Cover page

Title: Value Sensitive Algorithm Design

Author:

Haiyi Zhu, University of Minnesota, Twin Cities (Will Attend)

Loren Terveen, University of Minnesota, Twin Cities (Will Attend)

Keywords:

Algorithmic systems, automated or artificially intelligent decision making, Value Sensitive Algorithm Design

Abstract:

The most widely-used approaches to developing automated or artificially intelligent algorithmic systems are Big Data-driven and machine learning-based. However, there are two notable reasons why these approaches might fail: (1) Machine learning models often are based only on a single *prediction target*, which often is unable to capture the wide range of factors typically involved in any real-world problem. (2) Big Data-driven approaches rely largely on *historical* human judgements, which fail to capture and incorporate human insights into how the world can be improved in future. We propose a novel approach to the design of algorithms, which we call **Value-Sensitive Algorithm Design**. The Value Sensitive Algorithm Design method incorporates stakeholders' tacit knowledge and explicit feedback in the algorithm creation process. This increases the chance to avoid biases in design choices and compromises of key stakeholder values. Generally, we believe that algorithms should be designed to balance multiple stakeholders' needs, motivations and interests, and help achieve important collective goals. In our presentation, we shall describe our specific project "Designing Intelligent Socialization Algorithms for WikiProjects in Wikipedia" to illustrate our proposed methodology. We will illustrate our approach in some detail and report initial findings and lessons learned. We hope that this presentation contributes to the rich ongoing conversation concerning the use of algorithms in supporting critical decision-making in our society.

Overview

Automated or artificially intelligent algorithmic systems are assisting humans to make important decisions in a wide variety of critical domains. Examples include: helping judges decide whether defendants should be detained or released while awaiting trial (Corbett-Davies et al. 2017; Kleinberg et al. 2017); assisting child protection agencies in screening referral calls (Chouldechova et al. 2018); and helping employers to filter job resumes (O'Neil 2016).

The most widely-used approaches to develop decision-making or decision-supporting algorithms are driven by Big Data and are machine learning-based. The first step in the process is to define a *prediction target*. In relation to the examples cited above, this might consist of whether the defendant will commit a crime if released; whether the child will be removed from the home and placed in care, or whether a job applicant will receive or accept a job offer and be retained for a long time. The second step in the process of developing the algorithm is to use *historical data*, often in large volumes, for the purpose of training and validating the machine learning models. Finally, the validated models are applied to new data from incoming cases in order to generate predictive scores.

However, there are two main reasons why algorithms based on machine learning and Big Data might fail: (1) The model is often based only on a single *prediction target*. This might be an unreliable proxy for an unobserved or difficult-to-observe outcome such as severe maltreatment of children or suitability of applicants to jobs (Chouldechova et al, 2018). More important, **the single prediction target is often unable to capture the wide range of factors typically involved in any real-world problem.** (2) The Big Data-driven approach relies largely on *historical* human judgements, which are subject to historical stereotypes, discrimination, and prejudices. Using history to inform the future runs the risk of reinforcing and repeating historical mistakes and **fails to capture and incorporate human insights on how the world can be improved in future.**

We propose a novel approach to the design of algorithms, which we call **Value-Sensitive Algorithm Design**. Our approach is inspired by and draws on Value Sensitive Design (Friedman et al. 2013) and the participatory design approach (Muller & Kuhn 1993). We propose that the Value Sensitive Algorithm Design method should incorporate stakeholders' tacit knowledge and insights into the abstract and analytical process of creating an algorithm. This helps to avoid biases in the design choices and compromises of important stakeholder values. Generally, we believe that algorithms should be designed to balance multiple stakeholders' needs, motivations and interests, and help achieve important collective goals.

In the presentation, we shall describe our project "**Designing Intelligent Socialization Algorithms for WikiProjects in Wikipedia"** to illustrate our proposed methodology. We shall report on the lessons learned during the nine-month design and deployment period of the project, explain our approach to conducting value-sensitive design, and describe our attempts to address several outstanding challenges.

Designing intelligent socialization algorithms for Wikiprojects in Wikipedia

Retaining and socializing newcomers is a crucial challenge for the Wikipedia community. The number of active contributors in Wikipedia has been declining steadily since 2007, due at least in part to a sharp decline in the retention of new editors (Halfaker et al. 2012). WikiProjects, which are groups of contributors who work together as a team to improve Wikipedia, serve as important **socialization hubs** within the Wikipedia community. Prior work (Forte et al. 2012; Zhu et al. 2012b) suggests that WikiProjects provide three valuable support mechanisms for new members: (1) enabling them to locate suitable assistance and expert collaborators; (2) guiding and structuring their participation by organizing to-do lists, initiatives such as "Collaborations of the Week" and various task forces, and (3) offering new editors "protection" for their work, by shielding it from unwarranted reverts and edit wars**.**

However, matching new editors to WikiProjects is no trivial task. In the English version of Wikipedia alone there are currently about 2000 WikiProjects, and in an average month 38,628 new users register on the site. The goal of our research project is therefore to create algorithmic tools to **match new Wikipedia contributors to WikiProjects.** We are conducting this research using our Value-Sensitive Algorithm Design methodology, which consists of the following five steps:

- **Conduct empirical studies to understand stakeholders' motivations; the values and goals of importance to them, and potential trade-offs.** To achieve this, we reviewed prior studies and conducted our own research with Wikipedia editors in order to answer the following questions: What motivates new editors to participate in a WikiProject? What motivates WikiProjects to recruit new members? What collective outcomes are important for WikiProjects and for Wikipedia in general?
- **Identify algorithmic approach and develop algorithm prototypes.** We used the results of the first step to develop matching algorithms that: 1) satisfy the goals of new editors and WikiProjects by considering the match between the editor's interests and the project topic; 2) satisfy the goals of WikiProjects by excluding editors likely to produce low-quality work; and 3) satisfy collective outcomes by targeting new editors that are not only likely to be high-quality contributors, but can also help close Wikipedia's gender gap and improve topic coverage. We used a parallel prototyping approach and created four different types of algorithms: rule-based, category match, bond-based, and collaborative filtering.
- **Define methods for working with the community.** Online communities like Wikipedia constitute a rich laboratory for research: they make collaboration, social interaction, and production processes visible, and offer opportunities for experimental studies. However, there is also an unfortunate tradition of academic researchers treating these simply as platforms for their studies, rather than real communities with their own norms, values, and goals. Wikipedia studies often encounter resistance from Wikipedia editors, and may sometimes be halted before completion. We avoided these problems by working with stakeholders to develop a research protocol that is acceptable to the community. We essentially developed our research plan and algorithmic approach *in the Wiki.* In other words, we are developing and deploying our algorithms not just for, but with, the Wikipedia community.
- **Iterative and gradual refinement.** We iteratively tested and refined our algorithms. Over a duration of two months we sent weekly batches of recommendations to pilot test participants, conducted short surveys to seek their views on these and also interviewed project organizers and newcomers. We used their feedback to make significant changes to the design of the algorithms. After completing the pilot, we waited three months before declaring it a success and proceeding to the next step, to allow time for any unanticipated effects to manifest.
- **Continuous monitoring of short- and long-term effects.** We conducted experiments to systematically examine the short- and long-term effects of the algorithms. During the course of our experiments we maintained a dashboard of metrics, including indicators of important outcomes. We reviewed the dashboard regularly to ensure that progress was being made towards the goals and that there were no unintended consequences.

Discussion Points

During the research process, we identified a number of key challenges involved in conducting value-sensitive algorithm design. At the HCIC presentation, we will report on our efforts to date, some initial findings, and the lessons learned. At certain points, however, we might *offer more questions than answers*. Our hope is that this presentation contributes to the rich ongoing conversation concerning the use of algorithms in supporting critical decision-making in our society. In particular, we would like to pose the following questions for discussion:

1) How can we systematically translate stakeholders' concerns, interest and values into computational representations?

2) How can we address the gaps in technical literacy between most users/stakeholders (e.g. Wikipedia editors) and the algorithm designers? How can we explain and articulate our algorithms in ways that enable users and stakeholders to assess these and provide feedback for the purpose of improving them? What are effective and acceptable divisions of responsibility between algorithms, their designers, and community stakeholders?

3) How can we evaluate the algorithms created through the value-sensitive design approach?

4) What are the **trade-offs** (regarding efficiency, accuracy, fairness, acceptance, trust etc.) between the value-sensitive algorithm design approach, the Big Data-driven approach, and the pure human decision approach? How can we combine these different approaches in order to achieve the best outcomes?

Reference:

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic decision making and the cost of fairness." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797-806. ACM, 2017.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human decisions and machine predictions." *The Quarterly Journal of Economics*133, no. 1 (2017): 237-293.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." In *Conference on Fairness, Accountability and Transparency*, pp. 134-148. 2018.

O'Neil, Cathy. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

Friedman, Batya. "Value-sensitive design." interactions 3, no. 6 (1996): 16-23.

Muller, Michael J., and Sarah Kuhn. "Participatory design." *Communications of the ACM* 36, no. 6 (1993): 24-28.

Halfaker, Aaron, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. "The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline." *American Behavioral Scientist* 57, no. 5 (2013): 664-688.

Forte, Andrea, Niki Kittur, Vanessa Larco, Haiyi Zhu, Amy Bruckman, and Robert E. Kraut. "Coordination and beyond: social functions of groups in open content production." In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 417-426. ACM, 2012.

Zhu, Haiyi, Robert Kraut, and Aniket Kittur. "Organizing without formal organization: group identification, goal setting and social modeling in directing online production." In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 935-944. ACM, 2012.