Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms

Christian Sandvig*[1], Kevin Hamilton[2], Karrie Karahalios[2], & Cedric Langbort[2]

Paper presented to “*Data and Discrimination: Converting Critical Concerns into Productive Inquiry*,” a preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA.

* - Corresponding Author

[1]Department of Communication Studies, University of Michigan
4322 North Quad, 105 S State St
Ann Arbor, MI 48109 USA
csandvig@umich.edu

[2]Center for People and Infrastructures, Coordinated Science Laboratory
University of Illinois, Urbana-Champaign
1308 West Main Street, Urbana IL 61801 USA
{kham, kkarahal, langbort} @illinois.edu

Introducing “Screen Science”

In a pioneering commercial application of computing, in 1951 American Airlines partnered with IBM to attack the difficult logistical problems of airline reservations and scheduling (Copeland et al. 1995). The resulting computer system was named SABRE (the Semi-Automatic Business Research Environment), after the Air Force’s SAGE air defense system. SABRE was launched in 1960 and by 1964 it was hailed as the largest commercial computer network in existence (Redmond & Smith, 2000: 438). Airline reservations had previously been handled by a network of telephone call centers where actual seating charts of specific flights were reserved using push-pins. Double-booking was prevented by an elaborate paper process of multiple confirmations. With the SABRE system, by the mid-1970s travel agents across the country could complete a near-instantaneous booking for most airlines via special dedicated SABRE terminals. SABRE was a dramatic success for the computer and airline industries, reducing the time required to book a plane ticket from up to three hours from request to confirmation with paper and telephone to just a few minutes via computer. A later iteration of the SABRE system is still in use today, providing the technology behind Web sites like Expedia and Travelocity.

Yet as SABRE grew, American Airlines (owner of SABRE) also developed a new competitive strategy that its employees called “screen science” (Petzinger, 1996). To make its initial offering more useful, SABRE offered flight reservations for many airlines, not just American. Employees at American learned that users of the system tended to choose the first flight displayed in the results list even if it was not the optimal result for their query. In addition, the rules that could govern the display of flights were actually quite complex. The “screen science” group at American found that SABRE could favor American Airlines “by choosing criteria for the display algorithm that match distinctive characteristics of its own flights, such as connecting points and nonstop service” (Harvard Law Review, 1990: 1935).

American’s internal slang term “screen science” may have been braggadocio, because American’s manipulations of flight search results became increasingly unsubtle: it does not appear that there was much unsubtle about it. Travel agents and competitors noticed that the first flight returned was often an American flight that was much longer and more expensive than other alternatives, eventually leading the US Civil Aeronautics Board and the Department of Justice to launch antitrust investigations.

Surprisingly, in the face of public scrutiny the company did not deny its manipulations. Speaking before the US Congress, the president of American, Robert L. Crandall, boldly declared that biasing SABRE’s search results to the advantage of his own company was in fact his primary aim. He testified that “the preferential display of our flights, and the corresponding increase in our market share, is the competitive raison d’etre for having created the [SABRE] system in the first place” (Petzinger, 1996). We might call this perspective “Crandall’s complaint:” Why would you build and operate an expensive algorithm if you can’t bias it in your favor?

In response, in 1984 the Board decreed that the sorting algorithms for SABRE must be known, and passed a little-known regulation (codified at 15 CFR 255.4) entitled “Display of

Information," requiring (among other provisions) that: "Each [airline reservation] system shall provide to any person upon request the current criteria used in editing and ordering flights for the integrated displays and the weight given to each criterion and the specifications used by the system's programmers in constructing the algorithm." ([b][3]; see also Locke, 1989).[1]

Today, "algorithm transparency" is again a pressing societal problem, but it reaches far beyond airlines (Pasquale, 2011). Infrastructures of all kinds are being transformed into "smart" iterations which feature embedded computing power, telecommunications links, and dynamic real-time control (Graham & Marvin, 2001). At the core of these systems sit algorithms that provide functions like social sorting, market segmentation, personalization, recommendations, and the management of traffic flows from bits to cars. Making these infrastructures newly computational has made them much more powerful, but also much more opaque to public scrutiny and understanding. The history of sorting and discrimination across a variety of contexts would lead one to believe that public scrutiny of this transformation is critical. This paper addresses the question: How can such public interest scrutiny of algorithms be achieved?

Algorithms that Appear to Work Well May Still be Dangerous

We take the perspective that virtually any algorithm may deserve scrutiny. In the popular mind, algorithms like the Google search engine algorithm exist in order to satisfy their users, and so Crandall's pessimistic perspective that all algorithms are probably rigged might seem counter-intuitive. However, while it is true that a search algorithm that did not satisfy its users would be unlikely to continue operation for very long, it is important to note that most situations in which algorithms are employed permit the algorithms to satisfy multiple goals simultaneously. There are also many ways an algorithm might be "rigged" that are normatively problematic. We argue that public interest scrutiny of algorithms is required that will focus on subtle patterns of problematic behavior and that this may not be discernable directly or via a particular instance.

First, consider that algorithms can be manipulated in ways that do not disadvantage their users directly or obviously. For example, a user visiting google.com to locate expert health advice on a worrying medical symptom might be equally satisfied by advice from WebMD, the Mayo Clinic, the Centers for Disease Control, and Google Health. However, Edelman recently discovered what appear to be a series of hard-coded rules placing Google-provided services at the top of Google search results for some queries, despite Google's public statements that the company would never use such hard-coded rules as part of its algorithm (2010). In Edelman's searches, Google Health (a subsidiary of Google) was always returned first for health-related keywords. After Google's "screen science" received some publicity, Google modified its algorithm and ceased to return its own properties first. However, providing its own subsidiaries with free advertising or integrating them into Google search raises serious antitrust concerns (Edelman, 2014). In this scenario Google designed its algorithm in a way that could certainly be illegal, but the users of its search interface are not harmed directly and would be unlikely to perceive its search engine to be any less useful.[2]

[1] Earlier regulations went further than this and specified the variables that could be used in sorting the results.

[2] Indeed, although there is harm it is Google's competitors who are harmed and this economic damage of foreclosed innovation and competition would be very difficult for a user to perceive.

Second, algorithmic manipulation may be societally problematic and deserving of scrutiny even if it is not illegal. This is a significant observation because the majority of scholarship that has considered algorithmic discrimination in the past has done so from the perspective of law and regulation. Legal approaches to algorithmic monitoring have included the promulgation of regulations about algorithms (as the Aeronautics Board regulated SABRE) and the use of powerful tools of public scrutiny such as the subpoena. Yet the public may need to know something about how algorithms operate even if there is no recourse to legal tools and it is not likely that a crime has been committed.

An example that cements this point is the 2012 rise of the YouTube phenomenon known as the "Reply Girls."[3] YouTube's video hosting service allows video uploaders to indicate thematic relationships between videos via a function known as the "video response." If an uploader of Video B indicates via metadata that Video B is a "response" to Video A, it is presumably more likely that YouTube's recommendation algorithm will recommend Video B to people who have just watched Video A. This is intended to promote video conversations on the site and to increase the quality of algorithmic recommendations. However, many YouTube "partners" make money via advertising revenue-sharing arrangements offered by YouTube and thus they have a financial incentive to increase the audience of their videos no matter the content.

In 2012, a group of female uploaders discovered an effective strategy to increase audience and advertising revenue. These "reply girls" produced videos whose thumbnail depicted their own cleavage (Video B), then marked these videos as responses to content that was topical and was receiving a large number of views (Video A). As in our last example, the viewer who watches a video and is then next shown a picture of cleavage (and clicks on it) presumably may be satisfied by this transaction.[4] The audience for Video A might simply want to watch something next that holds their attention, and there may be many videos on YouTube that would satisfy this requirement.

However, since topical content is much more likely to be related to news and politics, the overall consequence of a prevalent "reply girls" strategy is that YouTube links news and politics to breasts. We might say that a recommender algorithm fully captured by "reply girls" responds with a recommendation for breasts whenever a viewer expresses interest in a topical political story. There is nothing illegal about this algorithm, and it is not the result of ill-intent by the algorithm's designers, nonetheless it is normatively problematic and deserves scrutiny by public interest researchers.[5] YouTube, in this situation, may inadvertently become a machine that produces misogyny and depresses political participation.

Note in this example that understanding the societally problematic behavior of YouTube requires systematic investigation of the recommender system, as simply finding that YouTube recommends a particular "reply girl" to you in one instance might occur for any number of reasons. In this instance the recommendation is a problem if it is a widespread pattern.

[3] http://knowyourmeme.com/memes/reply-girls

[4] We note, however, that there was a considerable backlash against "reply girls" and some viewers were not satisfied.

[5] It is not clear what the implications of "reply girls" are as this has not been studied, therefore this scenario is an extrapolation of possible consequences.

Ultimately, then, when we specify that public interest scrutiny of algorithms is important we do not mean scrutiny that can be satisfied by legal tools alone, or scrutiny that can be achieved by a single instance of trial-and-error use of an algorithm. Although individual trial-and-error may detect some forms of harm (Edelman, 2010), there are a variety of other problems that can only be uncovered via a systematic statistical investigation. Because algorithms often operate in a personalized way, individual investigations are unlikely to produce a broad picture of the system's operation across the diversity of its users.

As algorithms (and computers) become more common in the implementation of all technological systems, studying the world means studying algorithms. Yet the result of computerizing a process in an algorithm is often to reduce our ability to study it (Gillespie 2014). Worrisome algorithms may span a great range of inquiry from unlawful monopoly leveraging to public-interest-minded research about the broader media and information environment. To highlight this range of potential empirical research needs, a sampling of research questions with important public interest implications that demand systematic algorithmic scrutiny might thus include:

> **Fair Housing**: Are racial minorities less likely to find housing via algorithmic matching systems? (e.g., Edelman & Luca, 2014)
>
> **Economic Opportunity**: Does algorithmically-controlled personalization systematically restrict the information available to the economically disadvantaged? (e.g., Turow, 2013)
>
> **Payola**: A content distributor may make more money displaying some content vs. others. Are content recommendations influenced by this revenue?
>
> **Price Discrimination**: Do online markets unfairly make goods more expensive for particular demographics or particular geographic locations? (e.g., Valentino-Devries et al. 2012)

The Audit Study: Field Experiments That Detect Discrimination

To answer these questions, it is reasonable to turn to the most prevalent social scientific method for the detection of discrimination: this is known as the audit study (Mincy, 1993). Audit studies are typically field experiments in which researchers or their confederates participate in a social process that they suspect to be corrupt in order to diagnose harmful discrimination. In one common audit study design, researchers create a fictitious correspondence purporting to be from a job applicant seeking employment, and target it at real employers. In these studies, two or more equivalent resumes are prepared reflecting equal levels of education and experience. "The race…of the fictitious applicant is then signaled through one or more cues" such as the fictitious applicant's name, which might be manipulated between the two conditions of "Emily" and "Lakisha" to signal "Caucasian" vs. "African-American" (Pager, 2007, 109-110). The manipulation of the applicant's name is then used to measure of discrimination, allowing researchers to determine if identically qualified applicants receive differential treatment based on race. "[M]ost researchers view the audit methodology as one of the most effective means of measuring discrimination" (p. 114), and audit studies are frequently used as evidence in public policy debates and in court cases alleging discrimination (see also Riach & Rich, 2002).

Although the word "audit" may evoke financial accounting, the original audit studies were developed by government economists to detect racial discrimination in housing by the research unit of the US Department of Housing and Urban Development in the 1970s. Although the word "audit" has a similar dictionary meaning in both cases, the "audit study" as it evolved in social science is distinct from financial auditing. Audit studies are field experiments (Cook & Campbell, 1979) meaning a research design involving the random assignment to groups in a controlled setting in order to isolate causation, as in the above example employers may randomly receive a resume with either an African-American or Caucasian name. In a "field" experiment, as opposed to a traditional experiment, realistic settings and situations are used in order to ensure that the results are generalizable to real-world experience. However, as a consequence some forms of experimental control may be relaxed, making causal inference more difficult than it would be in a laboratory or classical experiment (Pager, 2009).

Audit studies of discrimination are often divided into two groups: correspondence tests and in-person audits (40). Correspondence tests include the example study of resumes given above, and are so called because they relied on fictional correspondence (initially, by postal mail). This remains an important and influential social scientific research method. For instance, a recent correspondence test that received wide press coverage found that professors were less likely to answer e-mails from an unknown student requesting an appointment if the sender was identifiable by name as a woman or a member of a racial minority group (Milkman et al. 2012).

In-person audits, in comparison, employ hired actors or research assistants called "testers" who might simulate applying for a job, requesting a mortgage, or buying a car in person (Pager, 2009: 42-43). The use of hired testers allows the investigation of discrimination that occurs in face-to-face situations, but because any two human testers are likely to vary in more ways than on just one variable of interest (e.g., race or gender), it is more difficult to cleanly assign causal inference in in-person audit studies. In-person audits are also subject to experimenter bias, as it is not possible to construct them in an effectively blind way it is sometimes alleged that researchers expecting a particular kind of discrimination might subtly produce it.

Instead of an audit of employers or real estate agents, we propose that the recently raised normative concerns that have been raised involving algorithmic discrimination (see Barocas, Hood, and Ziewitz, 2013; Gillespie 2014) demand an audit of online platforms. In essence, this means that a program of research should be undertaken to audit important Internet-based intermediaries with large data repositories (e.g., YouTube, Google, Facebook, Netflix, and so on) to ascertain whether they are conducting harmful discrimination by class, race, gender, and to investigate the operation of their algorithms consequences on other normative concerns. We find this to be a promising approach, yet such a research design also raises a number of important practical and ethical problems.

The Ethical Challenges Facing Audit Designs

Audit studies are quite distinct from other social scientific research designs. Although other social scientific designs employ field experiments, actors, deception, and even correspondence, audit studies of various kinds are often grouped together because the context of discrimination research produces particular, unique design goals and challenges. For example, audit studies are often used as evidence in lawsuits alleging employment or housing discrimination.

Because the goal of the audit study is often to prove that prohibited racial or gender discrimination exists for a specific practical purpose, many audit designs have tended toward simplicity in order to make the clearest possible case for discrimination to as wide an audience as possible. Rather than statistical pyrotechnics, audit studies still employ extremely simple coin-tosses to assign resumes or testers to experimental groups, and they may report results as simple percentages. Audit studies, because of their legal context, are designed to be so simple that even a lawyer or judge can understand them.

Audit studies are also remarkable in that they violate widely held precepts of ethical practice in science. To be blunt, audit studies waste the time of people who participate in them, even the innocent. Audit studies generate work by (e.g.,) producing fake job applicants and appointment requests. It is even more striking that the audit study typically intends to show that participants in it are immoral, or even that they are criminals (although audit studies done by academics do not aim to identify particular criminals, but rather to make a statement about the overall prevalence of illegal discrimination).

In order for an audit design to work, participants must not know they are being audited, violating the basic principles for the ethical conduct of science. Audit studies are not conducted with informed consent, and a review of published audit studies finds that many of them are silent about any form of debriefing. Yet audit studies continue to be held in high regard, likely because the bedrock principles for ethical conduct in behavioral science, the Belmont Report, emphasizes the idea that any research involves a calculation of risks and benefits and that normal ethical practices (such as consent) may need to be set aside for the greater good in specific research studies. That is, the evidence produced by audit studies is so important to society that the scholarly consensus has been that researchers must be allowed to waste a certain amount of other people's time and to operate without informed consent in order to secure meaningful evidence about these important social problems and crimes.

This special status of the audit study clearly only holds if it is used to investigate questions of great importance. In comparison, a Columbia business school professor used a modified correspondence audit design in 2001 to investigate customer service at New York City restaurants.[6] He mailed some restaurant owners a fabricated letter claiming he had received food poisoning at the restaurant. Some suspicious restaurateurs felt that the letters might be false and complained to Columbia. These complaints led to the Dean declaring the study to be "an egregious error in judgment" that was "part of an ill-conceived research project."[7] The owners of two dozen restaurants sued the professor and Columbia University for libel and emotional distress.[8] Belmont report principles clearly hold that ethical expectations about

[6] As this study was never completed, it is not clear what the design was. It may not have been technically an audit as no random assignment to groups was reported in the media. However, the brazen use of false correspondence remains a useful parallel to the audit designs we have discussed so far.

[7] http://www.nytimes.com/2001/09/08/nyregion/scholar-sets-off-gastronomic-false-alarm.html

[8] http://www.nytimes.com/2004/02/08/nyregion/following-up.html

informed consent and deception can only be waived if the knowledge to be gained is significant enough to warrant one of these unusual social science research designs.

A Proposal: Algorithm Audits

We propose that an analogous research design to the audit study can be used to investigate normatively significant instances of discrimination involving computer algorithms operated by Internet platforms. However, audits of Internet platforms must necessarily differ from the traditional design of the audit as a field experiment. The differences are many, but one central difference is that the field experiment audit was typically designed to target a societal phenomenon, providing an estimate of its prevalence across a number of individuals who were not, in themselves, very important. For instance, early housing audits in the 1970s estimated housing discrimination by auditing landlords, but the goal was not to make a conclusion about a particular landlord.

In comparison, algorithms of interest are located within particular Internet platforms and it is much more likely that an algorithm audit would seek to target a particular platform. While in theory we could ask a question like "what is the prevalence of (phenomenon) among all search result providers?" in practice each platform likely employs a unique application programming interface that makes interacting with it programmatically much less standardized than, say, applying for a job or an apartment. As another exampls, rating videos on Netflix or YouTube in order to assess recommendation algorithms requires a user account. The processes for obtaining a user account (as well as the information required) differ for each platform. If we concede that an effective audit needs more than just a few tests, it is likely that a software apparatus for audit will need to be built to investigate a particular platform.

It is also the case that the online markets that employ algorithms exhibit much more concentration than those in housing and employment. Google Search market share by country can be as high as 98%, making it silly or disingenuous to conceptualize an audit study as an investigation of "all search engine providers" as opposed to an investigation of Google.[9] This means that effective algorithm audits may be more similar to investigative reporting or prosecutorial investigations than broad representative surveys of a class of actors. This also means that algorithm audits must have an inherently oppositional character, creating problems that were not faced by traditional audit researchers. Algorithm audits are likely to generate organized and powerful corporate opposition from their targets, and particularly so from the guilty.

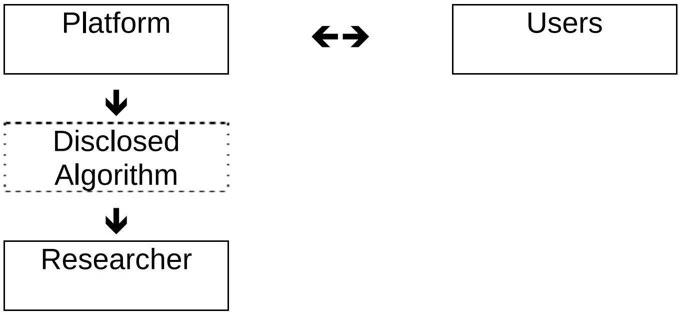
In this context, we will elaborate a variety of idealized audit study designs that might be used to investigate the normative problems introduced earlier in this paper. Some of these idealized study designs vary quite a bit from the field experiments used in traditional audit studies about fair housing and employment, but we feel the use of the word "audit" in its dictionary meaning is useful enough to stretch a point. Indeed, some of our proposals also employ the meaning of "audit" from financial accounting. We therefore offer these designs as algorithm audits: a starting point for a future conversation about how researchers might look inside the black box of the algorithm to pursue knowledge about pressing public problems.

[9] See: http://returnonnow.com/internet-marketing-resources/2013-search-engine-market-share-by-country/

We will now review several possible audit study designs in detail.

1. Code Audit (Algorithm Transparency)

In our introductory SABRE case and its context of the then-regulated US airline industry, all of the algorithms used to sort airline reservation system search results were eventually made public by an administrative law regulation. If researchers worried about algorithmic misbehavior could simply obtain a copy of the relevant algorithm, this could be a kind of algorithm audit. In order to easily compare this research design to our later proposed research designs, we will produce a simple diagram for each design. To begin, this design might be depicted this way:



Unfortunately, today Internet platforms consider their algorithms to be valuable intellectual property and also aim to conceal them using trade secret protection (Pasquale 2010; 2011). It seems unlikely that the relevant algorithms would be disclosed at all (especially illegal algorithms) unless such disclosure were somehow to be compelled. Although requiring algorithmic transparency might seem attractive from a public interest rationale, there are significant complications that may arise from such an effort.

A major problem is that the public interest disclosure of *just* algorithms might be likely to produce serious negative consequences. On many platforms the algorithm designers constantly operate a game of cat-and-mouse with those who would abuse or "game" their algorithm. These adversaries may themselves be criminals (such as spammers or hackers) and aiding them could conceivably be a greater harm than detecting unfair discrimination in the platform itself.

For example, a select few Internet platforms are already open about their algorithms, typically because they subscribe to a culture of openness influenced by the open source movement in software engineering. One such platform is Reddit (whose code is open source).[10] However, to prevent spambots from using the disclosed algorithm to attack Reddit, making its rating and commenting systems useless, a kernel of the algorithm (called the "vote fuzzing" code) must remain closed source and secret, despite Reddit's aspirations to transparency.[11]

Pasquale (2010) has proposed a solution to this problem wherein algorithms themselves could be disclosed to expert third parties who hold them in a safe escrow, permitting public interest scrutiny but not allowing the algorithm to become public. This is a noble proposal and it would dramatically improve this research design, rendering the auditing of code ordinary and

[10] http://amix.dk/blog/post/19588
[11] Thanks to Alex Leavitt for pointing this out.

thus this research design unobtrusive. (While providers might expect their algorithms to be audited, they might not know what auditors are looking for—certainly this is not the case in a legal process like a subpoena.)

Despite these positive features, in the final analysis we find this approach is unlikely to work. Algorithms differ from earlier processes of harmful discrimination (such as mortgage redlining) in a number of crucial ways. In one difference, the algorithms that affect large number of people (e.g., the Google search algorithm) are complicated packages of computer code crafted jointly by a team of engineers—they aren't a red line drawn on a map in ink. Even given the specific details of an algorithm, at the normal level of complexity at which these systems operate an algorithm cannot be interpreted just by reading it. That is, a disclosed algorithm is not going to contain a line like:

```
if ($race = NOT_CAUCASIAN) then { illegal_discrimination() };
```

It is possible that some mathematical tools adapted from the domain of information security may be brought to bear on a suspect algorithm to make determinations about its behavior, but this is unlikely to be straightforward.[12] Algorithms also increasingly depend on personal data as inputs, to a degree that the same programmatically-generated Web page may never be generated twice. If an algorithm implements the equation resulting from a multivariate regression, for instance, with a large number of variables it becomes virtually impossible to predict what an algorithm will do absent plugging in specific values for each variable. This implies that some badly-behaving algorithms may produce their bad behavior only in the context of a particular dataset or application, and that harmful discrimination itself could be conceptualized as a combination of an algorithm and its data, not as just the algorithm alone.

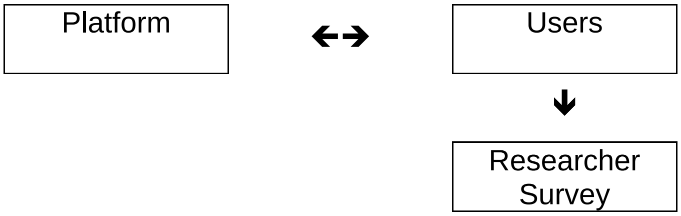
This point is made clear by recent controversies surrounding the publicly-disclosed portion of the Reddit ranking algorithm. Even with complete transparency about a particular part of the algorithm, expert programmers have been sharply and publicly divided about what exactly that part of the algorithm does.[13] This clearly implies that knowing the algorithm itself may not get us very far in detecting algorithmic misbehavior. Indeed we note that an important tactic employed by the algorithmic investigators on Reddit has been to plug data into the Reddit algorithm and look at what happens—re-uniting the algorithm with its data.

However, all is not lost. examining the outputs of a suspicious process through trial-and-error is exactly what classic audit studies were designed to do. While the code audit might be useful in some circumstances, we will now consider other research designs for algorithm audits that might address the limitations of performing code audits alone.

[12] Thanks to Jeffrey L. Vagle for this proposal.
[13] http://www.businessinsider.com/two-programmers-claim-reddits-voting-algorithm-is-flawed-2013-12

2. Noninvasive User Audit

Another form of algorithm audit—although it may be stretching a point to call it an audit—could be a noninvasive selection of information about users' normal interactions with a platform. If users agreed to answer questions about what they did online or (more thoroughly) agreed to share all of their search queries and results, for example, it might be possible to infer something useful about the operation of a platform's algorithm. Conceptually, this might be drawn:



Of course this design has the advantage of not perturbing the platform itself. As it involves only asking the users questions in a traditional social science survey format, it avoids behavior by the researcher that might appear malevolent to the platform and could be taken to be a terms of service violation.[14] Yet without the benefit of any manipulation or random assignment to conditions this is not an experimental design, and it may be quite difficult to infer causality from any particular results. For instance, a finding that a disadvantaged demographic tended to receive search results that are inferior in some way might be the result of a variety of causes.

In addition, this design introduces a serious sampling problem—the users queried in must exhibit enough variation on the dimensions of interest that they provide a useful test of a platform's harmful discrimination along those axes. If a sample of users doesn't include an important axis of discrimination (or the sample doesn't include enough of it), that manipulation will not be detected. This sampling problem is an extremely difficult one, implying that a weighted, representative racially diverse and economically diverse sample must be recruited to investigate racial and economic injustice. This could require great expense and effort when compared to other designs.

More significantly, a survey-based audit that relies upon any kind of self-report measure is likely to introduce significant validity problems that may be insurmountable. This is an unobtrusive research design with respect to the platform but it is an obtrusive research design for the users who participate. Unreliable human memories and cognitive biases will introduce important sources of error that must be controlled. Even after accounting for these, in other survey research designs demand bias has been found to cause error rates as high as 50% when comparing self-reported behavior to measured behavior.[15] Relying on users to self-report

[14] We did not include an audit design that involves surveying the platform because the oppositional nature of algorithm audits discussed earlier makes it unlikely that, for instance, criminal platforms would freely disclose their own crimes.

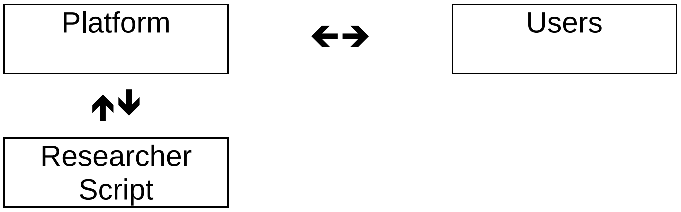
[15] For instance, an extremely unreliable high demand self-report question has been found to be "Did you vote in the last election?" because it carries with it a great deal of social pressure to say "yes." As a result, many respondents lie. Algorithm audits involving surveys might be salted with high demand questions

what platforms are doing to them is thus unlikely to work in some situations. Some more automated data collection (screen shots, saved Web pages, browser-plug ins) would mitigate this problem and might be a viable route to employing this research design at a large scale.

Still, to the degree that this design is obtrusive to users, it precludes the investigation of extremely important domains of algorithmic misbehavior such as health and financial transactions. Any domain that raises significant privacy concerns and any axes of discrimination that people are sensitive about discussing—these plainly include race and income, for instance—will be difficult to effectively investigate using this research design.

3. Scraping Audit

In a scraping audit, a researcher might issue repeated queries to a platform and observe the results. These queries might be as simple as requests for Web pages, just as a journalist for *The Atlantic* queried Netflix URL stems repeatedly to determine that there are 76,897 micro-genres of movies produced by the Netflix recommender algorithm in 2014, from "Visually-striking Foreign Nostalgic Dramas" to "Critically-acclaimed Emotional Underdog Movies."[16] This might be depicted as:



Note that the arrows linking the researcher and the platform go two directions in this diagram, signifying that the researcher is intervening in the platform in some way, even if this simply means by making requests. Although this distinction is subtle, in this design the researchers are positioned on the left because they are not acting like users. They may be accessing the platform directly via an API or they may be making queries that it is unlikely a user would ever make (or at least at a frequency a user is unlikely to ever make). They thus have a distinct relationship with the platform in this research design and it is not a relationship like that of a user.

In this design it is quite likely researchers will run afoul of a platform's terms of service and the US Computer Fraud and Abuse Act (CFAA). The CFAA has been criticized as over-broad legislation that criminalizes unauthorized access to any computer, where authorization can be defined by the Web site operator. The inherently oppositional nature of an algorithm audit makes it difficult to imagine that an operator will consent to scraping.

For instance, platform terms of service are often written to forbid the automatic downloading of any information from a Web site, even if that information is public. In one prominent use of the CFAA last year, Andrew Auernheimer was sentenced to 3.5 years in

that researchers are not aware of because the demand characteristics of these phenomena are not yet well-understood. Perhaps: "How many friends do you have on Facebook?" ("A lot!")

[16] http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

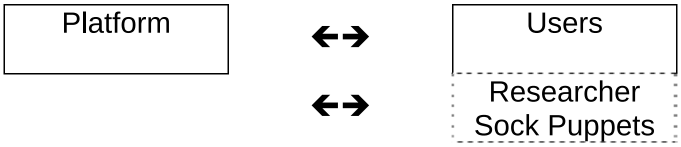
prison for sending HTTP GET requests to AT&T's public Web site.[17] Other parties found guilty of violating CFAA include a company that scraped publicly-available information from Craigslist.[18] These uses of the CFAA have been denounced by security researchers, who have commented that "We could all go to jail for security research at any moment, and a jury would happily convict us."[19] Such scraping remains common in the computer science community for a variety of purposes, yet nonetheless one of the co-authors of this paper recently received legal advice recommending that such scraping was almost certainly illegal under CFAA.

Terms of service also present a problem. While a Terms of Service document on a Web site may not have the force of law or even operate as a contract, some scholarly associations have taken the position that Internet researchers must not violate the Terms of Service agreements of Internet platforms, making many the results from designs like this one potentially impossible to publish in some venues.

Note that in this scenario there is also no randomization or manipulation, leading to the same problems with inferring cause that we encountered with the noninvasive user audit. However this might be a useful design when an audit must investigate information that is publicly available on the Web.

4. Sock Puppet Audit

A sock puppet audit is essentially a classic audit study but instead of hiring actors to represent different positions on a randomized manipulation as "testers," the researchers would use computer programs to impersonate users, likely by creating false user accounts or programmatically-constructed traffic. Although this obviously provides a great deal of control over the manipulation and data collection, it also creates a number of difficulties. Overall this design might be depicted as:



First, it involves a deception: the researcher is inventing false data and injecting it into the platform, hoping that the sock puppets cannot be distinguished from actual users. This is likely to produce the same legal (CFAA) problems as outlined in the discussion of the previous design.

However, it is also more likely that the injection of false data could be claimed to be harmful by the provider itself. It is an interesting unanswered research question to estimate what amount of false data injection would be required to prove an algorithm is misbehaving and yet still inject data in such a way that the algorithm being investigated is not itself perturbed by the injection of false accounts. Given the large size of many dominant Internet platforms, it is likely that this problem could be resolved if the legality of injecting false data were also resolved.

[17] http://www.wired.com/2013/03/att-hacker-gets-3-years/
[18] http://www.wired.com/2013/08/ip-cloaking-cfaa/
[19] https://twitter.com/0xcharlie/statuses/313669047466225667

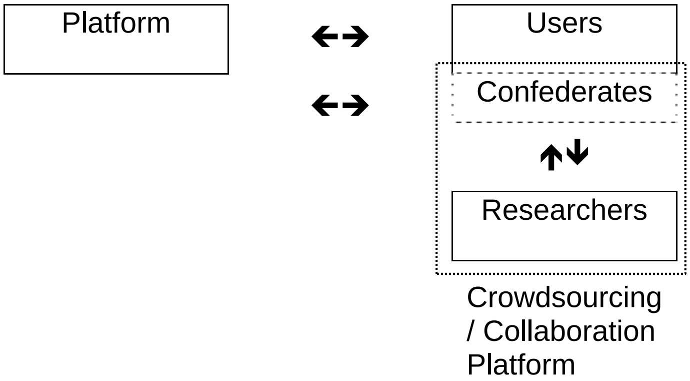
Note that many of the problems just outlined for this design would disappear if the researchers used an actual classic audit study design and hired human testers instead of using scripts. However, we anticipate that the problem of detecting misbehaving algorithms, unlike housing discrimination, is not likely to be detectable in any meaningful way without a large number of tests. We therefore expect that using research assistants or doing multiple tests ourselves is unlikely to produce enough data for a valuable audit of a significant problem, particularly where more than one variable might need to be considered at the same time, such as in an exploratory case where it is not clear what kind of harmful discrimination is likely to be occurring.

This design also would likely have benefits in that it allows the effective investigation of sensitive topics in otherwise difficult-to-access domains. Artificial user accounts can be used to investigate features of systems that are not public, and sock puppets can be assigned to normatively important categories of discrimination that may be difficult to talk about (e.g., HIV seropositive status, sexual orientiation, poverty).

However, due to the anticipated legal difficulties with this design we find it unlikely that this design will be practically workable in very many important situations. One exception might be that a virtuous algorithm provider might consent to an independent sock puppet audit as a public relations or trust-building tactic. For instance, while it was not a sock puppet audit, it appears that Netflix agreed to (or at least did not object to) the investigation of its subgenres (a scraping audit) operating in its recommendation algorithm by journalists at *The Atlantic Monthly.* Netflix may have seen the investigation as free publicity promoting an area of corporate pride (their recommender algorithm) where they had nothing to hide.

5. Crowdsourced Audit / Collaborative Audit

Finally, our fifth design slightly amends the previous one by using hired users as testers instead of computer programs. It might be depicted as:



We note that only a crowdsourcing platform like Amazon's Mechanical Turk is likely to allow a large enough group of "testers" (users) to be recruited to make this research design worthwhile. Simply hiring testers as classic audit studies have done will not work as these will not provide enough data quickly enough, as discussed previously. Using some form of semi-automated crowdsourcing also surmounts a variety of problems with the earlier research designs listed above. Although the previous four research designs may have utility in some situations, we find this approach to be the most useful and promising for future work in this area.

A sock puppet audit and a crowdsourced audit only differ in one respect: in the latter design the tester is a human. While it seems absurd that it would be illegal to use a computer program to act as a user yet legal to hire a person to act like a computer program, indeed this appears to be the case. The CFAA restrictions on unauthorized use are unlikely to be triggered by a genuine human user interacting with an online platform, especially a platform that accepts new user accounts from anyone on the Internet. A great deal of effort goes into preventing automated use of free Internet platforms by malicious bots and computer programs, and these are expressly prohibited in the Terms of Service given by platform providers. Yet an actual human being who is paid to do something with an Internet platform is unlikely to trigger any of these prohibitions.[20]

It is possible that a crowdsourced audit could be undertaken with no deception at all. A simple example study employing this design might pay Internet users from all over the world a small amount of money to perform particular Web searches on a search engine and save the results, providing these results to the researchers. The design would then vary location without deception. The use of predetermined Web searches seems so trivial a use of an online platform as to be ridiculous as a case of injecting false data. A similar research design has already been used by Rogers (2013) to, for instance, query Google for "human rights" from a variety of locations around the world. This was not an algorithm audit, as the goal was to diagnose the concept of "human rights" internationally by assuming that localized country-specific Google search results provided some useful proxy that measured sentiment in those countries.[21]

One drawback of this design is its cost, yet rather than employing Amazon's Mechanical Turk and payments made to the hired users, it might be possible to envision a "Collaborative

[20] Thanks to Kevin Werbach for pointing this out.
[21] And indeed interesting variations were found by country—see Rogers (2013).

Audit" where the testers are volunteers who are perhaps interested in public interest problems associated with algorithms. These collaborative auditors might volunteer to act as "secret shoppers" and could then avoid at least some of the sampling problems of a noninvasive user audit because, as confederates, they would not need to act like themselves and could explore controversial or sensitive topics and demographics by assuming them. This still leaves the ethical problem of injecting false data into the system, which would need to be considered in a manner similar to that proposed in the above discussion of the sock puppet audit.

An example of a kind of proto-collaborative audit might be BiddingForTravel (http://biddingfortravel.yuku.com/), a Web site where frequent users of "name-your-own-price" travel sites such as Priceline band together to collaboratively figure out the best strategy to bid for travel at different quality levels, in different cities. For instance, Priceline users visit BiddingForTravel just after using Priceline's "name-your-price" bidding function. They then report their own bids for a hotel room and their success or failure via an online form. Other users work together to analyze this aggregate information and offer each other travel advice.

One result of BiddingForTravel's informal collaborative audit has been an allegation made by BiddingForTravel users that the "average bid" displayed on the Priceline Web site as advice given to users to help them bid is likely to be false.[22] Combining a slightly more systematic approach with an involved user population like users of BiddingForTravel might increase the validity of conclusions reached by sites like BiddingForTravel, as the site is now fairly informal. At its most hopeful, the collaborative audit design might allow the creation of networks of concerned volunteers who provide accountability in a variety of algorithmic domains, from AngryPatients to WaryCreditCardUsers, perhaps in collaboration with researchers (as a form of "citizen algorithmic science") and/or nonprofit organizations dedicated to the public interest.

Conclusion: Regulating for Auditability

In this paper we have introduced the problem of "rigged" algorithms and normatively suspect algorithmic discrimination. We then introduced the social scientific audit study, a research design considered to be the most rigorous way to test for discrimination in housing and employment. After outlining some of the challenges of audit studies as they are traditionally done, we proposed the idea of "algorithm audits" as a research strategy that would adapt the social scientific audit methodology to the problem of algorithms. Although other algorithm audit designs are certainly possible, we outlined five idealized designs that empirical research projects investigating algorithms could take, discussing the major advantages and drawbacks of each approach. The five designs were the (1) code audit, (2) noninvasive user audit, (3) scraping audit, (4) sock puppet audit, and (5) collaborative or crowdsourced audit. Writing across these designs we also introduced a number of significant concerns and distinctions between them. (For instance, we examined the trade-offs between auditing *code* vs. making interventions at other points of the sociotechnical system.) While we hope these designs are useful as a guide and agenda for researchers interested in algorithmic discrimination, it is also important to pause at this point and reflect more broadly on the larger context that gives rise to these concerns and these designs.

[22] http://biddingfortravel.yuku.com/topic/2560/Hotel-FAQ#.U3rLsPldWSo

The behavior of American Airlines and SABRE which introduced this paper is probably unexceptional to scholars of business history. With history as a guide, one would expect invidious sorting not just as aggressive competition that may violate antitrust laws but also as violations of civil rights that resemble longstanding forms of injustice from real estate “redlining,” to unlawful exclusions from insurance and other services (Bowker & Star, 1999; Philips, 2005). It is the government’s role to deal with these problems, but embedding suspect algorithms in computing infrastructure makes many traditional forms of public scrutiny impossible.

Nonetheless, in order to know how to think about corrupt algorithms there is no need to digress into a discussion of human nature, reenacting Locke vs. Hobbes. The question at issue in this paper has not been whether we would expect algorithm providers to be good or evil, but what mechanisms we have available to determine what they are doing at all. One insight that arises from a comparison of our proposals to “traditional” audit studies of housing discrimination originating in the 1970s is that the social problems we are addressing are often instances of those problems we have been addressing for a long time. The migration of these longstanding problems to online environments, however, has been matched by a wholesale transformation of the legal and cultural expectations surrounding the activities that might help us to detect discrimination.

For example, while the problem of racial discrimination by landlords seems structurally to be quite the same across 40 years, in the 1970s we tolerated audit study designs that lied promiscuously while aggressively wasting the time of the people who participated in them. These deviations from the norms of practice for other research studies were deemed acceptable for the greater good. Today, however, an identical design done via computer immediately runs afoul of an overly-restrictive law (the CFAA) and overly restrictive scholarly guidelines (e.g., requiring researchers to obey the terms of service for Internet platforms).

Indeed, the movement of unjust face-to-face discrimination into computer algorithms appears to have the net effect of protecting the wicked. As we have pointed out, algorithmic discrimination may be much more opaque and hard to detect than earlier forms of discrimination, while at the same time one important mode of monitoring—the audit study—has been circumscribed. Employing the “traditional” design of an audit study but doing so via computer would now waste far fewer resources in order to find discrimination. In fact, it is difficult to imagine that a major internet platform would even register a large amount of auditing by researchers. Although the impact of this auditing might now be undetectable, the CFAA treats computer processor time as and a provider’s “authorization” as far more precious than the minutes researchers have stolen from honest landlords and employers over the last few decades. This appears to be fundamentally misguided.

As a consequence, we advocate for a reconceptualization of accountability on Internet platforms. Rather than regulating for transparency or misbehavior, we find this situation argues for “regulation toward auditability.” In our terms, this means both minor, practical suggestions as well as larger shifts in thinking. For example, it implies the reform of the CFAA to allow for auditing exceptions that are in the public interest. It implies revised scholarly association guidelines that subject corporate rules like the terms of service to the same cost-benefit analysis that the Belmont Report requires for the conduct of ethical research—this would acknowledge that there may be many instances where ethical researchers should disobey a platform provider’s stated wishes.

But more broadly, regulating for auditability also implies an important third-party role for government, researchers, concerned users, and/or public interest advocates to hold Internet platforms accountable by routinely auditing them. This would require financial and institutional resources that would support such an intervention: a kind of algorithm observatory acting for the public interest.

It also suggests that research-sponsoring organizations might promote the investigation of a variety of basic research questions that can be brought to bear on such a process—questions that often have yet to be asked, such as: How difficult is it to audit a platform by injecting data without perturbing the platform? What is the minimum amount of data that would be required to detect a significant bias in an important algorithm? What proofs or certifications of algorithmic behavior could be brought to bear on public interest problems of discrimination? A program of research on auditable algorithms could make great contributions via simulation and game theoretic modeling of these situations.

Remarkably, the proposals and research designs advanced in this paper are fundamentally both anti-capitalist and pro-competition. The need for these investigations into corporate information architectures could be embraced by those who wish to work against corporations. However, in history those most likely to investigate some forms of malfeasance like the ones we have been discussing have been a misbehaving company's competitors—this was also the source of the antitrust complaints filed against American Airlines and the SABRE system.

Finally, a shift of perspective to "accountability by auditing" implies a normative discussion about what the acceptable behavior of Internet platforms ought to be. While some instances of discrimination are obviously unfair and illegal, algorithm audits might produce situations where a comparison to standards of practice are required (as is done in financial auditing) yet no consensus or "accounting standard" for algorithms yet exists. Discovering how algorithms behave on the Internet might then lead to a difficult but important discussion that we have so far avoided: how do we as a society want these algorithms to behave?

Acknowledgement

The first author would like to thank Christopher Yoo, Jeffrey L. Vagle, Kevin Werbach, Frank Pasquale, and audience members at the University of Pennsylvania Law School for comments on this project.

Sources Cited

Barocas, Solon; Hood, Sophie; and Ziewitz Malte. 2013. Governing Algorithms: A Provocation Piece. A paper presented to *Governing Algorithms*, NYU School of Law, New York, NY. http://governingalgorithms.org/resources/provocation-piece/

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.

Copeland, D.G., Mason, R.O., and McKenney, J.L. (1995). Sabre: The Development of Information-Based Competence and Execution of Information-Based Competition. *IEEE Annals of the History of Computing*, vol. 17, no. 3: 30-57.

Edelman, Benjamin G. 2014. Leveraging Market Power Through Tying and Bundling: Does Google Behave Anti-Competitively? *Harvard Business School NOM Unit Working Paper* No. 14-112. Available at SSRN: http://ssrn.com/abstract=2436940

Edelman, Benjamin G. and Luca, Michael. 2014. Digital Discrimination: The Case of Airbnb.com *Harvard Business School NOM Unit Working Paper No. 14-054*. Available at SSRN: http://ssrn.com/abstract=2377353

Gilbert, Eric; Karahalios, Karrie; and Sandvig, Christian. 2010. The Network in the Garden: Designing Social Media for Rural Life. *American Behavioral Scientist 53*(9), 1367-1388.

Gillespie, Tarleton. 2014. "The Relevance of Algorithms."in: Tarleton Gillespie, Pablo J. Boczkowski and Kirsten A. Foot (eds.), *Media Technologies: Essays on Communication, Materiality, and Society.* Cambridge: MIT Press.

Graham, S. & Marvin, S. (2001). *Splintering Urbanism: Networked Infrastructures, Technological Mobilities, and the Urban Condition*. New York: Routledge.

Harvard Law Review. (1990, June). Note. The Legal and Regulatory Implications of Airline Computer Reservation Systems. *103 Harvard Law Review* 1930.

Milkman, K. L., Akinola, M., & Chugh, D. 2012. Temporal Distance and Discrimination: An Audit Study in Academia. *Psychological Science 23*(7): 710-717.

Mincy, Ronald. 1993. "The Urban Institute Audit Studies: Their Research and Policy Context." In Fix and Struyk, eds., Clear and Convincing Evidence: Measurement of Discrimination in America. Washington, DC: The Urban Institute Press, pp. 165-86.

Pager, Devah. 2007. “The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future.” The Annals of the American Academy of Political and Social Science, Vol. 609, No. 1, pp. 104-33.

Pager, Devah. 2009. Field Experiments for Studies of Discrimination. In: E. Hargittai (ed.) *Research Confidential: Solutions to Problems Most Social Scientists Pretend They Never Have,* pp. 38-60*.* Ann Arbor, MI: University of Michigan Press.

Pasquale, F. 2010. “Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries,” 104 *Northwestern University Law Review* 105.

Pasquale, F. (2011). Restoring Transparency to Automated Authority, *Journal on Telecommunications and High Technology Law, 9*(235).

Petzinger, T. (1996). *Hard landing: the epic contest for power and profits that plunged the airlines into chaos*. New York: Random House.

Redmond, K. C. & Smith, T. M. 2000. *From Whirlwind to MITRE: The R&D Story of the SAGE Air Defense Computer.* Cambridge, MA: MIT Press.

Riach, Peter A., and Judith Rich, 2002. “Field Experiments of Discrimination in the Market Place.” The Economic Journal, Vol. 112, No. 483, November, pp. F480-518.

Rogers, Richard. 2013. *Digital Methods.* Cambridge, MA: MIT Press.

Sandvig, Christian; Hamilton, Kevin; Karahalios, Karrie; and Langbort, Cedric. 2013. Re-Centering the Algorithm. Paper presented to Governing Algorithms, NYU School of Law, New York, NY, USA. http://governingalgorithms.org/wp-content/uploads/2013/05/4-response-karahalios-et-al.pdf

Turow, Joseph. 2013. *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. New Haven, CT: Yale University Press.

Valentino-Devries, J., Singer-Vine, J., Soltani, A. 2012, December 24. Websites Vary Prices, Deals Based on Users’ Information. *The Wall Street Journal*. http://online.wsj.com/news/articles/SB10001424127887323777204578189391813881534

Author Biographies

Kevin Hamilton, Associate Dean of Research, College of Fine and Applied Arts and Associate Professor of New Media and Painting, University of Illinois at Urbana-Champaign.

> Hamilton works as a historian, critic and artist in digital media forms to illuminate the role of mediation in modern approaches to self and society. His artworks have been shown in festivals, museums and public spaces across Europe and North America. He holds the M.S. in Visual Studies from MIT and the B.F.A. from the Rhode Island School of Design. His work has been supported by the National Science Foundation and the National Endowment for the Humanities. http://complexfields.org/

Karrie Karahalios Associate Professor, Computer Science and Co-Director, Center for People & Infrastructures, University of Illinois at Urbana-Champaign.

> Karahalios is a computer scientist whose work focuses on the interaction between people and the social cues they perceive in networked electronic spaces. Karahalios completed a S.B. in electrical engineering, an M.Eng. in electrical engineering and computer science, and an S.M. and Ph.D in media arts and science at MIT. She received the Alfred P. Sloan Research Fellowship and the Faculty Early-Career Development Award from the US National Science Foundation (NSF CAREER) in the area of human-centered computing. http://social.cs.uiuc.edu/people/kkarahal.html

Cedric Langbort Associate Professor, Aerospace Engineering and Co-Director, Center for People and Infrastructures, University of Illinois at Urbana-Champaign.

> Langbort is a game theorist whose research focuses on distributed decision and control theory and its application to large-scale public infrastructures. He received the Faculty Early-Career Development Award from the US National Science Foundation (NSF CAREER) in the area of game theoretic approaches to cyber-security. He was previously a Postdoctoral Scholar at the Center for the Mathematics of Information, Caltech and holds the Ph.D. in Theoretical and Applied Mechanics from Cornell University. http://aerospace.illinois.edu/directory/profile/langbort

Christian Sandvig, Associate Professor, Communication Studies and School of Information, University of Michigan.

> Sandvig studies the development of new communication infrastructures and their consequences for public policy. He is a Faculty Associate of the Berkman Center for Internet & Society at Harvard University. He received the Ph.D. in Communication from Stanford University and previously served as Markle Foundation Information Policy

Fellow at Oxford University. He has been named a “next-generation leader in technology policy” by the American Association for the Advancement of Science and he received the Faculty Early-Career Development Award from the US National Science Foundation (NSF CAREER) in the area of human-centered computing. http://www.niftyc.org/