Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2017: 1–12

DOI: 10.1177/2053951717718854
journals.sagepub.com/home/bds
SAGE

Toward human-centered algorithm design

Eric PS Baumer

Abstract

As algorithms pervade numerous facets of daily life, they are incorporated into systems for increasingly diverse purposes. These systems' results are often interpreted differently by the designers who created them than by the lay persons who interact with them. This paper offers a proposal for human-centered algorithm design, which incorporates human and social interpretations into the design process for algorithmically based systems. It articulates three specific strategies for doing so: theoretical, participatory, and speculative. Drawing on the author's work designing and deploying multiple related systems, the paper provides a detailed example of using a theoretical approach. It also discusses findings pertinent to participatory and speculative design approaches. The paper addresses both strengths and challenges for each strategy in helping to center the process of designing algorithmically based systems around humans.

Keywords

Human-centered design, interpretation of algorithms, performance metrics, theory, speculative and critical design

> *Premature optimization is the root of all evil (or at least most of it)*. (Knuth, 1974: 671)

Algorithms are designed to perform. A given algorithm is considered better when its results show an improvement according to some agreed-upon performance metric. These metrics assess aspects such as accuracy (e.g., which algorithm correctly identifies more spam email, a Naïve Bayes classifier or a support vector machine?), speed (e.g., which sorts faster, insertion sort or bubble sort?), or computational resources (e.g., which requires less bandwidth, a Skype conversation or a Google hangout?). Such metrics establish agreed-upon goals against which algorithm designers can compare their results.

These metrics' primacy, however, can also preclude other aspects of an algorithm's performance. Significant work has documented disconnects between the functioning of algorithmically based systems and the social interpretations thereof (e.g., Ananny, 2011; Eslami et al., 2015; Gillespie, 2011; Rader and Gray, 2015). Among the many factors involved, the metrics used to assess algorithms' performance are based on those aspects that are most readily computationally quantifiable. As a result, those metrics may or may not closely align with human and lay interpretations of what said algorithms do and mean.

This paper suggests addressing these disconnects through a practice termed *human-centered algorithm design* (HCAD). This practice applies techniques from human-centered design to the technical components of algorithmic systems. This paper presents the author's experiences designing, implementing, and evaluating several algorithmically based systems around issues related to political framing (Chong and Druckman, 2007; Entman, 1993). From these experiences, the paper distills three strategies by which algorithm design could become more human centered.

First, *theoretical* approaches can incorporate a wealth of concepts and theories from behavioral and social sciences on topics pertinent to algorithmic systems. While some work has used such an approach in research contexts (e.g., Hopkins and King, 2010;

Lehigh University, USA

Corresponding author:
Eric PS Baumer, Lehigh University, Computer Science & Engineering, 379 Packard Lab, 19 Memorial Dr W, Bethlehem, PA 18015, USA.
Email: ericpsb@lehigh.edu

Murnane and Counts, 2014; Recasens et al., 2013), it is less often used for developing the algorithmic components of interactive systems. This paper will draw on the authors' own work in the area of political framing (Baumer et al., 2015a) to demonstrate this approach, describing both successes and limitations (cf. Agre, 1997).

Second, current conceptions of what computation systems can do or should do may limit the design of such systems. Thus, *speculative* approaches may provide a means of overcoming how we currently conceive of algorithmically based systems. Drawing on insights from a field study led by the author (Baumer et al., 2014b), the paper will describe how speculative design (Baumer et al., 2014a; Bleecker, 2009; Blythe, 2014; Dunne and Raby, 2013; Linehan et al., 2014; Sterling, 2005; Tanenbaum, 2014) can be used to explore various alternatives for living with algorithms.

Finally, both theoretical constructs and speculative design can at times be removed from human practice. Thus, *participatory* approaches offer another avenue. Participatory design has as established history, both in technology design and elsewhere (Ehn, 1988; Muller and Druin, 2012; Muller and Kuhn, 1993). Using data from the same field study (Baumer et al., 2014b), this section describes both some potential challenges and unique opportunities with participatory approaches.

These three approaches do not exhaustively map the space of possible strategies. Rather, they provide a sense for the range of possibilities. The discussion considers the potential and limits of these approaches in addressing the relationship between the technical and social dimensions of algorithms. The paper concludes with some considerations about the practicality of following such approaches.

Related work: The interpretation of algorithmic systems

Misalignments often occur between the functioning of algorithmic systems and how those systems are interpreted. The term "algorithm" itself only recently entered the public consciousness (Sandvig, 2014), due in part to increasing attention in popular media. For instance, Facebook regularly tweaks the algorithm that curates users' news feeds to test how different variants impact subsequent behavior. This fact was pointedly highlighted in study of emotional contagion (Kramer et al., 2014). Despite this practice having existed for quite some time previously, the Kramer et al. study resulted in significant backlash, including some calls to try abandoning Facebook (Baumer et al., 2015b).

Numerous other examples exist. At one time, the Google Play store page for Grindr, a dating app for gay men, recommended as similar an app to determine if the user lives near a sex offender (Ananny, 2011). Flickr's automatic photo tagging been observed suggesting that African-American faces be labeled as "apes" (Hern, 2015). Google's search correction feature has been noted to ask those who search for "she invented" if they actually intended to search for "he invented" (Lenssen, 2007a, 2007b; Zimmer, 2007). Searches for African-American names are more likely to show ads suggesting that that person has an arrest record (Barocas and Selbst, 2016; Sweeney, 2013).

From a design perspective, these and other similar examples would likely have been difficult to predict, perhaps even impossible. When confined to a laboratory or scientific setting, results along these lines could be seen as aberrant and interpreted as such. However, when algorithms become incorporated into interactive, public-facing systems, their results become interpreted by people who likely have less knowledge of the technical implementation details of these algorithms. In such situations, what becomes most striking is not the results themselves but what they are interpreted to mean about, e.g. deviant behaviors (Ananny, 2011), social norms around gender and race (Barocas and Selbst, 2016; Lenssen, 2007a; Sweeney, 2013; Zimmer, 2007), or your "true" friendships (Eslami et al., 2015).

Furthermore, these lay interpretations are rarely, if ever, accounted for during the development of such algorithms. Lay interpretations often enter the picture only after an algorithmically based system is publicly released. The process of algorithm design is driven primarily by technical constraints, performance on benchmarks, etc. One of the goals in the work proposed here is to link these lay conceptions of algorithmic analysis into the design process for systems driven by algorithms.

The case for *human*centered

This paper intentionally avoids use of the term "user centered." No one would dispute the value that user-centered design has brought to human–computer interaction (HCI) and related fields. However, the process of design should not hinge entirely on the construct of "the user" (Redström, 2006) but on potential relationships to, with, and through technology. Those impacted in the above cases—gay men who are implicitly associated with sex offenders (Ananny, 2011), African-Americans who are suggested as having an arrest record (Barocas and Selbst, 2016; Sweeney, 2013), and women whose pregnancies are predicted based on their purchase histories (Baumer, 2015; Duhigg, 2012)—would not be included among users of algorithmically based systems. Although the examples described in the case study below focus primarily on users of the systems being designed, the strategies

distilled from these experiences could apply to a wide variety of different subject positions in relation to algorithmically based systems (Baumer and Brubaker, 2017).

Computational supports for frame reflection

The example systems in this paper come from a larger project exploring computational approaches to identifying the language of political framing (Chong and Druckman, 2007; Entman, 1993). The goal was not analytic in nature; this project did not involve identifying frames per se. Rather, the goal was reflective. We[1] wanted to develop algorithmically based interactive technologies that could draw attention to, and promote critical thinking about, framing, i.e. frame reflection (Schön and Rein, 1994).

Background: Political framing

"Facts have no intrinsic meaning. They take on their meaning by being embedded in a frame [. . .] that organizes them and gives them coherence" (Gamson, 1989: 157). Frames can be invoked by "keywords, stock phrases, stereotype images, sources of information" (Entman, 1993: 52) and "metaphors, exemplars, catchphrases, depictions, and visual images" (Gamson and Modigliani, 1989: 3; Price et al., 2005). These and other linguistic or rhetorical devices provide an interpretive lens or "package" (Gamson and Modigliani, 1989) through which to perceive and make sense of facts or events.

Most work on framing essentially compares the impact of different frames, exposing study participants to two (or more) different frames and assessing the impact of those frames on participants' opinions (Brewer, 2002; Druckman et al., 2012; Druckman and Nelson, 2003; Hart, 2011; Maibach et al., 2010; Quattrone and Tversky, 1988; Schuldt et al., 2011; Sniderman and Theriault, 2004; Tversky and Kahneman, 1981).

A complementary line of work seeks to shift framing from a source of subconscious influence to the focus of conscious inquiry. In what Schön and Rein (1994) call *frame reflection*: "assumptions, views of the world, and values that have heretofore remained in the background, giving shape to foreground inquiry but keeping, as it were to the shadows, become foreground issues, open to discussion and inquiry in their own right" (Rein and Schön, 1996: 94).

While it can result in more productive dialog, engaging in frame reflection is no mean feat. Framing is so effective in part because it operates largely subconsciously (Gamson and Modigliani, 1989; Lakoff and Turner, 1989). It can be difficult to notice that a frame is even present, let alone interrogate it critically or explore alternatives.

The project described here asked: can we leverage data-intensive techniques from natural language processing and computational linguistics to help bring the language of framing to conscious attention? Doing so, we suggested, could help serve as a scaffold for frame reflection. The remainder of this section recounts experiences with two such systems, the strategies employed in their design, and those strategies' efficacies.

Frame reflection in action: FrameCheck

Imagine that you are reading a news article online. While reading that article, a web browser plug-in highlights a few key words and phrases most related to framing. This subsection describes work related to a natural language classifier designed to perform just this task (Baumer et al., 2015a).

The design process explicitly employed a theoretical strategy. We began by inventorying the linguistic and rhetorical devices mentioned in sociological, political, or communication work related to framing (Brewer, 2002; Chong and Druckman, 2007; e.g., Entman, 1993; Fairclough, 1999; Gamson and Modigliani, 1989; Price et al., 2005). We also conducted an in-lab study where we asked human participants to read an article and highlight the terms most relevant to framing. In debriefing interviews, participants then explained which terms they highlighted, which they did not highlight, and their reasoning behind each (for details, see Baumer et al., 2015a).

The question became: how would we (or could we even) develop computational analogs for each of these linguistic or rhetorical devices? Some were fairly straightforward. For example, keywords and catchphrases (Entman, 1993; Gamson and Modigliani, 1989) can be operationalized as n-grams, i.e. one-, two-, or three-word phrases that consistently occurred together. Grammatical construction, e.g. active versus passive voice (Fairclough, 1999), can be identified using existing parsing tools (De Marneffe et al., 2006). Some devices required more sophisticated techniques, drawing on computational work for identifying, e.g. metaphorical language (Turney et al., 2011). Others had to be abandoned entirely. For example, we could find no suitable algorithmic technique for identifying stereotyped language.

From a technical standpoint, this strategy proved relatively successful. We were able to create a classifier that, when tasked with identifying framing in political news articles, agreed with human annotators about as often as the human annotators agreed with one another. Specifically, the classifier's average F1 scores

(using 10-fold cross validation) were statistically indistinguishable from average F1 scores comparing each human annotator against the others. The results warranted publication in a top computational linguistics venue (Baumer et al., 2015a).

From a theoretical standpoint, the work also made some interesting contributions. Prior work in political communication had noted that "straightforward guidelines on how to identify [. . .] a frame in communication do not exist" (Chong and Druckman, 2007: 106). The results of our experiments suggest that the presence of framing could be effectively identified using only lexical features (i.e., n-grams). That is, word choice mattered more than grammatical construction, figurative language, or other aspects suggested as important by the theoretical literature. This system was trained on data collected from lay annotators, though. Analysis of framing annotation by expert trained coders (e.g., Card et al., 2015) could yield different results.

The theoretical strategy. These connections demonstrate the potential of a theoretical approach to HCAD. Behavioral and social sciences provide numerous theories to help understand how people work. The above example shows how such theories can serve at least two valuable functions within algorithm design.

First, theories can be prescriptive. Most data sets include a potentially overwhelming number of possible features. About a given individual, one might know age, gender identity, height, weight, personality, drug use, annual income, charitable donations, frequency of physical activity, cholesterol levels, purchasing histories, credit score, marital status, parental status, etc. One could derive similarly myriad features from our annotated data about framing. When developing a model, how does one choose which features to include in the model? A naïve approach would simply test every possible feature. Not only might this prove computationally intractable, but testing more features increases the likelihood of identifying spurious relationships. A data-centric approach would likely emphasize feature selection (Guyon and Elisseeff, 2003), which quantitatively determines which features in a model are most informative. Again, though, feature selection provides less guidance as to *why* a given feature should be included in, or excluded from, a model. Instead, FrameCheck explicitly and clearly drew on social scientific theories of framing, augmented by a human subjects study, to inform feature selection. This example demonstrates how theory can be used prescriptively to guide algorithm design.

Second, theories can be descriptive. Given the results produced by an algorithm, what do they mean? Theories can help sort through potential interpretations for the results of such systems. The evaluation of FrameCheck was specifically arranged to allow for testing different theoretical hypotheses about which linguistic features of framing mattered most. By comparing the algorithm's performance against data collected from human annotators, the results were able to speak back to these open questions about how to identify framing (Chong and Druckman, 2007).

While leveraging social and behavioral theories can be beneficial, the nature of the transformations necessary for their incorporation in algorithmic analysis also poses serious challenges (Agre, 1997). For instance, the technique used here to identify metaphorical language focuses on mismatches in levels of abstraction between a term and its modifier. The phrase "dark paint" involves a concrete modifier and a concrete object, while the phrase "dark thoughts" matches a concrete modifier with an abstract object; the latter is thus more likely metaphorical (Turney et al., 2011).

Low-level Details about Behavioral Theories

Such codifications happen with virtually any kind of categorizing representational system, computational or otherwise (Bowker and Star, 1999). However, as those representations become the subject of algorithmic manipulation, they can become farther and farther removed from an intuitive understanding of the phenomena they are meant to capture. Burrell (2016) describes the application of deep learning techniques to computer vision tasks, such as identifying handwritten digits 0 through 9. The resultant neural networks cue in on visual features that bear little apparent resemblance to how humans perceive their own process of telling the difference between, say, a 1 and a 7.

A similar situation occurs in FrameCheck's use of lexical features. The most influential values for this feature show that words in close proximity to prepositions or conjunctions—"to," "and," "in," "of," etc.—are most likely to be classified as invoking framing. While they resonate strongly with comments made during our formative human subjects study, such patterns differ significantly from the keywords and catchphrases (Entman, 1993; Gamson and Modigliani, 1989) that these lexical features were intended to identify. Burrell (2016) notes difficulties in interpreting the complex and sometimes inscrutable features identified by algorithms. The example here shows how even seemingly simple features, such as the specific words used in a sentence, can be algorithmically transformed into something other than the theoretical attribute they are meant to represent.

This situation raises some serious difficulties for theoretical approaches. Social science's extant commitment to theory-driven work often requires that researchers formulate and test hypotheses within a theoretical framework, e.g. framing (Chong and Druckman, 2007; Entman, 1993; Gamson and Modigliani, 1989). The algorithmic transformation of theoretical concept to

computationally identifiable feature problematizes, or perhaps even renders impossible, this kind of theoretical hypothesis testing. If the lexical features to which an algorithm attends do not clearly map onto the keywords and catchphrases described in theory (Chong and Druckman, 2007; Entman, 1993; Gamson and Modigliani, 1989), what can we say conclusively about their importance in identifying framing?

Collectively, these points highlight differences between the manner in which human centering occurs during the *design* process and during the *evaluation*. Despite drawing on social scientific concepts, the classifier was still evaluated using traditional performance measures for machine learning algorithms, namely accuracy, precision, recall, and F1 score. Optimizing these performance metrics, however, may elide the very attributes on which the theoretical strategy hinges. Furthermore, while likely necessary for publication in technical venues, such performance metrics may highlight different kinds of issues than evaluation with human users. For instance, an overly aggressive classifier may achieve high recall, finding most of the words related to framing, but highlight so many terms that it becomes overwhelming. Such examples suggest that, while valuable, metric-based performance evaluation should not and cannot serve as the ultimate evaluation in human-centered approaches to algorithm design. As an alternative, the next subsection shows how field studies with human users can elucidate such nuances about the real-world functioning of algorithmic systems.

Frame reflection on action: Reflext

The design of a second system began with a similar theoretical motivation: to identify and draw attention to patterns of language relevant to political framing (Chong and Druckman, 2007; Entman, 1993; Gamson and Modigliani, 1989) as a means of supporting frame reflection (Schön and Rein, 1994). However, this design process followed a less prescriptive, more speculative approach (Baumer et al., 2014a; Bleecker, 2009; Blythe, 2014; Dunne and Raby, 2013, 2001; Linehan et al., 2014; Sterling, 2005; Tanenbaum, 2014). What if the system avoided technically codifying commitments to particular linguistic or rhetorical devices? What if the implementation embodied fewer assumptions about what text means, either denotatively or connotatively? What if, instead, the algorithm was designed to present prevalent language patterns but to leave their exact meaning open to interpretation (Pinch and Bijker, 1987; Sengers and Gaver, 2006)? At some level, we as system designers needed to choose which patterns of language the system would identify. However, we stopped short of ascribing any particular meaning or interpretation to those patterns, leaving that task in the hands of those using the system.

This approach resulted in an interactive text visualization called Reflext (Baumer et al., 2014b). This section first presents a technical overview of how Reflext works, then describes a subset of users' experiences that arose during our roughly six-month field trial of the tool. This method complements the metric-oriented evaluation of FrameCheck. It then draws out several resonances and tensions between these experiences and our speculative design strategy. Finally, while this work did not directly involve users in the design process, many of the experiences that emerged provide useful guidance for future work on participatory approaches to HCAD.

Reflext implementation. This system's design hinged on a fundamental tension: how could we decide which patterns of language to identify without making overly prescriptive commitments about the constitution of framing or those patterns' meanings? Ultimately, we chose a technique called selectional preference learning (Resnik, 1993; Ritter et al., 2010), which quantifies the tendency for co-occurrence between sets of linguistic predicates and arguments. For instance, the verb "to drink" tends to prefer, i.e. to select for, animals as its subject and potable liquids as its object. Reflext applied this technique to political news articles and political blog posts. What kinds of terms would likely appear with, i.e. be selected for by, such words as health care, Congress, privacy, war, etc.?

The visualization for Reflext allowed users to select a given term from a word cloud like interface (not depicted here) and see how different sources discussed that term. For instance, Figure 1 shows how the term "contraception" was discussed (i.e., the term's selectional preferences) in content about health care (in purple across the top) and in content about abortion (in blue around the bottom). The image shows how the selectional preference algorithm was mapped onto a visual representation, where strength of preference corresponds to size. This example demonstrates Reflext's support for interpretive flexibility (Pinch and Bijker, 1987; Sengers and Gaver, 2006). The visualization exposes linguistic patterns, e.g. in the context of health care, one is more likely to "provide," "pay," or "cover" contraception, while in the context of abortion, one is more likely to "ban," "withdraw," or "deny" contraception. However, in contrast to most natural language processing tools, the user must interpret what these patterns mean.

We deployed Reflext in a qualitative field between May and November 2012, which was the main campaign season for the US 2012 national elections (president, congress, etc.). For full details about both

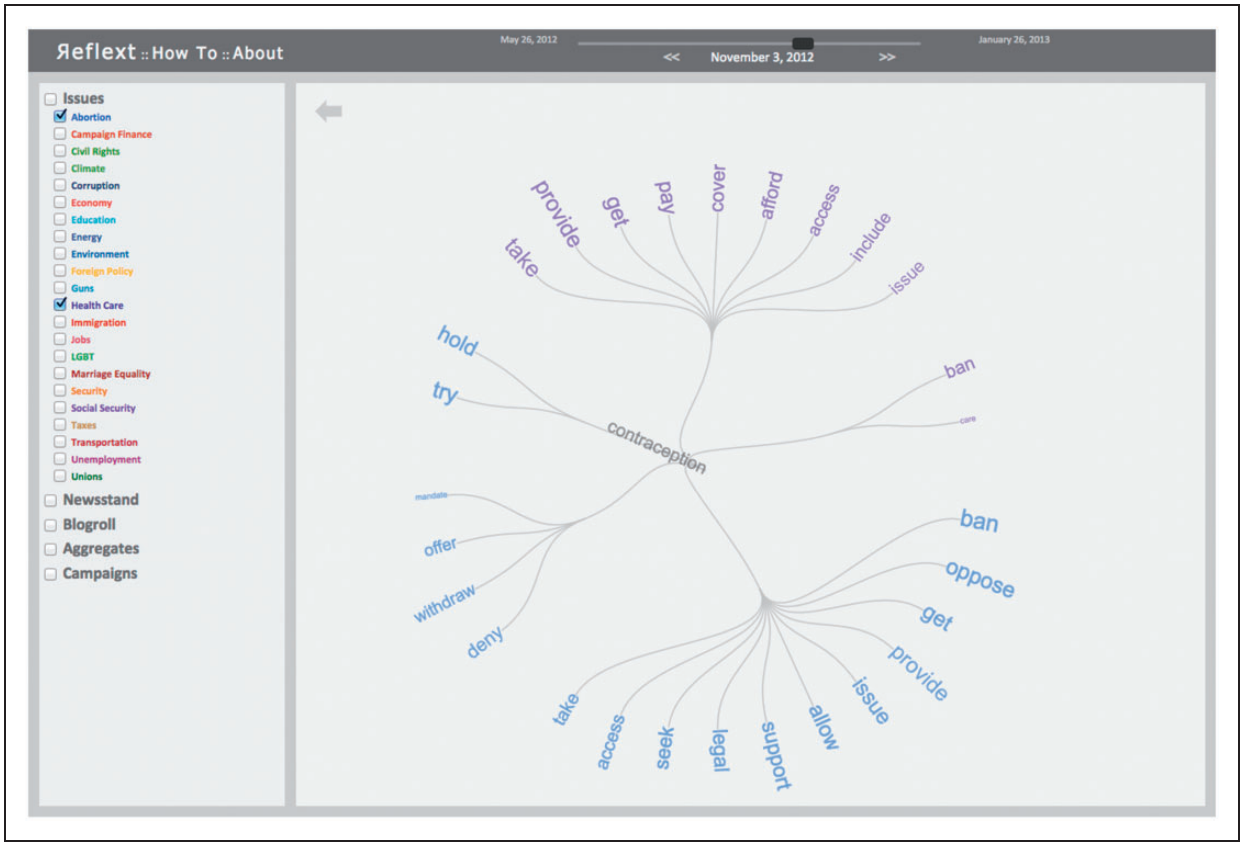

Figure 1. A screen shot from Reflext. This image shows selectional preferences for the term "contraception" in content about health care (in purple across top) and in content about abortion (in blue across bottom). The visualization shows groups of words that were statistically likely to appear in similar grammatical relationships, e.g. verbs appearing with "contraception" as their direct object.

Reflext's functionality and the field study, please see Baumer et al. (2014).

Our initial analysis of interview data from this field study focused on how Reflext supported (or did not support) practices consistent with frame reflection (Schön and Rein, 1994). However, during the interviews, we also noticed moments when participants described specific expectations about what the system *should* be showing them. These expectations and disconnects highlight the relationship between the actual functioning of algorithmically based systems and beliefs about both what computers *can* do and what they *should* do. Here, I refer to this construct as *computational imaginaries*, which I suggest plays a key role in speculative design.

Computational imaginaries. Prior work has described the idea of social imaginaries (Anderson, 1983; Castoriadis, 1975; Taylor, 2004). An individual person does not and cannot know how every other individual in a society behaves. Thus, people form social imaginaries, descriptions of how they believe others within a culture and/or our society behave. Here, I apply this same sensibility to beliefs about technology. Most people do not and cannot know everything about how a given technology functions. At some point, people use what they do know to extrapolate imaginaries, both about how technology does function and about how it might be able to function (Bucher, 2017). Such imaginaries—beliefs about what is easy, difficult, or even possible—may not align with current thinking, either technical or philosophical.

In one such disconnect, many participants described trying to use Reflext to identify bias. For instance, one respondent wanted to know "what percentage of the talk is from one side and what percentage of the talk is from the other." Another wanted "to see infographs at like a higher level. Like overall, just bias, not about specific terms or specific stories, but just [. . .] There's still more straight news that is biased." A third respondent described how he opened the analysis of Fox News, known for its fairly conservative viewpoint, and of the *New York Times*, which usually takes a relatively liberal stance. He then selected the term "Obama" to see how each source discussed the then-president running for reelection. "I figured that there would be some more unpleasanter [sic] words from the right leaning paper [Fox News] about Obama and more

nice words from the NY Times.'' These and other responses evince a normative desire for unbiased political coverage. Failing that ideal, algorithmically based systems should, according to these respondents, at least call attention to the presence of bias.

Such expectations expose some of the tensions at work in computational imaginaries. First, the detection of bias is an active and challenging area of computational linguistic research, exacerbated by the fact that humans often do not agree with one another as to what constitutes bias (Recasens et al., 2013). Thus, participants' desire for a system that identified ''overall just bias'' does not align with what is currently algorithmically possible. Second, according to most theories of framing, a neutral, unbiased description does not exist (Entman, 1993). ''Facts have no intrinsic meaning'' (Gamson, 1989: 157). That is, the political communication literature rules out the possibility of the kind of neutral, unbiased, ''straight news'' participants sought. Thus, a disconnect occurs not only between participants' expectations and Reflext's functionality, but also between participants' expectations and what is even possible, both theoretically and technically.

This concept applies quite broadly. Due to their details being obscured as trade secrets, computational imaginaries are necessary to interpret algorithmically based systems ranging from the Facebook news feed (Eslami et al., 2015; Rader and Gray, 2015), to Google Play's app recommender system (Ananny, 2011), to Twitter's trending topics (Gillespie, 2011).

In many ways, the computational imaginaries I describe here resemble Bucher's (2017) notion of the algorithmic imaginary. As Bucher (2017: 2) puts it: ''the algorithmic imaginary is [. . .] the way in which people imagine, perceive and experience algorithms and what these imaginations make possible.'' Thus, ''the algorithmic imaginary does not merely describe the mental models that people construct about algorithms but also the productive and affective power that these imaginings have'' (Bucher, 2017: 12).

While they share much in common, two important aspects distinguish computational imaginaries, as described here, from Bucher's (2017) algorithmic imaginary. First, computational imaginaries include a normative dimension of what computers *should* do. Participants' exhortations of how Reflext in particular should function stem largely from its political context, which implies specific democratic norms. In contrast, Bucher's (2017) work on interpretations of the Facebook newsfeed algorithm does not identify a similar normative component. Further work is required to ascertain whether this normative dimension of computational imaginaries arises more from the context and purpose of the system in question or from the notion of an algorithm per se. A second distinguishing aspect comes from the current paper's focus on design and, in particular, the use of a speculative approach.

The speculative strategy. Imagination plays a prominent role in critical and speculative design approaches (Dunne and Raby, 2013, 2001). This branch of design is concerned not necessarily with producing objects that are useful so much as provocative. For instance, Biojewellery (Thompson et al., 2006) crafts engagement rings out of bone that is grown from donor couples' wisdom teeth. The ''Object for Lonely Men'' series (Toran, 2001) provides other examples, such as a robot that throws plates as if having an argument, or a small fan in one's bed that mimics the sensation of sleeping next to a heavy breather. To reiterate, rather than produce functional products or prototypes, these design provoke dialog about the normative role of such technologies.

Since they involve extrapolating from existing circumstances to imagine possible futures, speculative approaches are less constrained by the realm of what is (currently) technically feasible. This freedom can facilitate thinking through the ramifications of different design variants without them necessarily existing (Baumer et al., 2014a; Blythe, 2014). Such advantages become more pronounced in the case of algorithm design, where the boundaries of what is possible can seem to change quite rapidly (although see Dreyfus, 1992). Instead of technical feasibility, what comes to the fore are the assumptions and values embedded in technology. Dunne and Raby (2001) describe critical design as a sort of value fiction. In the process of critical design, a designer identifies a value held by the dominant societal mainstream; subverts, negates, or otherwise alters that value; and then uses the altered value as the basis for a design.

Reflext applies this speculative approach to algorithm design. Traditional natural language processing and computational linguistics techniques often seek to provide an answer about, e.g. expressions of sentiment (Pang and Lee, 2008), topics discussed (Blei et al., 2003), or the grammatical decomposition of a sentence (De Marneffe et al., 2006)? This orientation toward achieving *the* answer emphasizes measurable performance metrics—accuracy, F1 score, etc.—as noted above. These metrics also align with values traditionally emphasized by HCI design, such as usability, speed, efficiency, responsiveness, a transparent interface, etc. (Card et al., 1983).

In contrast, Reflext can be described as a system that raises questions. Unlike sentiment analysis, topic modeling, or syntactic parsing, Reflext offers no prescriptive interpretation of the patterns it identifies. Instead, it asks users to determine what the differences might mean between, say, the discussion of contraception in

the context of health care versus in the context of abortion. This approach aligns with other speculative or ludic designs, such as the Drift Table (Gaver et al., 2004). This coffee table features a small, round display in the center showing aerial photos. It can only be "controlled" by placing heavy objects on the table to alter the direction that the photos slowly drift through the display. The Drift Table exchanges the values of speed, efficiency, and clear control for slowness, playfulness, and curiosity. Similarly, Reflext exchanges the values of performance and efficiency for provocation and exploration.

While this speculative approach successfully generated an artifact that deviated from the dominant norms, it also led to disconnects between participants' expectations and Reflext's functionality. Specifically, the interpretive flexibility (Pinch and Bijker, 1987; Sengers and Gaver, 2006) at the heart of our design became a point of tension and, at times, confusion. Participants in our study regularly requested a more prescriptive analysis and representation. As one participant put it: "Like if I pick out taxes, [I would want to see] how are taxes talked about in a left-leaning paper compared to a right-leaning paper [...] and I'll get it in like an objective kind of quantitative way." Another participant "would want to have frequency numbers, charts and tables that shows me how these two terms work together [...] Just anything to kind of spell out, yeah we can do this but *what does it mean* [emphasis added]?" These requests about prescribed meaning echo participants' perceptions and experiences around bias, where they desired a system that would explicitly state the degree and kind of bias present in an article.

Again, these computational imaginaries belie expectations both about what a computational system *could* do and about what it *should* do. With respect to technical feasibility, natural language processing has developed numerous techniques well adapted for analyzing text to identify important patterns. Ultimately, though, a human must interpret what those patterns mean (Rhody, 2012). Thus, participants' requests for an algorithm to "spell out [...] what does it mean" diverge from current state of the art. Furthermore, the above statements carry a normative implication that the system *should* perform these interpretive functions. These participants imply that, in order to be valuable, an algorithmically based system not only must identify patterns of interest, it must also ascribe human-interpretable meaning to those patterns. In other words, while our design made a commitment to interpretive flexibility, it did not necessarily convince our participants to adopt this same value.

This disconnect raises an important final point about a speculative strategy for algorithm design. Speculative approaches do not always generate specific designs to be widely adopted or even necessarily to be implemented at all. Although such designs often could be and sometimes are implemented, speculative approaches serve primarily as a conceptual device to help explicate and interrogate the values embedded in existing technologies. That is, speculative designs are just as much a discursive intervention as a technical one (Agre, 1997).

While Reflext offers one example, similar approaches could be applied to more well-known algorithms. What if the Facebook news feed showed items it believed you were *least* likely to click on or to like? What if Amazon's recommendations were designed not to encourage additional purchases but to make a user feel a particularly way about their socioeconomic status (e.g., affluent, low-brow, middle-class) compared to other users? What if Twitter's trending topics selected those terms or hashtags most likely to be confusing or misinterpreted out of context? These variants may not meet the economic goals of Facebook, Amazon, or Twitter. Making users feel variously low-brow or affluent would not likely increase their spending. Highlighting confusing hashtags might not increase Twitter's monetizable user traffic.

However, imagining such alternatives offers at least two important outcomes. First, it shows how the current state of computational systems both is contingent and is nonexclusive with other possibilities. Second, it highlights how abstract concepts and values, such as likability, socioeconomic status, or interpretability, are in some ways already algorithmically encoded into these systems.

While not employed here, a third strategy may be useful in addressing disconnects between algorithm designers' intent and lay persons' expectations.

The participatory strategy. Although quite different from one another, both the theoretical strategy and the speculative strategy share at least one commonality. Neither incorporates people who might be impacted by a system as participants in the design process.

Participatory design emerged largely from Scandinavian countries as a form of empowering laborers, giving them some measure of control in the design of their workplace settings. Participatory techniques gained significant traction in information technology design, especially for workplace technologies (Muller and Druin, 2012; Muller and Kuhn, 1993). Not all instances, though, carry the original movement's political orientation and weight (Beck, 2002).

One of the unique challenges in participatory design deals with differential expertise. In his work developing software for typographers, Ehn (1988) describes how he and his design team needed to gain some amount of proficiency in typographic practices. Although Ehn

does not discuss it at length, the same imbalance occurs in the other direction; typographers needed to understand something of software engineering and interface design. However, Ehn also acknowledges limits of this technique. One cannot simply grant a participant-designer the equivalent of a graduate-level training in computer science. Similarly, one cannot easily bestow upon a computer scientist an expert understanding of the subtle nuances in journalistic news reporting.

Consider, for instance, the way that participants reasoned about size in the visualization. Following a mass shooting in Aurora, CO, one participant recounted being surprised not to see the term ''Aurora'' more prominently in the visualization. When she ''finally found a really tiny Aurora somewhere on the list, [she] thought 'well it should've probably been bigger' because it was such a hot story.'' Another participant similarly asserted that ''the word that's biggest [. . .] is the word that's the most important word of the discussion.''

We see here a conflation between frequency and importance, a conflation on which most natural language processing algorithms hinge. However, numerous words that occur very frequently—''the,'' ''is,'' ''to,'' ''this,'' ''and''—likely have little importance. Computational approaches often employ stopword lists (Leskovec et al., 2014), which screen out such high frequency terms, as a means of focusing on what are supposedly more content-oriented terms.

Essentially, this point highlights differences between how computers count and how people count. If a shooting occurs in Aurora, the location will likely be named once in a given news article, perhaps twice. Other terms, even nonstopwords, will likely occur far more frequently, regardless of whether the article is *about* such words.

Applying a participatory approach from the outset may not necessarily resolve such tensions. However, it might identify them sooner and potentially incorporate them into the design. For instance, the data-driven, probabilistic selectional preference technique could have been replaced or supplemented with sentiment analysis, named entity recognition, or other methods more in line with participants' intuitions about the important aspects of a given article. Moreover, a participatory approach could facilitate a dialectic exchange among those who might interact with the system and those who are designing and implementing it. Doing so would help better understand the formation of, and potentially how to intervene in the process of developing, computational imaginaries.

Implications and practicalities

This paper presents the author's experiences with applying human-centered strategies to the design of algorithmically based systems. It discusses both strengths and challenges in this approach. However, the ''toward'' in the paper's title should be taken seriously. The steps described here provide important groundwork, but they initially move along a longer term trajectory. To conclude, I discuss some important considerations of what will likely need to be done if we are to foster a more developed HCAD practice.

One major concern revolves around what is actually meant by the term ''theory.'' As described above, even when social scientific theories are used to inform algorithm design, the specific features on which an algorithmically based system focuses may vary in how closely they resemble those theoretical concepts. Moreover, what actually constitutes a theory can vary across disciplines, sometimes drastically so. For instance, the theories of framing applied above bear little similarity to the concerns of theoretical computer science (complexity, formal logic, automata, etc.). Thus, understanding, not to mention grappling with, theoretical concerns across multiple disciplines requires expertise across multiple disciplines.

This point draws attention to a second major concern. Rhetoric around interdisciplinary research often lauds training that results in so-called T-shaped people, who have a breadth of knowledge across a wide variety of subjects as well as deep expertise in one area. In practice, though, strict focus on a single depth of expertise can become limiting. For instance, the computer scientist who has a surface understanding about social scientific theories of framing will have limited guidance in terms of feature selection. Instead, it may prove more effective to cultivate what one might call ''Π-shaped'' people, ones who have a wide breadth of knowledge but also deep expertise in multiple disparate disciplines. Clearly, such an approach becomes difficult given traditionally disciplinary publication venues, grant funding programs, departments, research institutes, etc. Thorough discussion of such concerns, however, far exceeds the scope of this paper.

Finally, gaps in understanding exist not only for designers across disciplines but also for those impacted by algorithms. While an algorithm's designers do not necessarily need or want others to understand *how it works*, they likely want to reach consensus on *what it means*. This paper provides strategies to address misalignments or disconnects between algorithmically based systems' technical functioning and their interpreted meaning. Theoretical approaches can help ensure from the outset that algorithm design is grounded in a thorough understanding of behavioral and social phenomena. Speculative approaches can help highlight the values and assumptions on which algorithmically based systems are predicated. Participatory strategies can help identify and

incorporate lay interpretations earlier in the design process. The crucial question then becomes: to what extent and in what ways are these various strategies effective or ineffective?

This question returns to the paper's central motivation about evaluation. The work presented above variously used metric-driven evaluation and human subject studies at different stages. While traditional algorithmic performance metrics and human-centered design approaches are both necessary, neither is sufficient alone. In as much as evaluation is central to design, HCAD must be paired with human-centered evaluation methods. Doing so does not preclude, say, running two variants of an algorithm to compare performance metrics. It does, however, require combining such metrics with more time- and labor-intensive human subject studies. Whether via ''Π-shaped'' people or multidisciplinary teams, a confluence of human-centered expertise and technical expertise is necessary to assess exactly when (i.e., during which phases of the design process) and how human subject studies should be deployed.

It is tempting to ask whether such design approaches could have avoided algorithms that, e.g. demonstrate bias along the lines of race, gender, or sexual orientation (Ananny, 2011; Barocas and Selbst, 2016; Hern, 2015; Sweeney, 2013; Zimmer, 2007). Instead, I suggest we ask how we would know. What kinds of evaluation techniques must we develop to identify such situations? While some algorithmic tools purport to identify such biases (Bolukbasi et al., 2016), questions of bias are ultimately questions of interpretation, of meaning making. It is these questions of meaning making that most clearly elucidate the limitations of traditional algorithm design and evaluation methods. Similarly, algorithmic support for the processes of meaning making is where human-centered approaches stand to contribute the most.

Acknowledgments

Thanks to Geri Gay and Francesca Polletta, the investigators on that grant; to our study participants; and to all the student collaborators who made invaluable contributions. Thanks also to Samir Passi and to Jed Brubaker for useful conversations.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This material is based upon work supported by the NSF under Grant No. IIS-1110932 and was completed largely while the author was affiliated with Cornell University.

Note

1. The first person plural in this section refers to the larger team that worked on this project. The first person singular refers specifically to the current paper's author.

References

Agre PE (1997) Toward a critical technical practice: Lessons learned in trying to reform AI. In: Bowker GC, Gasser L, Star SL, et al. (eds) *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ananny M (2011) The curious connection between apps for gay men and sex offenders. *The Atlantic*, April.

Anderson B (1983) *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.

Barocas S and Selbst AD (2016) Big data's disparate impact. *California Law Review* 104: 671–732.

Baumer EPS (2015) Usees. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Seoul, South Korea, pp. 3295–3298. ACM Press.

Baumer EPS, Bie M, Bonsignore EM, et al. (2014a) CHI 2039: Speculative research visions. In: *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (alt. CHI)*, Toronto, ON, pp. 761–770.

Baumer EPS and Brubaker JR (2017) Post-userism. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Denver, CO, pp. 6291–6303.

Baumer EPS, Cipriani C, Davis M, et al. (2014b) Broadening exposure, questioning opinions, and reading patterns with Reflext: A computational support for frame reflection. *Journal of Information Technology and Politics* 11(1): 45–63.

Baumer EPS, Elovic E, Qin Y, et al. (2015a) Testing and comparing computational approaches for identifying the language of framing in political news. In: *Proceedings of the annual meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Denver, CO, pp. 1472–1482.

Baumer EPS, Guha S, Quan E, et al. (2015b) Missing photos, suffering withdrawal, or finding freedom? How experiences of social media non-use influence the likelihood of reversion. *Social Media + Society* 1(2): 1–14.

Beck EE (2002) P for political: Participation is not enough. *Scandinavian Journal of Information Systems* 14(1): 77–92.

Bleecker J (2009) *Design Fiction: A Short Essay on Design, Science, Fact and Fiction*. Venice Beach, CA: Near Future Laboratory.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Blythe M (2014) Research through design fiction: Narrative in real and imaginary abstracts. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Toronto, ON, pp. 703–712.

Bolukbasi T, Chang K-W, Zou JY, et al. (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Lee DD, Sugiyama M,

Luxburg UV, et al. (eds) *Advances in Neural Information Processing Systems (NIPS)*, pp. 4349–4357. Curran Associates, Inc. Available at: http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf.

Bowker GC and Star SL (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Brewer PR (2002) Framing, value words, and citizens' explanations of their issue opinions. *Political Communication* 19(3): 303–316.

Bucher T (2017) The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication and Society* 20(1): 30–44.

Burrell J (2016) How the machine ''thinks'': Understanding opacity in machine learning algorithms. *Big Data and Society* 3(1): 1–12.

Card D, Boydstun AE, Gross JH, et al. (2015) The media frames corpus: Annotations of frames across issues. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Beijing, China, pp. 438–444.

Card SK, Moran TP and Newell A (1983) *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Castoriadis C (1975) *The Imaginary Institution of Society*. (K Blamey, Trans.). Cambridge: MIT Press.

Chong D and Druckman JN (2007) Framing theory. *Annual Review of Political Science* 10(1): 103–126.

De Marneffe MC, MacCartney B and Manning CD (2006) Generating typed dependency parses from phrase structure parses. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, pp. 449–454.

Dreyfus HL (1992) *What Computers Still Can't Do*. Cambridge: MIT Press.

Druckman JN, Fein J and Leeper TJ (2012) A source of bias in public opinion stability. *American Political Science Review* 106(2): 430–454.

Druckman JN and Nelson KR (2003) Framing and deliberation: How citizens' conversations limit elite influence. *American Journal of Political Science* 47(4): 729–745.

Duhigg C (2012) How companies learn your secrets. *The New York Times Magazine*, February.

Dunne A and Raby F (2013) *Speculative Everything*. Cambridge: MIT Press.

Dunne T and Raby F (2001) *Design Noir: The Secret Life of Electronic Objects*. Berlin: Birkhäuser.

Ehn P (1988) *Work-Oriented Design of Computer Artifacts*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Entman RM (1993) Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4): 51–58.

Eslami M, Rickman A, Vaccaro K, et al. (2015) ''I always assumed that I wasn't really that close to [her]'': Reasoning about invisible algorithms in news feeds. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Seoul, South Korea, pp.153–162.

Fairclough N (1999) Global capitalism and critical awareness of language. *Language Awareness* 8(2): 71–83.

Gamson WA (1989) News as framing. *American Behavioral Scientist* 33(2): 157–161.

Gamson WA and Modigliani A (1989) Media discourse and public opinion on nuclear power: A Constructionist approach. *The American Journal of Sociology* 95(1): 1–37.

Gaver WW, Bowers J, Boucher A, et al. (2004) The drift table: Designing for ludic engagement. In: *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI)*, Vienna, Austria, pp.885–900.

Gillespie T (2011) Can an algorithm be wrong? *Limn* 1(2).

Guyon I and Elisseeff A (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research* 3(3): 1157–1182.

Hart PS (2011) One or many? The influence of episodic and thematic climate change frames on policy preferences and individual behavior change. *Science Communication* 32(2): 1–24.

Hern A (2015) Flickr faces complaints over ''offensive'' auto-tagging for photos. *The Guardian*, 20 May.

Hopkins D and King G (2010) A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54(1): 229–247.

Knuth DE (1974) Computer programming as an art. *Communications of the ACM* 17(12): 667–673.

Kramer ADI, Guillory JE and Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* 111(29): 8788–8790.

Lakoff G and Turner M (1989) *More Than Cool Reason: A Field Guide to Poetic Metaphor*. Chicago, IL and London: University of Chicago Press.

Lenssen P (2007a) Did you mean: ''He invented''? Available at: http://blogoscoped.com/archive/2007-05-07-n56.html (accessed 14 November 2016).

Lenssen P (2007b) Google stops ''did you mean: He invented''. Available at: http://blogoscoped.com/archive/2007-05-24-n36.html (accessed 14 November 2016).

Leskovec J, Rajaraman A and Ullman JD (2014) Data mining. In: *Mining of Massive Datasets*. Cambridge University Press, pp.1–19.

Linehan C, Kirman B, Blythe M, et al. (2014) Alternate endings: Using fiction to explore design futures. In: *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI EA)*, Toronto, ON, pp. 45–48.

Maibach E, Nisbet MC, Baldwin P, et al. (2010) Reframing climate change as a public health issue: an exploratory study of public reactions. *BMC Public Health* 10(1): 299.

Muller MJ and Druin A (2012) Participatory design: The third space in HCI. In: Jacko J (ed.) *The Human-Computer Interaction Handbook*. Vol. 4235, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 1–70.

Muller MJ and Kuhn S (1993) Participatory design. *Communication of the ACM* 36(6): 24–28.

Murnane EL and Counts S (2014) Unraveling abstinence and relapse: Smoking cessation reflected in social media. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Toronto, pp.1345–1354.

Pang B and Lee L (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1–2): 1–135.

Pinch TJ and Bijker WE (1987) The social construction of facts and artifacts. In: Bijker WE, Hughes TP and Pinch TJ (eds) *The Social Construction of Technological Systems*. Cambridge: MIT Press, pp. 17–50.

Price V, Nir L and Cappella JN (2005) Framing public discussion of gay civil unions. *Public Opinion Quarterly* 69(2): 179–212.

Quattrone GA and Tversky A (1988) Contrasting rational and psychological analyses of political choice. *The American Political Science Review* 82(3): 719.

Rader E and Gray R (2015) Understanding user beliefs about algorithmic curation in the Facebook news feed. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Seoul, South Korea, pp.173–182.

Recasens M, Danescu-Niculescu-Mizil C and Jurafsky D (2013) Linguistic models for analyzing and detecting biased language. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria, pp.1650–1659.

Redström J (2006) Towards user design? On the shift from object to user as the subject of design. *Design Studies* 27(2): 123–139.

Rein M and Schön DA (1996) Frame-critical policy analysis and frame-reflective policy practice. *Knowledge and Policy* 9(1): 85–104.

Resnik P (1993) *Selection and Information: A Class-Based Approach to Lexical Relationships*. Philadelphia, PA: University of Pennsylvania.

Rhody L (2012) Topic modeling and figurative language. *Journal of Digital Humanities* 2(1).

Ritter A, Mausam and Etzioni O (2010) A latent dirichlet allocation method for selectional preferences. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 424–434.

Sandvig C (2014) Seeing the sort: The aesthetic and industrial defense of ''the algorithm''. *Media-N* 11(1): 35–51.

Schön DA and Rein M (1994) *Frame Reflection: Toward the Resolution of Intractable Policy Controversies*. New York: Basic Books.

Schuldt JP, Konrath SH and Schwarz N (2011) ''Global warming'' or ''climate change''?: Whether the planet is warming depends on question wording. *Public Opinion Quarterly* 75(1): 115–124.

Sengers P and Gaver B (2006) Staying open to interpretation: Engaging multiple meanings in design and evaluation. In: *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)*. pp.99–108. University Park, PA: ACM.

Sniderman PM and Theriault SM (2004) The structure of political argument and the logic of issue framing. In: Saris WE and Sniderman PM (eds) *Studies in Public Opinion*. Princeton, NJ: Princeton University Press, pp. 133–165.

Sterling B (2005) *Shaping Things*. Cambridge: MIT Press.

Sweeney L (2013) Discrimination in online ad delivery. *Communications of the ACM* 56(5): 44–54.

Tanenbaum J (2014) Design fictional interactions. *Interactions* 21(5): 22–23.

Taylor C (2004) *Modern Social Imaginaries*. Dunham, NC: Duke University Press.

Thompson I, Scott N and Kerridge T (2006) *Biojewellery: Designing Rings with Bioengineered Bone and Tissue*. London: Oral & Maxillofacial Surgery, King's College London.

Toran N (2001) Object for lonely men. Available at: http://noamtoran.com/NT2009/projects/object-for-lonely-men (accessed 14 November 2016).

Turney PD, Neuman Y, Assaf D, et al. (2011) Literal and metaphorical sense identification through concrete and abstract context. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, Scotland, Vol. 2, pp.680–690.

Tversky A and Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481): 453–458.

Zimmer M (2007) Google: ''Did you mean: 'He invented'?'' Available at: http://www.michaelzimmer.org/2007/05/09/google-did-you-mean-he-invented/ (accessed 12 July 2016).

This article is a part of special theme on Algorithms in Culture. To see a full list of all articles in this special theme, please click here: http://journals.sagepub.com/page/bds/collections/algorithms-in-culture.