Nicholas Diakopoulos, Northwestern University, nad@northwestern.edu
Whitepaper for Trustworthy Algorithms Workshop, October 6, 2017

Towards Algorithmic Transparency

As various facets of society come to adopt algorithmic decision making (ADM) it raises questions of how to maintain accountability and responsibility for those decisions, particularly when they can have negative impacts on individuals or society. One mechanism that can help expose processes, algorithmic or otherwise, is transparency. Transparency entails the disclosure of particular pieces of information so that interested stakeholders can monitor, check, criticize, or intervene in those processes[1]. For applications such as safety or health, targeted transparency policies (i.e. regulations) specify the precise pieces of information that must be made available publicly, such as crash test safety ratings for cars, nutritional information for food, or cleanliness grade ratings for restaurant inspections[2]. Transparency is an enabler of, but not *the* unitary solution to algorithmic accountability.

I would argue that there is substantial value for thinking through what types of transparency policies may be effective when applied to algorithms. Transparency as a strategy presents a pragmatic approach for thinking about the spectrum of information that might be disclosed about an algorithm, and the ways in which that disclosed information may in turn affect stakeholder behavior. Initial research in this area has enumerated a model or palette of potential information that might be disclosed about an algorithm, including across aspects of data, model, inference, and interface[3]. Of course, transparency information by itself does not *do* anything, but its creation and publication can impact behavior based on what people understand from that disclosed information. Such effects may include, among other things: changing algorithmic development processes and behavior in order to produce transparency information (which could engender a more ethical approach to engineering)[4], setting functional expectations according to disclosed information (which could perhaps improve usability), and providing the ability to raise awareness when there is a violation of expectations. Raising awareness for when ADMs produce decisions that are at odds with expectations set by transparency disclosures provides a mechanism for accountability -- the compulsion of an explanation to detail the nature of that deviation. Still, transparency is not a panacea. There are, for example, risks related to manipulation of systems using disclosed information, as well as the possibility for evasion by steering around known disclosure requirements[5].

A question that is still very much open is the extent to which transparency information may impact individual perceptions such as trust. Some research indicates that in situations when algorithms violate personal expectations for system behavior, additional explanation of system decisions can serve to repair trust in the system[6]. Importantly, such transparency information will not be attended to by *most*

---

[1] Deuze, M., 2005. What is journalism?: Professional identity and ideology of journalists reconsidered. Journalism, 6(4), pp.442–464.
[2] Fung, Archon, Mary Graham, and David Weil. 2007. Full Disclosure: The Perils and Promise of Transparency. New York: Cambridge University Press
[3] N. Diakopoulos and M. Koliska. Algorithmic Transparency in the News Media. Digital Journalism. 2016.
[4] N. Diakopoulos, S. Friedler. How to Hold Algorithms Accountable. MIT Technology Review. https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/
[5] Ananny, M. & K. Crawford (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society.
[6] Kizilcec, Rene F. 2016. "How Much Information?: Effects of Transparency on Trust in an Algorithmic

individuals, but only those with a stake in paying additional attention to a system due to some break with expectation or negative ADM affecting them personally. Transparency information may also be attended to by actors in civil society such as journalists or others who have professional roles associated with monitoring ADMs.

Algorithmic transparency is going to look different in different use cases and domains. For instance, in cases where personal liberty at stake, such as criminal justice (an area I would consider a high-stakes decision for an individual) transparency policies may include provisions to compel code detailing how an algorithm is implemented. When there is reason to believe that an implementation may be faulty or not adhere to standard expectations of operation, such disclosure may provide additional trust and accountability of the system[7]. Such a scenario does not suffer from the same concerns over manipulation or gaming as in other scenarios such as, say, online media where code-level disclosure may be unnecessary. In such cases, transparency information related to benchmarks may be more appropriate. For instance, it is not important to me to see the code for how Twitter identifies bots that are manipulating information on their platform, but it is instead important to see benchmarks for their overall accuracy for detection and expulsion, as well as how that activity is trending over time.

Within journalism, where I have spent a majority of my time thinking about algorithmic accountability and transparency, the demands for transparency vary quite a bit between, for instance, an editorial project that publishes data and code, and a product like a news app or website that algorithmically shapes exposure to news. Demands differ depending on the degree to which there is a human in the loop, and the degree to which there are business concerns surrounding disclosure. In the case of investigative work, data journalists tend toward transparency because the credibility of their knowledge claims depends on the veracity of those claims. This operates similarly to the way academics describe methods to buttress interpretations of scientific data -- no one would believe the results if a scientist didn't show or at least describe the methods. The need for transparency is also greater in fully autonomous systems that publish information directly gleaned from models. In contrast, if a journalist uses an algorithm internally to a newsroom to find a lead, which they then carefully check and do hours of additional reporting on, then the journalist can reasonably stand behind the published information and may not feel compelled to disclose the details of the algorithm that informed their process. This is all to argue that the way algorithmic transparency is implemented needs to be sensitive to human factors, and may vary substantially even within the same domain according to the task for which an ADM is being employed.

There are many open research questions that need to be pursued if effective algorithmic transparency policies are to be crafted for specific scenarios. Information modelling and user research needs to consider how disclosed transparency information may be used or acted upon. Threat models should consider how a piece of information might allow an algorithm to be gamed, manipulated, or circumvented. Pragmatic questions about the cost, effort, and time required to produce transparency information will need to be answered. And effective presentation strategies will be needed so that disclosed information can be presented to users in salient ways that are appropriately placed for their own decision making processes.

Interface." Proceedings Conference on Human Factors in Computing Systems (CHI)

[7] Lauren Kirchner. ProPublica Seeks Source Code for New York City's Disputed DNA Software. ProPublica. September, 2017. https://www.propublica.org/article/propublica-seeks-source-code-for-new-york-city-disputed-dna-software