process. When making decisions that are ethical, it is important to understand the process, potential positive and negative consequences. AH and ADS can be characterized by the How-What-Why triad of questions enabling analysts to think about the consequences. In the VALCRI system, analysts can choose the features, algorithms and number of clusters, how these choices can affect the outcome produced. Often artifacts of data collection and preprocessing can induce undesirable correlations that the algorithms pick up during data mining. Some of the features may be highly correlated with sensitive features such as race, ethnicity and religion etc. These issues are difficult to identify by just looking at the raw data and predictions. When analysts are using any system, analyst tends to consider holistic properties of a system at high level of abstractions (the main function of the VALCRI) in order to make sense of relationships at the lower levels of abstraction. Through this preliminary and exploratory investigation, we outline how ethically important functional relationships and system invariants may be identified.

AH is often used in the context of causal systems where the functional relationships between variables are known *a priori* before development. Whereas in intelligence analysis systems such as VALCRI, the functional relationships comprise interconnections between fragments of data that explain a situation, can only be constructed *post hoc,* while one is using VALCRI during an investigation. From this AH and ADS, we identified the relationship between the different stages within the ML process, rather than the investigative analysis process. There are a number of avenues of future work that we wish to explore. We seek to: improve the AH, and ADS representation of the machine learning computational processes; investigate how this approach helps in ethical decision making process; apply the semantic mapping and other representation design principles to devise an EID based on based on the functional relationship we identified to translate the VALCRI's key functional relationships into visual representations for ethical decision making.

ACKNOWLEDGEMENTS

The research reported in this paper has received funding from the European Union 7th Framework Programme FP7/2007-2013, through the VALCRI project (2014-2018) under grant agreement no. FP7-IP-608142, awarded to B.L. William Wong, Middlesex University London, and Partners.

REFERENCES

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. http://doi.org/10.1177/2053951715622512

Carol Ormand. (2011). Developing Student Understanding of Complex Systems in the Geosciences. Retrieved from https://serc.carleton.edu/NAGTWorkshops/complexsystems/introduction.html

Centre for Internet and Human Rights. (2015). The Ethics of Algorithms: from radical content to self-driving cars - Final Draft Background Paper. *GCCS*, (1), 1–18. Retrieved from https://www.gccs2015.com/sites/default/files/documents/Ethics_Algorithms-final doc.pdf

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, *29*(3), 245–268. http://doi.org/10.1007/s13347-015-0211-1

Diakopoulos, N. (2014). Digital Journalism Algorithmic Accountability. http://doi.org/10.1080/21670811.2014.976411

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. Retrieved from https://arxiv.org/pdf/1702.08608.pdf

Gillespie, T. (2012). The relevance of algorithms. *Media Technologies: Essays on Communication, Materiality, and Society*, (Light 1999), 167–194. http://doi.org/10.7551/mitpress/9780262525374.003.0009

Information Commissioner's Office. (2017). General Data Protection Regulation (GDPR), 43 pages. Retrieved from https://ico.org.uk/media/for-organisations/data-protection-reform/overview-of-the-gdpr-1-13.pdf

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–54. http://doi.org/10.1146/annurev.psych.57.102904.190100

Lintern, G. (2013). Tutorial: Work Domain Analysis, (1999). http://doi.org/10.1201/b14774

Naikar, N. (2013). *Work domain analysis : concepts, guidelines, and cases*. CRC Press.

O'Neil, C. (2016). *Weapons of Math Destruction*. Penguin UK.

Pasquale, F. (2015). The Black Box Society. *Cambridge, MA: Harvard University Press*, *36*, 32. Retrieved from http://www.hup.harvard.edu/catalog.php?isbn=9780674368279

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decisionmaking and system management. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*(2), 234–243. http://doi.org/10.1109/TSMC.1985.6313353

Reising, D. (2000). The Abstraction Hierarchy and its Extension beyond Process Control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(1), 194–197. http://doi.org/10.1177/154193120004400152

Sacha, D., Zhang, L., Sedlmair, M., Lee, J. A., Peltonen, J., Weiskopf, D., … Keim, D. A. (2017). Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 241–250. http://doi.org/10.1109/TVCG.2016.2598495

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. *Handbook of Human Factors and Ergonomics*, *2*, 1926–1943. http://doi.org/10.1207/s15327108ijap0204_5

Tang, J., Liu, J., Zhang, M., & Mei, Q. (2016). Visualizing Large-scale and High-dimensional Data. http://doi.org/10.1145/2872427.2883041

Wagner, B. (2016). Algorithmic regulation and the global default : Shifting norms in Internet technology. *Etikk I Praksis - Nordic Journal of Applied Ethics*, *10*(1), 1–9. http://doi.org/10.5324/eip.v10i1.1961

Zarsky, T. (2013). Transparent Predictions. *University of Illinois Law Review*, (4), 1503–1569.

Ziewitz, M. (2015). Governing Algorithms Myth, Mess, and Methods. *Science, Technology, & Human Values*.