Article

SM+S
social media + society

Social Media + Society
January-March 2022: 1–18

DOI: 10.1177/20563051221077021
journals.sagepub.com/home/sms
SAGE

Bartleby: Procedural and Substantive Ethics in the Design of Research Ethics Systems

Jonathan Zong[1] and J. Nathan Matias[2]

Abstract

The lack of consent or debriefing in online research has attracted widespread public distrust. How can designers create systems to earn and maintain public trust in large-scale online research? Procedural theories inform processes that enable individuals to make decisions about their participation. Substantive theories focus on the normative judgments that researchers and participants make about specific studies in context. Informed by these theories, we designed Bartleby, a system for debriefing participants and eliciting their views about studies that involved them. We evaluated this system by using it to debrief thousands of participants in a series of observational and experimental studies on Twitter and Reddit. We find that Bartleby addresses procedural concerns by creating new opportunities for study participants to exercise autonomy. We also find that participants use Bartleby to contribute to substantive, value-driven conversations about participant voice and power. We conclude with a critical reflection on the strengths and limitations of reusable software to satisfy values from both procedural and substantive ethical theories.

Keywords

research ethics, data collection, privacy, consent, debriefing, opt out, non-participation

Introduction

A series of high-profile research scandals in the past decade has led to calls for improvements and standardization in research procedures. In 2014, after public outrage about a study that altered the contents of hundreds of thousands of Facebook news feeds (Kramer et al., 2014), many called for studies to include informed consent, debriefing, and a chance for participants to opt out (Grimmelmann, 2015). In 2020, residents of the state of Illinois sued IBM when researchers included their online photographs, without consent, in a research dataset initially prepared by Yahoo (Olivia Solon, 2019; Stoller, 2020). Yet surveys of research practices in social computing have found that many academics avoid consenting or informing participants because they believe it to be impractical (Vitak et al., 2016).

Scholars, critics, and policymakers have argued that these research projects failed by ignoring individual autonomy (Grimmelmann, 2015). By preventing people from learning about research and data collection, researchers failed to give individuals a chance to choose whether to participate in a study or choose how their data would be used. Critics argued that the needs of research and the autonomy of participants could be maintained with the right procedures in place. After all, the communication technologies that enable large-scale data collection have also enabled new design possibilities for innovations in research procedures (Grimmelmann, 2015). Since then, lawmakers in the European Union (EU) and the United States have passed regulations that require data processors (but not academic researchers) to inform people about the data they collect and provide them with a chance to have it removed (California Consumer Privacy Act [CCPA], 2018; General Data Protection Regulation [GDPR], 2018). In parallel, researchers have suggested that more studies include a debriefing stage, where participants are told the details of a study and given a chance to opt out (Desposato, 2018; Grimmelmann, 2015).

[1]Massachusetts Institute of Technology, USA
[2]Cornell University, USA

Corresponding Author:
Jonathan Zong, MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA.
Email: jzong@mit.edu
Twitter: @ohnobackspace

To advance research ethics procedures that protect individual autonomy, prevent abuses of power, and promote public trust, we introduce Bartleby: a system that delivers research ethics procedures for large-scale online studies. Bartleby provides a user interface that researchers can customize to the details of their study. Using Bartleby, researchers can automatically send each of their study participants a message directing them to a website where they can learn about their involvement in research, view what data researchers collected about them, and give feedback. Most importantly, participants can use the website to opt out and request to delete their data. The system is named after the titular character in Herman Melville's short story *Bartleby, the Scrivener*. Over the course of the story, Bartleby opts out of completing various requests. Instead, he states simply that he "would prefer not to" (Melville, 1853).

We designed Bartleby in response to public criticism of academic research conducted by Facebook and IBM. Because these studies lacked informed consent, people who were included in the research data likely never found out whether or not they were in the study. Scholars have suggested that the ethics of these studies could have been improved with minimal effort by debriefing participants—in other words, notifying them of their participation and offering a chance to opt out (Grimmelmann, 2015). With Bartleby, researchers can automate debriefing with few adjustments to their existing research processes. By creating and deploying Bartleby in the field, we also demonstrate that large-scale debriefing can be simple and practical, despite claims to the contrary.

We also present the Bartleby system as a case study for critical thinking about the design of research ethics procedures. What does it mean for a research ethics system to be successful? Because the purpose of debriefing is to protect participants' right to autonomy, it can serve moral and procedural purposes regardless of whether any individual participant exercises their rights by opting out. For that reason, we include an extended discussion of the system in light of two kinds of ethical theories drawn from feminist and political philosophy: procedural and substantive theories. Procedural theories are concerned with the abilities and limitations of scalable, repeatable procedures to protect individual autonomy. Substantive theories are concerned with the values upheld by the research, and the use of power in deciding those values.

As US-based researchers, our frameworks for thinking about the role of autonomy in research ethics may differ from those of a global audience. Scholars have questioned the idea of universally-applicable research ethics, arguing instead that "ethical codes are never universal and are geographically sensitive" (Zhang, 2017). While our work is primarily informed by the institutional environment of US ethics regulation and university review boards, we acknowledge the importance of respecting different values and expectations that arise when conducting research on social media platforms with global reach.

In this article, we summarize the design challenge of creating ethics systems, describe the design of the Bartleby system, and present empirical evidence from two 2020 field studies on Twitter and Reddit involving 4,766 and 1,342 participants. We also review the design of the system through the lens of procedural and substantive theories of ethics. In addition to presenting the Bartleby system, our work demonstrates how procedural and substantive theories can guide the design and evaluation of research ethics systems.

Debriefing Participants and Opting Out of Social/Behavioral Research

Debriefing is a research ethics procedure that happens at the end of a study, after data collection has concluded (The CITI Program, 2018). During debriefing, researchers notify participants that they were involved in research and disclose information about study procedures. Participants are informed about the data researchers collected, and have an opportunity to exercise their agency by opting out and withdrawing their data from the study.

Debriefing serves a distinct purpose from other procedures. For example, informed consent is a procedure that happens before a study begins. Participants are given information about what will happen if they are involved in research, and can give or withdraw their consent based on that information. While debriefing and informed consent are not mutually exclusive, debriefing is unlike informed consent in that it can be used in study designs where participants are unaware of their inclusion in data collection. In this article, we focus on debriefing due to the fact that this kind of research is increasingly common on social media platforms. While there is a broader ongoing conversation in research ethics about whether the increasing prevalence of large-scale data collection on unaware individuals is acceptable (Yeshimabeit Milner, 2019), our work starts from the premise that there exist at least some cases where valid methodological reasons prevent the use of informed consent. For instance, it is impractical to seek prior informed consent for a study that observes Twitter discussions between certain dates, because researchers cannot know who will participate in discussions ahead of time. Furthermore, knowledge about the study might also bias the behavior of subjects who are being observed, affecting the validity of the results. In some studies, consent procedures might introduce selection bias—some people might be more likely to consent than others due to group membership. Indeed, researchers in political science and philosophy have studied the moral significance of ethics procedures other than prior informed consent—such as proxy consent (Humphreys, 2015) and hypothetical consent (Enoch, 2017)—due to their necessity in certain practical situations.

Debriefing has most commonly been used offline in fields like psychology and behavioral economics to manage the ethics of deception-based studies (Hertwig & Ortmann,

2008). Because this research has usually happened in-person in a laboratory, participants know that they are a part of research. However, they are not informed about the true nature of the research until researchers debrief them in-person, providing an immediate opportunity to ask questions and address harms. Online research creates a new situation where participants may not be outright deceived, but potentially never become aware of their inclusion in research unless debriefed. We use the term *non-consented research* to distinguish this situation from other research designs where debriefing has previously been used.

Although debriefing is a well-understood ethics procedure, it is rarely used in non-consented online social science research (Desposato, 2018). In non-consented research, debriefing serves two important purposes: informing participants and creating opportunities for them to opt out. Informing research subjects of their participation, providing opt out opportunities, and opening communication between subjects and researchers all serve to increase the agency of participants over their involvement in research. For observational studies, debriefing can be thought of as retroactive informed consent. It could be argued that debriefing has equivalent moral significance to informed consent in observational studies, because data could potentially remain unprocessed until consent is given and can be deleted when consent is withdrawn. However, for field experiments, post hoc debriefing cannot undo participant exposure to interventions. Debriefing is not an equivalent replacement for upfront informed consent in field experiments, though it is one of the better alternatives when consent is not an option.

As an ethics procedure, debriefing cannot always address participant concerns about the normative content of the research. In prior empirical work on participant perceptions of research ethics, adding debriefing to study designs that lacked consent did not significantly change the perceived acceptability of research (Desposato, 2018). More debriefing, then, does not necessarily resolve ethical issues with studies that are objectionable for reasons other than whether or not they were consensual.

Nonetheless, scholars have argued that making debriefing more commonplace in online research would be an achievable yet significant advancement in research ethics. Writing about the Facebook Emotion Contagion study, Grimmelmann argues that "standardized debriefings could easily have been given via email or private message to the users who were unwittingly drafted into the studies" (Grimmelmann, 2015). Systems like Bartleby that provide debriefing at scale contribute to research ethics by lowering the technical burden required for researchers to implement ethics procedures. Grimmelman argues that "as it becomes easier to do more for participants, researchers should, because there is less and less reason not to" (Grimmelmann, 2015). Prior work has highlighted how shared tools that assist ethical compliance have benefited participants (Bravo-Lillo et al., 2013). Bartleby makes it less likely that researchers will forgo debriefing due to impracticality or technical burden, which will create more opportunities for participants to exercise individual autonomy. While greater access to debriefing opportunities will not solve all problems in research ethics, debriefing improves participant agency compared with the alternative.

Related Work: Researching Research Ethics

Many scholars across disciplines including philosophy, history, law, and medicine have contributed scholarship on research ethics. In this article, we are engaging with prior work on empirical research ethics alongside research on the design of digital ethics procedures.

Empirical Research Ethics

Many empirical approaches to research ethics focus on understanding the views of people who participate in research. Scholars have observed that public scandals such as the Facebook Emotion Contagion study can arise from differing expectations between researchers, technologists, and the public (Hallinan et al., 2020). Participants contended that conducting research on the platform exceeded Facebook's purpose and worried about manipulation and abuse of power by researchers. Ethics procedures can allow participants to hold researchers accountable by providing ways to opt out and report misconduct. Procedures can also provide ways for researchers to receive feedback from research subjects, which can help researchers better understand how a given study might violate participant expectations.

Prior work has highlighted the importance of contextual factors—such as privacy expectations and perceived benefit—in subjects' willingness to participate in research. On Twitter, many people are unaware that their public posts can be included in research without consent (Fiesler & Proferes, 2018). However, attitudes regarding research vary depending on the specifics of the study. Subjects are more willing to participate in research that has clearer benefits, although this does not offset the loss of trust experienced when researchers do not seek consent (Desposato, 2018).

Importantly, participants may not always share expectations, definitions, and categories with researchers. Empirical work has found that commonsense notions of consent held by laypeople "[contradict] prevailing normative theories of consent," and sometimes mistakenly permit deceptive practices like fraud (Sommers, 2020). Consequently, researchers may be surprised by participant expectations. Indeed, research ethics regulations exist to protect the public from scientific norms that differ from public expectations (Rothman, 2017).

By designing and evaluating Bartleby, we contribute to empirical research ethics by providing an additional method for participants to express their preferences and beliefs to researchers. Many studies in empirical research ethics survey

participants to see how they would respond in hypothetical scenarios, or focus on the details of a particular case. In contrast, feedback or opt-out decisions elicited through Bartleby are situated in the context of the particular real-world study being debriefed.

Design for Research Ethics

Research is increasingly delivered digitally, and researchers and designers have investigated how to design software and procedures relevant to research ethics. The field of Human-Computer Interaction (HCI) can offer valuable contributions to scholarly conversations about research ethics, since so many matters in research ethics hinge on the design and user experience of ethics procedures.

Many designers work to make research ethics processes cost-effective for university bureaucracies, maximizing research output, and minimizing compliance risks. Many companies offer white-label institutional review board (IRB) systems, which universities buy and apply their own branding to. Examples of commercial products include Cayuse IRB, Kuali Protocols, iRIS, and others (Geier, 2019). These systems provide user interfaces for researchers to submit protocols for IRB review, and for IRB staff to quickly review large numbers of protocols. Case management systems like EthicsPoint (Hyatt, 2005) also provide a process for participants to anonymously report researcher misconduct.

One way to scale research is to design standardized consent forms for participants, which can be evaluated for readability using graphic design principles and cognitive measures (Arora et al., 2011; Bhansali et al., 2009; Hochhauser, 2000; Terblanche & Burgess, 2010). Researchers have also empirically evaluated the design of consent forms for participant comprehension and awareness of legal implications (Akkad et al., 2006; Tait et al., 2005; Wright, 2012). For example, studies have found that shorter consent forms are better understood by participants (Dresden & Levitt, 2001; Epstein & Lasagna, 1969). However, critics note that efficient forms may not always lead to outcomes that protect participants. For example, researchers have found that people sign consent forms even when they are designed illegibly, concluding that consent forms do more to facilitate submission to authority than protect participant autonomy (Jacob, 2007).

Designers have also worked to help researchers deliver digital equivalents of paper-based research ethics procedures. In psychology, designers of survey software have used pop-up windows to deliver debriefing information for consented participants who exit surveys before completion (Kraut et al., 2004). Researchers have also considered the tradeoffs of using digital signatures to legally document informed consent (Barchard & Williams, 2008). More recent work has explored eConsent systems that fully replace paper documentation (Coiera & Clarke, 2004; Kim et al., 2017). Researchers have evaluated eConsent's effectiveness in terms of factors such as trust, scalability, and user experience (Chen et al., 2020). As scholars of Feminist HCI have observed, affirmative consent involves more than simply providing users with a checkbox (Im et al., 2021). Wilbanks notes that eConsent procedures are usually implemented as single-point transactions, and proposes to "transform informed consent into an ongoing relationship of trust-based permission" in a digital context (J. Wilbanks, 2018). This involves not only delivering consent procedures digitally, but also designing interactive experiences to ensure participant comprehension and ongoing engagement in consent procedures.

As part of Wilbanks' work, Sage Bionetworks released an open-source toolkit of interface components for designers to adapt into informed consent user experiences. This work has also been adapted into Apple's ResearchKit framework for iOS developers (J. T. Wilbanks, 2020). Typical user experiences with these tools involve a series of interactive concept assessments before subjects are presented with a consent form to sign.

Novel ethics procedures are not guaranteed to increase protections for participant autonomy. As Wilbanks points out, "it is just as possible to use the visual interface to obscure [concepts] as it is to . . . reveal them" (J. Wilbanks, 2018). Outside of research ethics, consent management platforms have been widely adopted by tech companies in response to the European Union's General Data Protection Regulation (GDPR). Much design effort in consent management has gone toward nudging people to consent through "dark patterns" or misleading user experience designs that undermine autonomy (Nouwens et al., 2020). Even without misleading designs, procedures that introduce incentives, barriers, or irrelevant information into a consent process can easily influence people to give up personal data (Athey et al., 2017).

Any research ethics system that relies on individual choice also struggles with a "consent dilemma" (Solove, 2020). Scholars have argued that this model of "privacy self-management" overburdens individuals with an impossible task of never-ending decisions within a rapidly-changing, complex information landscape (Solove, 2020). If people check a box out of resignation at the impossibility of privacy management, their recorded privacy preferences could be inconsistent with their actual preferences or behavior.

To overcome the limitations of systems based on individual autonomy, researchers have explored collective governance schemes for research ethics. For example, researchers have convened a representative group of citizens to discuss the details of genetic testing. If the representative body approves the research on the behalf of the group, individuals are offered a choice to consent to be governed by their deliberations (Desposato, 2018; Koenig, 2014). Similarly, community IRBs are formed by participants and work in partnership with institutional IRBs to review and negotiate over potential research (Bronx Community Research Review Board What is CERA?, 2016; Community IRBs & Research Review Boards: Shaping the Future of Community-Engaged Research, 2012; Liat Racin, 2016; Puneet Chawla Sahota, 2009).

Another thread of work in HCI seeks to address problems of autonomy by restructuring the relationship between participants and researchers. The CivilServant system supports moderators of online communities in designing studies with the help of researchers, and provides processes for "community debriefings" involving public discussions of research results (Matias & Mou, 2018). Research that is co-designed with participants, who are directly affected by and exert agency over how research is conducted, falls under the broader category of participatory research (Cornwall & Jewkes, 1995). However, communities are often heterogeneous and can contain multiple conflicting parties. For example, some community members may oppose the power held by a dominant group that may be working with researchers. Researchers working with online communities must navigate how they are positioned in relation to multiple conflicting social actors (Keegan & Matias, 2016).

By designing and evaluating Bartleby, we are advancing a body of design research akin to ResearchKit and eConsent that develops scalable user interfaces for common research ethics procedures. By automating and scaling the debriefing procedure, we extend rights-based autonomy protections to a large number of online research participants.

Bartleby: Research Debriefing System

Bartleby is a system that automates debriefing for large-scale social and behavioral experiments online. Bartleby consists of a message-sending script to invite participants to start the debriefing process, and a website that provides debriefing information and an opt out form. It is available as open source software under an MIT license at *https://github.com/jonathanzong/bartleby*.

Design Values

HCI designers and researchers routinely grapple with the values, ethics, and politics of technologies (Shilton, 2018). In the process of creating Bartleby, we listed values that we think research ethics systems should be designed toward. Different systems may reflect these values in different ways, and to varying degrees. Whatever their goal, the designers of any research ethics system will encounter questions of informedness, agency, and scale.

Informedness. The goal of debriefing procedures is to provide people with the capacity to make an informed decision about research participation. Informedness is thought of as a state of understanding that people can achieve given enough information and guidance (Rothman, 2017). Debriefing interfaces can inform participants about the purpose of the research and what data were collected about them. They may also provide ways for participants to ask questions to further guide decision-making. The process of informing and clarifying is normally facilitated by informed consent, which makes it essential to debriefing in non-consented research.

Agency. Debriefing procedures provide participants with control over their involvement in research. Participants in debriefing must be able to withdraw from the experiment by opting out. They can also give feedback to researchers and address any harms that may arise. These mechanisms of accountability ensure participants can exercise their right to individual autonomy. University IRBs also typically provide language in consent forms that lets participants know that they can contact the IRB if they suspect researcher misconduct. Debriefing procedures can provide pathways for participants to access other, more powerful procedures if necessary. Even if these other procedures are never activated, making them available to participants can increase their agency.

Scale. As more people are included in online research, researchers interested in exploring the potential for large sample sizes to contribute to knowledge must also think about how to protect the autonomy of large numbers of people. As Gillespie notes, scale is more than size; "scale is about . . . how a process can be proceduralized such that it can be replicated in different contexts, and appear the same" (Gillespie, 2020). In our work, we see questions of scale arising whenever ethics procedures are called on to manage the autonomy of more people than a research team could interact with on an individualized basis, without the help of automation. Ethics procedures can outline standard practices that can be repeated efficiently and reliably for many people, across many studies. Procedures must navigate the tension between being customizable enough to adapt to different people and studies, while being standardized enough to be reusable.

Debriefing User Experience

The Bartleby user flow has three steps: the invitation message, the login page, and the debriefing page.

Invitation Message. Participants enter the debriefing process when they receive an invitation message notifying them about the research study (Figure 1) and inviting them to debrief. The message includes basic information about the purpose of the study and a link to the debriefing website. If the participant clicks the link in the message, they arrive at the login page for the study.

Login Page. Each study that uses Bartleby for debriefing will have its own login page. On this page, participants can view a more detailed explanation of the research question of who the researchers are, and why they are being debriefed (Figure 2). At the bottom of the page, they can click the button to log in with their social media account. Bartleby uses the Reddit and Twitter APIs to log users in. This means that

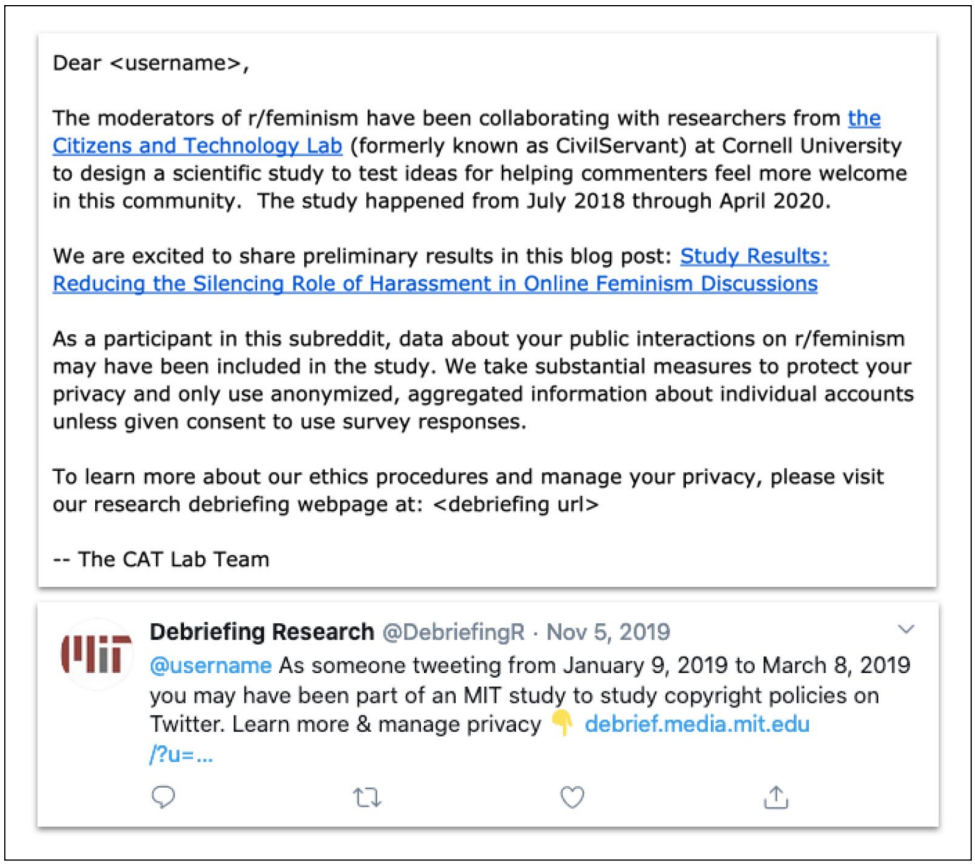

Figure 1. Debriefing invitation messages sent to participants eligible for debriefing. The messages included a link to the debriefing login page for each study.

the platforms handle authentication, and Bartleby only handles a username and public account information once authenticated. Once logged in, the system matches their username against a list of study participants stored in a database. If the account attempting to log in is not in this list—for example, if a study participant forwarded their debriefing message to an account that was not included—that account is ineligible for debriefing. Ineligible participants are redirected to a page stating that they were not included in the study. However, eligible participants will be redirected to the debriefing page containing information pertaining to their account (Figure 3).

Debriefing Page. The debriefing page confirms the participant's involvement in the study and documents the specific actions that researchers took (Figure 3a). Participants can view a table showing exactly what data the researchers collected (Figure 3b). The table comes with a description of why the data collection was necessary, and how data will be used in the future. Below the table, there is a section on opting out of the study. Participants can read information about the effects of opting out, and decide whether to check the opt-out checkbox (Figure 3c). They can return and update their decision at any point before the date listed at the bottom of this section. The cutoff time is intended to allow researchers to submit their work for publication without constantly needing to check back. If participants are curious to know the results of the study, researchers may include instructions for following up (Figure 3d). This is especially useful for studies where results have practical value for the studied population. In the case that participants feel that they have experienced harm from being included in the study, the debriefing interface includes contact information for the university IRB (Figure 3e). Finally, participants can fill out an optional survey to give feedback on the study and the debriefing interface at the bottom of the page (Figure 3f).

Researcher User Experience

Bartleby requires some initial configuration to set up a database, store database credentials in config files, define API keys to interface with Reddit and Twitter, and set up a public-facing web server to serve the login and debriefing pages for each study.

To use Bartleby to debrief a study, the researcher populates the Bartleby database with records of people who were included in the study. These records include the minimal

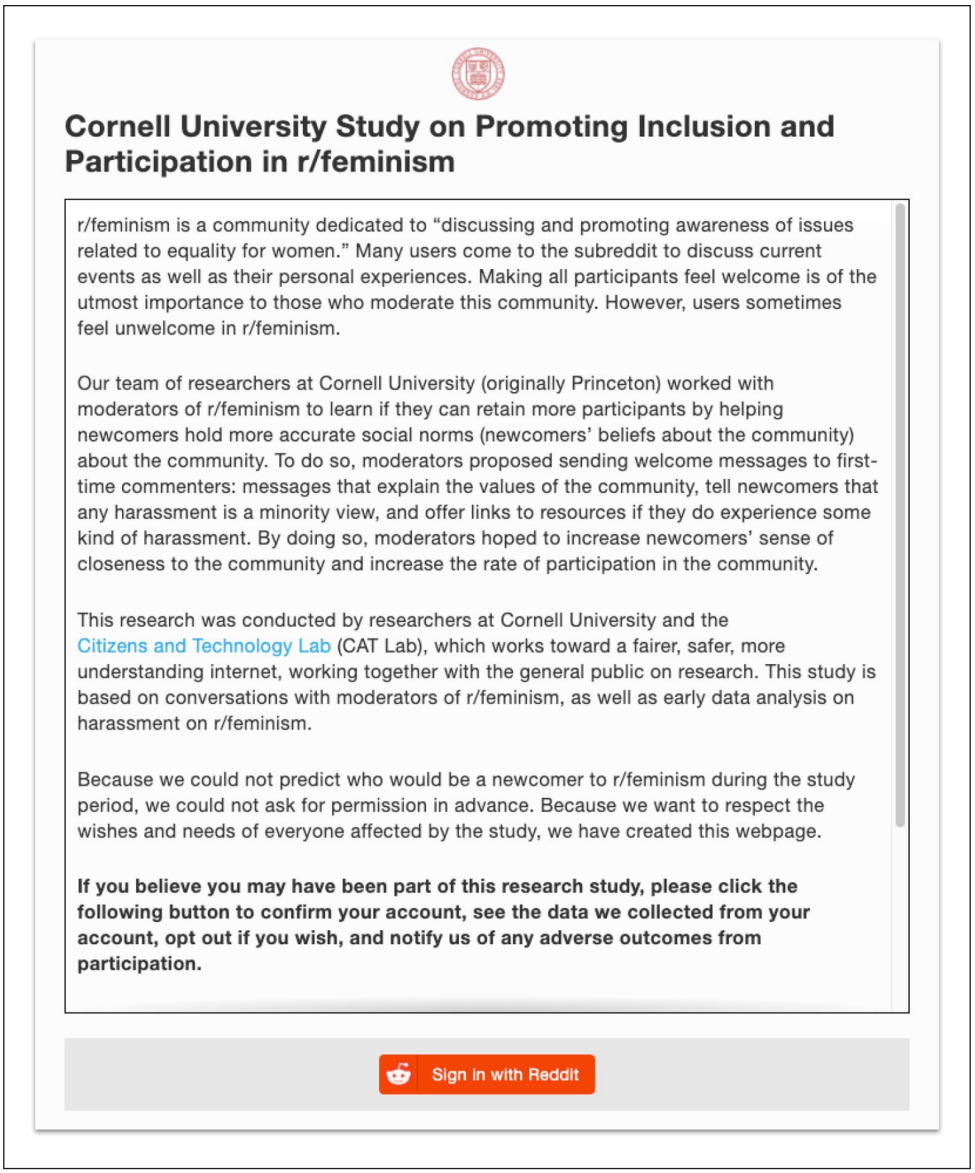

Figure 2. Experiment-specific login page for the debriefing software. People can log in with their social media account to verify their inclusion in the study and view debriefing information.

amount of information required to debrief: the user id on the platform where the study occurred (i.e., Twitter handle or Reddit username), and the user's associated study data.

The researcher also defines templates for the invitation message, login page, and debriefing page. These templates are specific to a particular study. For example, the login page template should describe the study, and the debriefing page template should include information on how to report ethics violations at the researcher's institution.

When the researcher is ready to debrief, they use a script to send an invitation message for each user record in the database. When a user logs in, their debriefing page template will be filled in with the study data associated with their username in the database.

Supporting Multiple Studies

Because we intend for researchers to use the system on an ongoing basis, Bartleby supports debriefing for multiple concurrent studies. Researchers can define experiments in the Bartleby database, and associate eligible participant accounts with those experiments. Each experiment has its own base URL on the web server. Base URLs start with a randomly generated unique identifier, so that the existence of other experiments cannot be revealed by sequentially guessing URLs from a known experiment. Each experiment is also associated with a template directory, which contains template files to render for the experiment's landing and debriefing pages. In these template files, researchers can customize

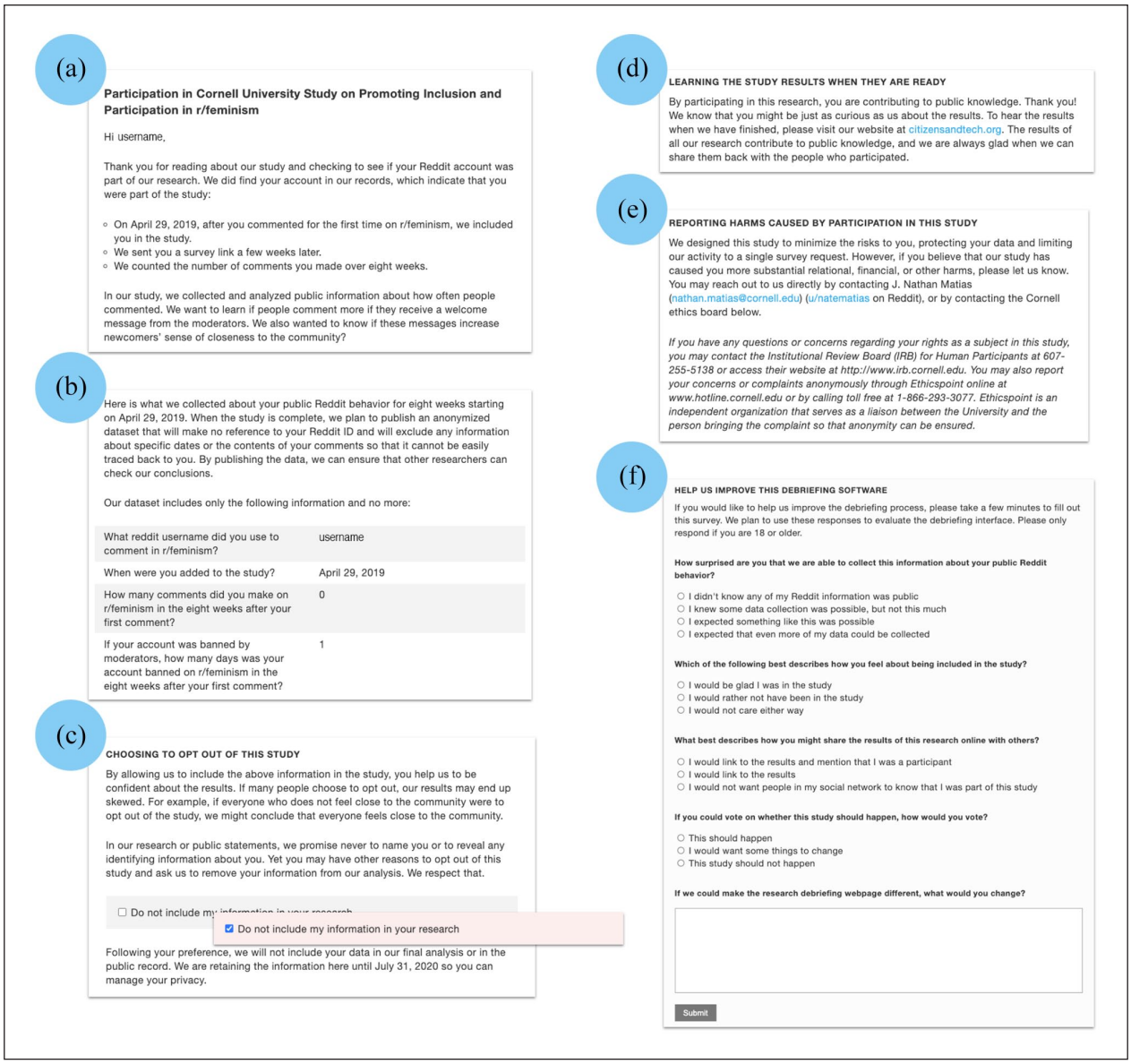

Figure 3. The components of the debriefing page. (a) Information about inclusion in the study and the purpose of the research. (b) Table showing what data were collected on the participant, with description of anonymization procedures. (c) Information on risks and benefits, with checkbox for opting out. (d) Details on following up about the study results. (e) Contact information for university review board. (f) Survey on improving the debriefing webpage.

the language for each experiment page. When a participant logs into a debriefing page, Bartleby renders the data collected on that account for that study into a table on the page.

Data Removal Procedure

In digital systems, data collection and circulation are usually ongoing over long periods of time. Because people's preferences change over time, maintaining consent over time is a general problem in the use of research data, as datasets are shared publicly by researchers and adapted for new and unexpected purposes. In our field deployments, we found that everyone who engaged with the system did so in the first few days after receiving the debriefing invitation and never returned. In practice, this method does give people some flexibility to change their mind but encounters difficulties upholding autonomy past a certain amount of time.

When people opt out of our studies, we overwrite the entries in the research dataset for that person so their data are not included and researchers retain a record that a participant opted out. We also mark people as opted-out in email lists and other databases to prevent them from being included in future analyses or communications related to the study.

Table 1. Accounts Participating in Twitter Debriefing.

Period	Participants	Contacted	Logged in	Login rate	Opted out	Opt-out rate
25 August–5 November 2019	4,766	3,631	3	0.00083	1	0.00028

How other researchers handle data removal procedures for their own studies will depend on contextual judgments of risks and benefits to participants or society. Risks will be influenced by the sensitivity of the data and possibility of de-anonymization. Analysis of benefits will need to consider the minimum amount of data needed to facilitate scientific reproducibility. In the event that certain subpopulations are more likely to opt out, data removal could potentially introduce skew affecting the validity of the results. In this case, researchers must consider whether and how to communicate about uncertainty while respecting participant privacy preferences.

Field Deployment of the Debriefing System

To test Bartleby's effectiveness, we used the system to debrief two large-scale causal studies of online behavior. For both studies, we used the Bartleby system to message participants about their involvement in the study after it was complete. Participants could log in to the Bartleby system to receive more information about the data we collected and its intended use. They could also choose to opt out of the study and provide further feedback.

Debriefing an Observational Study on Twitter

In a large-scale observational quasi-experiment on Twitter, we collected 5,171,111 public posts made by 9,818 accounts that had received legal action for allegedly violating copyright law (Citizens and Technology Lab, 2021; Matias et al., 2020). This study was designed by a group of legal scholars, social scientists, and computer scientists based on prior ethnographic and survey research, without consulting participants about the specific study design. To identify eligible accounts, we scraped public records of legal notices, linked them with specific Twitter account IDs, and queried the public Twitter API to retrieve their public statements. We collected public tweets over the course of the 23 days before and after they received the notice. To support analysis, we then created an anonymized, aggregated record of the number of tweets per day for each account, removing reference to specific days from the final dataframe. The final analysis examined differences in the daily rate of tweets before and after receiving a legal notice.

We developed this observational study because we believe a randomized trial would violate the principles of beneficence and justice. A field experiment that randomly assigned people to different law enforcement conditions could disproportionately expose some people to tens of thousands of dollars in legal penalties. Because we were not assigning people to receive the intervention, we could not consent people before they received a legal notice. Furthermore, we hypothesized that receiving a notice would cause people to participate less on Twitter. If we had sought consent afterward, our requests might only be seen by people who were not deterred by legal action, leading us to mistakenly underestimate the damage of copyright enforcement to people's participation online. This study was reviewed by the MIT IRB, who granted us permission to waive informed consent and required us to debrief participants. We were also required to store all public Twitter data and legal notices in an encrypted datastore.

We used Bartleby to send debriefing invitation messages to a random sample of 4,766 study participants. The message sender script sent tweets from an account that presented itself as a research debriefing account, with a university logo as its profile picture. The tweets @-mentioned the participant and included a link to the Bartleby page for the study. We did not use direct messages because most Twitter accounts are configured to only receive direct messages from accounts that they follow. While these tweets are publicly viewable on the debriefing account, they will not appear in the timelines of recipients' followers unless those followers also follow the debriefing account. The additional privacy risk the debriefing account introduces is small because there are already public records databases of copyright notices. Because our study is relatively low risk, we decided the benefit of sending debriefing invitations was worthwhile for this particular study. Out of the 4,766 accounts we designated for debriefing, numerous accounts were not contactable because they had been suspended, deleted, or had otherwise become unavailable in the time between data collection and debriefing. We successfully sent debriefing invitations to 3,631 accounts. As reported in Table 1, three accounts logged into the debriefing system and one opted out.

During the debriefing process, we became aware of Twitter's "Quality Filter" feature, which filters notifications from "duplicate Tweets or content that appears to be automated" (@EmilLeong, 2016). It is possible that some participants did not receive notifications about the debriefing tweet. However, the filter would not have affected all participants, and the algorithm's exact criteria are opaque. Because this feature would affect all attempts at debriefing on Twitter, our results are still of value for understanding Twitter debriefing and opt out rates.

Debriefing a Field Experiment in a Reddit Community

We also used Bartleby to debrief participants in a field experiment hosted by an online feminism discussion community

Table 2. Accounts Participating in Reddit Debriefing.

Period	Participants	Contacted	Logged in	Login rate	Opted out	Opt-out rate
25 June–31 July 2020	1,342	1,177	10	0.0085	3	0.0025

Table 3. Direct Message Replies to Reddit Debriefing Invitation.

Period	Participants	Contacted	Replies	Unique users	User reply rate
25 June–7 June 2020	1,342	1,177	23	22	0.019

(Citizens and Technology Lab, 2020). In this study, which included feminist and anti-feminist participants, we randomly assigned first-time commenters to a control group or to an intervention group that received a private message. Several weeks later, we both collected data on public participant activity and sought consent for participation in a survey. After an observation period, we created an anonymized, aggregated dataset of participant activity in the community and merged it with survey responses. We then conducted analyses within the full observation sample and within the subset of participants that completed the survey. Finally, we used Bartleby to send participants a link to a debriefing experience, providing them an opportunity to opt out.

In this study, we followed procedures of co-design and participatory hypothesis testing (Matias, 2016; Matias & Mou, 2018), where community representatives were involved in the research process from inception to debriefing. We held a day-long workshop with community representatives to identify the general research area, co-designed the study over several months with the community, and presented the final study design for approval by the community before submitting it to review by the Princeton and Cornell University IRBs.

In the final design, we decided to use a debriefing process for participants who experienced our intervention and whose public activity was observed for the study. We and community representatives decided together on debriefing rather than consent because the intervention was designed to support first-time participants, because the minimal, short-term risks were reversible, and because we could not anticipate who would participate before the study began.

Using Bartleby, we sent direct messages to all 1,342 study participants. These debriefing invitation messages included study information and a link to the Bartleby login page for the study (Figure 1). Out of the 1,342 participant accounts, a small number of accounts had been deleted in the time between data collection and debriefing. We successfully sent debriefing invitations to 1,177 accounts. As reported in Table 2, 10 accounts logged into the debriefing system and 3 opted out.

In addition to receiving feedback from those who logged into the debriefing system, we also received 23 direct messages from 22 participants via replies to the debriefing invitation message (Table 3). Some messages were positive, thanking us for conducting the study. Others included harsh criticism and profanities. We observed that many messages to the debriefing account were directed at the moderators of the Reddit community, whom participants did not distinguish us from as researchers.

Discussion

Procedural and Substantive Theories in Research Ethics

How academics think about research ethics is shaped by the underlying ethical theories they are working with. For example, when US institutions established their approach to research ethics with the Belmont Report, they were responding to human rights violations through the lens of existing theories of research ethics (United States, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). Mid-20th century ethicists described research ethics as a balancing act between the individual autonomy of participants and the common good that scientists were expected to pursue (Rothman, 2017). Consequently, when developing the model of research ethics in the United States, the authors of the Belmont Report cited principles protecting individual rights including respect for persons. They also advanced principles that guided scientists toward the common good, including beneficence and justice (United States, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). By working from principles of individual autonomy and the common good, the authors of the Belmont Report were able to develop guidelines, regulations, and bureaucracies that govern research ethics in the United States to this day.

Feminist philosophers Mackenzie and Stoljar have identified two overarching kinds of ethical theories at play in discussions of autonomy and the common good: procedural theories and substantive theories (Mackenzie & Stoljar, 2000). These theories have been particularly important for feminist ethics, which has sought to reconcile individual autonomy with the collective concerns of structural oppression. Similar distinctions have been made between theories of justice in political philosophy (Rawls, 1971), and between procedural ethics and "ethics in practice" in medicine

(Guillemin & Gillam, 2004). *Procedural theories* focus on the details of procedures—standardized, repeatable steps that can be automated by a software system or a bureaucracy—that protect individual rights, such as a right to individual autonomy. Procedural theories are often called "content-neutral" because these theories do not treat the content of a person's specific actions and decisions as relevant to whether they are moral. Procedural theories focus on how well those actions and decisions happened within the structure of pre-defined ethics procedures that protect autonomy, such as informed consent. *Substantive theories* focus on the idea that procedural theories "must be supplemented by some non-neutral condition" (Mackenzie & Stoljar, 2000). In other words, substantive theories argue that the actions and decisions people make, even within procedures such as informed consent, must account for moral ideas such as the common good—ideas that cannot necessarily be standardized into a procedure. For example, if a participant gives informed consent to a study that is against their own interests or that poses a threat to others in society, substantive theorists would question whether that consent is sufficient moral justification for the study to happen.

Researchers and their institutions apply procedural theories to research ethics when they design and implement ethics procedures. For example, researchers comply with regulatory requirements by submitting their plans to an IRB for review. IRBs will often require researchers to implement specific procedures, such as informed consent, designed to protect participant autonomy and enforce oversight processes. When US researchers "believe that approval by IRBs is sufficient for addressing ethical considerations" (Chiauzzi & Wicks, 2019), those beliefs about research ethics can be described in purely procedural terms. When regulators mandate certain procedures and institutional arrangements, the power of researchers over participants is guided and restrained by researchers' focus on procedural compliance. When new needs arise as research methods change, regulators revise policies governing ethics procedures—such as in 2018 when the US government made changes to the Common Rule (US Department of Health and Human Services, 2017).

Procedural theories can guide valuable progress in research ethics. University IRB staff have struggled with the fact that the "Common Rule . . . does not provide appropriate guidance for the realities of research with online data" (Vitak et al., 2017). Even in the absence of institutional requirements, online researchers and their participants can benefit from improvements in the design and use of ethics procedures for large-scale behavioral research online. For example, many offered procedural criticisms of the Facebook Emotion Contagion study, arguing that it should have included consent or debriefing (Grimmelmann, 2015). By introducing a large-scale debriefing procedure using a system like Bartleby, the researchers would have made the morally significant improvement of offering an opportunity to exercise autonomy where none was previously afforded.

Researchers apply substantive theories to research ethics when they have value-driven conversations about a study's content, design, and other ethically-relevant issues—regardless of any procedures employed. For example, a substantive ethics conversation on the Facebook study would cover topics including mental health risks, collective risks associated with large-scale attempts at social influence, and the nature of people's relationship with Facebook. These substantive conversations often depend on the content and context of a specific study. These conversations are also affected by potentially different views about normative concepts (such as the common good) held by participants and researchers.

Even when researchers and participants agree on the importance of normative concepts in substantive ethics (such as harm, beneficence, and justice) they can still disagree on how to understand those concepts and their relative importance. As prior work has shown, participants are active agents, with their own agendas separate from researchers' plans (Howard & Irani, 2019). As these conflicts are ignored or negotiated, the power structures and power imbalances that enable those moves are also a concern of substantive ethics. Researchers applying substantive theories would ask how participants are able to exercise voice and power (if at all) in normative discussions and decisions about the design, implementation, and uses of research.

While substantive theories help people ask context-specific questions about a study, these theories also enable conversations about power relations between participants and researchers that can apply across multiple studies. Guided by these substantive questions, researchers and communities have worked to redesign how interpersonal and institutional power relations are structured in research (Matias & Mou, 2018). Participatory research, co-design processes (Sasha Costanza-Chock, 2020), refusal (Benjamin, 2016), and empirical work on participant expectations (Desposato, 2018; Fiesler & Proferes, 2018) all provide ways to surface substantive issues in research. Because none of these models can prescribe solutions to normative questions in research, they should not be treated as procedural checkboxes.

Just as researchers work to balance or weave together individual autonomy and the common good, we should also see procedural and substantive theories as complementary. Rather than supplanting each other, these theories provide resources for combining equally-important considerations inherent in research ethics. For example, substantive theorists make an important critique that the content-neutrality of procedural theories provides a necessary but insufficient account of the ethics of a situation. Rather than abolish procedures that respect individual autonomy, researchers should do further work to consider questions of power.

Because designing systems that automate ethics procedures often involves creating structures of power (Winner, 1980), attempts to design new systems for research ethics can benefit from both procedural and substantive ethical

theories. In this article, we use the two kinds of theories to conduct complementary analyses that situate the Bartleby system within a broader design space of research ethics systems that might balance individual and collective concerns differently.

Procedural Ethics Evaluation of Bartleby

When we ask whether introducing debriefing procedures via Bartleby increases the ability of participants to exercise their individual autonomy, we are applying procedural theories to the evaluation of research ethics systems. Bartleby protects autonomy by offering each participant access to the information and interface controls required to make an informed opt out decision. In most non-consented online research, research participants have no opportunity to exercise autonomy over their participation. They are often never even told that researchers collected their data. Because Bartleby makes it easier for researchers to implement and scale debriefing, the system creates new opportunities for participants to exercise autonomy that did not previously exist.

In our field deployments, we found that some participants logged into the system and made an active choice to either remain in the study or opt out. Because procedural theories are content-neutral, the debriefing system is considered successful in procedural terms regardless of what their choice was. It does not matter whether they opted out or remained in the study; the system is successful because they made an informed choice that otherwise would not have been possible.

We argue that Bartleby protects participant autonomy even if no people use the system. To illustrate this with an example, consider the case of a facial recognition dataset where people can have their images removed on an opt out basis. Imagine that everyone is offered a reasonable opportunity to opt out, but nobody chooses to exercise that ability. Now imagine a different project where nobody is ever offered the ability to opt out. The outcomes are the same (no one opts out), but the protections to individual autonomy in these two situations are very different. The fact that Bartleby offers participants a choice that would not otherwise be available is morally significant to the ethics of the study.

Opt out procedures are successful if the people who would have opted out, when given the best opportunity to do so, actually do opt out. Because opt out rates must be interpreted in the context of both risks and benefits to participants, and the overall size of the study, there is no normatively desirable opt out rate. For a given study, low participation in debriefing could accurately reflect participant preferences (especially for studies with low risk). We believe this is the case for our studies, which collected low-risk data and are motivated by reasonable common good arguments. The opt out rates observed in our studies may also not be considered low for a study with more total participants, because the absolute number of people who opt out would be much higher.

Bartleby can also be used with studies that pose greater risk to participants. In these cases, we might hypothesize that opt out rates will be higher.

Ethics procedures may introduce barriers to autonomy if a system does not do enough to include people in the process. The counterfactual of whether or not people would have opted out under other circumstances is difficult to measure, because it is difficult to study non-participants. For example, people who declined to use Bartleby would likely not respond to a survey about why they declined. In the case of non-consented studies that would use Bartleby, the alternative to debriefing is not informed consent but rather an absence of procedures that protect individual autonomy.

Although our university ethics board received no complaints about either study during our debriefing tests, Bartleby would also be considered successful if people had reported us for researcher misconduct. Guided by procedural theories, we consider a procedure successful when accountability mechanisms are available and activated when needed. If people had complained about the study after hearing about it through Bartleby, or if there had been a public scandal because people found out about the study, or if the study had been shut down, or if we had lost our research positions—all of these would have been procedural successes.

In addition to improving the implementation and adoption of ethics procedures, the Bartleby system could also support empirical research on the design of ethics interfaces. Researchers using Bartleby could, for example, conduct field experiments that vary features of the debriefing interface to test their effectiveness and usability. Researchers might also interview participants to learn about the relationship between the content of research and the effectiveness of ethics procedures—for example, whether more people opt out of studies perceived to be higher risk. As consumer privacy regulations like the GDPR continue to prioritize consent procedures, research from Bartleby and related systems could influence the design of ethics procedures beyond academic research.

Substantive Ethics Evaluation of Bartleby

Although the two studies in our field deployment used debriefing procedures in the same way, they were different in significant substantive ways. For instance, the participants' prior relationships to each other and to the researchers differed between the two studies. A procedural account of these two studies would describe them very similarly—through a procedural lens, the studies used the exact same debriefing process. The lens of substantive ethics reveals differences in the exercise of power between participants and researchers that may lead to differences in normative judgments on the ethics of those studies.

For researchers thinking in terms of substantive theories, the normative content of the research is important to ethics—not simply the content-neutral procedures that were implemented. If researchers and participants both agree that

benefits of the study outweigh the risks of using the data, the research might proceed uncontested. However, interpretations of risk and benefit are subjective. When researchers make decisions without involving participants, they wield a large amount of power over participants through this exclusion. In our analysis, we narrate the different substantive ethics concerns surfaced from the two studies where we tested Bartleby.

In the Twitter study, which included individuals who have experienced copyright takedowns, we did not have access to organizations or entities that could speak and act on behalf of participants. Consequently, we were unable to include participants in the design of the study or account for their voices and perspectives beyond evidence from exploratory qualitative research with people similar to those in our study. As a result, the Bartleby system offered the study's only opportunity for input or power from research participants.

Since few people in the study responded to debriefing, we have relied on our own intuitions and the oversight of our IRB boards to address substantive questions. The Twitter study was observational, collected minimal data, and used appropriate anonymization and data storage practices. Harms from unintended disclosure of our research dataset would also be minimal—the count of tweets someone made in a time period is not sensitive in the same way that, for example, medical information is sensitive. As a result, we can reasonably argue that the study is minimal risk. The study also has a reasonable common good argument, as knowledge about the effects of automated legal notices on the exercise of speech rights could inform future policy discussions. Because the study is relatively innocuous, we might be less concerned that Bartleby had low usage.

In contrast, the Reddit study involved a community in conflict—a large feminism discussion community where anyone can join discussions and where anti-feminists attempt to disrupt conversations. People who participate in this community accept governance by community, so we worked with moderators to co-design research questions and study procedures. Because moderators understand the norms and preferences of the community, we obtained what Humphreys calls *proxy consent* (Humphreys, 2015), asking moderators to grant consent on behalf of the community they represent. Since community representatives reviewed, influenced, and approved the study design, they contributed to decisions about substantive ethics concerns. Independently of the number of people who used Bartleby, our community representatives' knowledge of norms and values gave us more confidence that our research was respecting and managing risk for all participants, including those who did not debrief.

Working with community members is a helpful way to surface substantive issues in research, but participatory research should not be understood as just another one-size-fits-all procedure. Community-based proxy consent, for example, is not sufficient for all studies because communities are usually heterogeneous. It is not always clear how researchers should position themselves when different groups within the community disagree (Keegan & Matias, 2016).

In the Reddit study, debriefing enabled us to maintain a respect for individual autonomy alongside our community engagement. The study included people who self-identify as feminists, people who do not, and people who identify as anti-feminists. Among those who identified as feminists, some did not identify closely with the community, arguing that the community's moderators align with ideological positions that they do not agree with. Our co-design process was also unable to include the views of anti-feminists whose purpose was to disrupt the community and cause them harm. As researchers, we held the normative position that harassment and disruption were not legitimate goals to uphold in our research. Consequently, when deciding to conduct research focused on protecting community from harassment, we committed to a power structure that aligned us with the community's moderators. But we also wished to respect individual autonomy.

Within this complex situation, the Bartleby system provided opportunities for voice and agency among individuals who disagreed with our ethical judgments and those of the community. When debriefing participants, we learned that some people disagreed that the moderators' use of power was legitimate. Several participants who had been banned by moderators sent private messages to the Bartleby system complaining about their treatment by the community. When they saw the study and were included in its procedures, they were reminded of community moderators' continued power over them. They then used the debriefing system to object not to our study, but to that deeper structure of power. Procedurally, those messages were irrelevant. Substantively, they were essential.

These messages also suggest potential risks to researchers arising from increased visibility due to the use of systems like Bartleby. Although US research ethics regulation was written to protect participants from abuse of power by researchers, internet scholars have written about the risk of online communities conducting organized harassment and abuse against researchers. Writing about the challenges of researching far-right online spaces, Massanari notes that "[power] asymmetry is, in part, due to the visibility of those being targeted and the relative invisibility of those who are perpetrating the attacks" (Massanari, 2018). Indeed, researchers have chosen not to debrief research subjects for reasons ranging from "a demonstrated propensity for online harassment" (Munger, 2017) to "[an indicated] strong desire to be left alone" (Hudson & Bruckman, 2004). In these cases, debriefing could cause more harm than benefit. Harms to researchers from online harassment disproportionately impact marginalized scholars based on factors including gender, race, chosen research topic, and career expectations of online visibility (Gosse et al., 2021; Massanari, 2018; Stein & Appel, 2021). The decision to use Bartleby must

account for these power dynamics, especially when institutions are still learning how to protect researchers from online harassment. When we decided to study and debrief the Reddit community, we followed guidance from a Data & Society report on "Best Practices for Conducting Risky Research" (Marwick et al., 2016). For example, team members had discussions about possible risks and followed cybersecurity guides to remove personal information from the internet to reduce risk of doxxing. These examples of situations where debriefing would be inappropriate demonstrate that even well-designed procedures require substantive conversations about when they should be used.

As we found, debriefing systems can also contribute to conversations about substantive ethics even if they seem like mere box-checking exercises. As we saw in the Reddit study, the Bartleby system protected individual autonomy while surfacing substantive issues by enabling participants to voice concerns. We found that debriefing can make researchers aware of contrasting values held by different participants, informing how researchers think and act on our normative values and uses of power.

Passive Non-Participation, Sovereignty, and Non-Alienation

In our deployment of Bartleby, we observed aspects of online, non-consented research that drew our attention to two risks to autonomy: default inclusion and passive non-participation.

In many offline studies in controlled settings, people are not enrolled into studies by default without their consent. If people ignore requests for informed consent, they will not be included in research without their active involvement. However, in online non-consented research, people are *included by default* in research as they participate in online public spheres. People must actively opt out in order not to participate in research.

Non-consented research and debriefing procedures introduces the possibility of *passive non-participation* (Casemajor et al., 2015). People who log into the debriefing system will either decide to opt in as active participants or opt out to become active non-participants. In contrast, passive non-participants do not respond to debriefing and do not actively reason about participation. When people are included in research but are either unaware of their involvement or disinterested in managing a relationship they did not initiate, they do not engage in research ethics on the terms laid out by researchers employing procedures or automated systems.

In our field deployments of Bartleby, we observed that most people were passive non-participants—that is, they did not log into the system to complete debriefing. Distinguishing between opt out and passive non-participation is important because critics of Bartleby may question whether the system truly protects the autonomy of passive non-participants. Opt-out is morally significant because it involves an act of communication, which makes the result of their informed autonomous decision known to us. Passive non-participants do not communicate with us, so we have no way of knowing whether they made an informed decision. We cannot distinguish participants who saw the debriefing invitation and chose to ignore it from participants who never saw the debriefing invitation at all. Access to the ability to opt out via the debriefing invitation is what protects autonomy. So if participants never saw the invitation, we might say that morally it is similar to if the invitation was never made.

How should we think about how to protect the autonomy of passive non-participants? Enoch, an ethicist, argues that concerns about autonomy can actually reflect two distinct concerns: a concern for sovereignty, and a concern for non-alienation (Enoch, 2017). *Sovereignty* concerns are about individuals having control over choices that affect them. When theorists and activists apply the standard of affirmative consent (Im et al., 2021) to issues such as sexual politics, they are responding to the concern for sovereignty. Once someone has asserted their sovereignty and communicated their decision about a request for consent, that decision is final. *Non-alienation* concerns are instead articulated in terms of a person's deep commitments. Medical caregivers must navigate the concern for non-alienation when, for example, treating someone who is unconscious and cannot affirmatively consent. Imagine the unconscious person has religious commitments that disallow certain medical interventions. Going against those commitments, even if it might save their life, would violate their autonomy. This is because people's deep commitments are intimately tied to their sense of self.

A theory of distinct autonomy concerns helps us untangle our worries about passive non-participation in debriefing when using Bartleby. According to Enoch, when we talk about autonomy we are sometimes concerned with sovereignty, sometimes with non-alienation, and sometimes with both (Enoch, 2017). If we are primarily concerned with sovereignty, debriefing procedures are not enough to address this concern when passive non-participation is possible. Because people cannot be forced to participate in procedures, we do not know what sovereign decision they would have made about their own participation.

If we are primarily concerned with non-alienation, passive non-participation might be less of a problem depending on the normative content of the research. For example, it is unlikely that counting how many tweets someone made during a time period or advancing public understanding of the effect of DMCA takedowns on free speech is against anyone's deep commitments. It may be against someone's preferences, but likely does not threaten their sense of self. Making feminism discussion communities more welcoming could be against anti-feminists' deep commitments. However, as people who participate in that community, anti-feminist commenters have made a sovereign decision to subject themselves to the governance structure and community expectations of that forum.

Because Bartleby is a system designed to protect participant autonomy in online non-consented research, its design must navigate the issues that arise in this category of research. For situations where we are mostly concerned with sovereignty, Bartleby can only protect the autonomy of active participants (whether or not they opt out). For situations where we are mostly concerned with non-alienation, the autonomy of passive non-participants is sensitive to substantive issues in the normative content of the research. In most situations, researchers will want to protect autonomy out of concern for both sovereignty and non-alienation. For the reasons we have discussed so far, using Bartleby is an improvement over not using it in such cases.

Debriefing and Spam Filters

Ethics procedures might constitute a form of spam. In our Twitter field deployment, the platform's algorithmic "Quality Filter" potentially affected whether subjects received our debriefing invitations in a way that we are currently unable to quantify. The feature is designed to filter out notifications from automated spam accounts. Twitter does not distinguish between debriefing messages and other automated communications for purposes such as marketing or disinformation, a normative stance motivated by values of authenticity in content moderation. From the platform's perspective, these messages are "high-volume unsolicited . . . mentions," which constitute platform manipulation (Platform manipulation and spam policy, 2019).

Anti-spam efforts likely complicate the problems of passive non-participation. Spam filters create situations where someone could potentially have exercised their right to opt out, but was prevented from doing so by an algorithm that the neither researchers nor participants control. This problem is widely applicable beyond social media and includes any digitally mediated communication—such as email-based recruitment, where algorithms can down-weight ethics procedures in the inbox or relegate them to a spam folder.

In contrast to those for whom the platform filters out debriefing messages, other passive non-participants see the debriefing invitation but choose not to respond. For these subjects, debriefing is an unwanted burden on their time and attention. If more researchers adopt a norm of debriefing, the volume of requests could grow substantially. Scholars have noted that the more entities collect and use personal data, the less feasible it is for individuals to manage their privacy separately with each entity (Solove, 2012). This creates a difficult trade-off for researchers, who want to reach people who will be helped by debriefing (especially those who would be helped but are prevented from engaging due to anti-spam algorithms) but do not want to lose trust by gaining reputations as spammers.

These issues suggest a need for researchers to situate thinking about autonomy in a broader discussion about socio-technical systems. For researchers thinking about debriefing procedures, spam filters highlight the fact that participants' individual choices—usually thought of as a product of individual autonomy—are a product of the interaction between individual autonomy, platform algorithms, and possible value-driven differences between what researchers and participants think of as spam. The burden created by a large volume of individual debriefing requests also highlights the need for more research into collective approaches for managing autonomy.

Conclusion

With Bartleby, we contribute an open-source research ethics system that provides an interface for delivering debriefing procedures alongside large-scale online research. Researchers who use Bartleby in non-consented research will offer participants more opportunities to exercise autonomy than would otherwise be available. In evaluating this contribution, we underscored the importance of bringing multiple complimentary theories to bear when interrogating the both the promise and limitations of research ethics system design. At a time when few researchers invite any kind of public voice into the research process, we believe that similar creative conversations among design, empirical research methods, and feminist philosophy can advance research ethics and increase public trust in research.

Acknowledgements

The authors thank Haley Schilling, Nicholas Proferes, David Karger, Jane Im; members of the Citizens and Technology Lab, particularly Eric Pennington, Julia Kamin, and Max Klein; the MIT Visualization Group; and the anonymous reviewers.

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

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the NSF Graduate Research Fellowship and the Paul and Daisy Soros Fellowship for New Americans, and by open access funding from MIT Libraries.

ORCID iD

Jonathan Zong https://orcid.org/0000-0003-4811-4624

References

Akkad, A., Jackson, C., Kenyon, S., Dixon-Woods, M., Taub, N., & Habiba, M. (2006). Patients' perceptions of written consent: Questionnaire study. *BMJ*, *333*(7567), Article 528.

Arora, A., Rajagopalan, S., Shafiq, N., Pandhi, P., Bhalla, A., Dhibar, D. P., & Malhotra, S. (2011). Development of tool for the assessment of comprehension of informed consent form in healthy volunteers participating in first-in-human studies. *Contemporary Clinical Trials*, *32*(6), 814–817.

Athey, S., Catalini, C., & Tucker, C. (2017, June). *The digital privacy paradox: Small money, small costs, small talk* (Working paper no. 23488). National Bureau of Economic Research. http://www.nber.org/papers/w23488

Barchard, K. A., & Williams, J. (2008). Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods*, *40*(4), 1111–1128. http://link.springer.com/10.3758/BRM.40.4.1111

Benjamin, R. (2016, June). Informed refusal: Toward a justice-based bioethics. *Science, Technology, & Human Values 41*: 967–990. https://journals.sagepub.com/doi/10.1177/0162243916656059

Bhansali, S., Shafiq, N., Malhotra, S., Pandhi, P., Singh, I., Venkateshan, S. P., Siddhu, S., Sharma, Y., & Talwar, K. K. (2009). Evaluation of the ability of clinical research participants to comprehend informed consent form. *Contemporary Clinical Trials*, *30*(5), 427–430.

Bravo-Lillo, C., Egelman, S., Herley, C., Schechter, S., & Tsai, J. (2013, May). *You needn't build that: Reusable ethics-compliance infrastructure for human subjects research*. CREDS 2013(Workshop). https://www.microsoft.com/en-us/research/publication/you-neednt-build-that-reusable-ethics-compliance-infrastructure-for-human-subjects-research/

Bronx Community Research Review Board What is CERA? (2016). http://bxcrrb.org/cera-project/what-is-cera/

California Consumer Privacy Act (CCPA). (2018, October). https://oag.ca.gov/privacy/ccpa

Casemajor, N., Couture, S., Delfin, M., Goerzen, M., & Delfanti, A. (2015, September). Non-participation in digital media: Toward a framework of mediated political action. *Media, Culture & Society*, *37*(6), 850–866. https://doi.org/10.1177/0163443715584098

Chen, C., Lee, P.-I., Pain, K. J., Delgado, D., Cole, C. L., & Campion, T. R. (2020). Replacing paper informed consent with electronic informed consent for research in academic medical centers: A scoping review. *AMIA Summits on Translational Science Proceedings*, 2020, 80–88. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233043/

Chiauzzi, E., & Wicks, P. (2019). Digital trespass: Ethical and terms-of-use violations by researchers accessing data from an online patient community. *Journal of Medical Internet Research*, 21(2), Article e11985. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6403524/

The CITI Program. (2018). *Social-behavioral-educational (SBE) comprehensive*. https://about.citiprogram.org/en/course/human-subjects-research-2/

Citizens and Technology Lab. (2020, June). *Freedom of expression in marginalized groups online*. https://citizensandtech.org/research/expression-in-marginalized-groups/

Citizens and Technology Lab. (2021, May). *Automated law enforcement*. https://citizensandtech.org/research/automated-law-enforcement/

Coiera, E., & Clarke, R. (2004). e-Consent: The design and implementation of consumer consent mechanisms in an electronic environment. *Journal of the American Medical Informatics Association*, *11*(2), 129–140.

Community IRBs & Research Review Boards: Shaping the Future of Community-Engaged Research [Tech. Rep.]. (2012). Albert Einstein College of Medicine, The Bronx Health Link and Community-Campus Partnerships for Health. https://ccphealth.org/community-irbs-research-review-boards-shaping-the-future-of-community-engaged-research-2/

Cornwall, A., & Jewkes, R. (1995). What is participatory research? *Social Science & Medicine*, *41*(12), 1667–1676. https://doi.org/10.1016/0277-9536(95)00127-s

Desposato, S. (2018). Subjects and scholars' views on the ethics of political science field experiments. *Perspectives on Politics*, *16*(3), 739–750. https://www.cambridge.org/core/journals/perspectives-on-politics/article/subjects-and-scholars-views-on-the-ethics-of-political-science-field-experiments/1A152D85B81C5F9D7FE82CA37494FEDC

Dresden, G. M., & Levitt, M. A. (2001). Modifying a standard industry clinical trial consent form improves patient information retention as part of the informed consent process. *Academic Emergency Medicine*, *8*(3), 246–252.

@EmilLeong. (2016, August). *New ways to control your experience on Twitter*. https://blog.twitter.com/en_us/a/2016/new-ways-to-control-your-experience-on-twitter.html

Enoch, D. (2017, October). Hypothetical consent and the value(s) of autonomy. *Ethics*, *128*(1), 6–36. https://www.journals.uchicago.edu/doi/10.1086/692939

Epstein, L. C., & Lasagna, L. (1969). Obtaining informed consent: Form or substance. *Archives of Internal Medicine*, *123*(6), 682–688.

Fiesler, C., & Proferes, N. (2018, January). "Participant" perceptions of Twitter research ethics. *Social Media + Society*, *4*(1), 1–14. https://doi.org/10.1177/2056305118763366

Geier, E. (2019, May). *5 important considerations for evaluating IRB systems*. https://www.badrabbit.com/news/2019/5/2/evaluating-irb-systems

General Data Protection Regulation. (2018). https://gdpr-info.eu/

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 1–5. http://journals.sagepub.com/doi/10.1177/2053951720943234

Gosse, C., Veletsianos, G., Hodson, J., Houlden, S., Dousay, T. A., Lowenthal, P. R., & Hall, N. (2021). The hidden costs of connectivity: Nature and effects of scholars' online harassment. *Learning, Media and Technology*, *46*(3), 264–280. https://doi.org/10.1080/17439884.2021.1878218

Grimmelmann, J. (2015). The law and ethics of experiments on social media users. *Colorado Technology Law Journal, 13*, Article 219. https://osf.io/cdt7y

Guillemin, M., & Gillam, L. (2004). Ethics, reflexivity, and "Ethically Important Moments" in research. *Qualitative Inquiry*, *10*(2), 261–280. https://doi.org/10.1177/1077800403262360

Hallinan, B., Brubaker, J. R., & Fiesler, C. (2020). Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media & Society*, *22*(6), 1076–1094. https://doi.org/10.1177/1461444819876944

Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior*, *18*(1), 59–92. https://doi.org/10.1080/10508420701712990

Hochhauser, M. (2000). The informed consent form: Document development and evaluation. *Drug Information Journal*, *34*(4), 1309–1317.

Howard, D., & Irani, L. (2019). Ways of knowing when research subjects care. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. http://doi.org/10.1145/3290605.3300327

Hudson, J. M., & Bruckman, A. (2004). "Go Away": Participant objections to being studied and the ethics of chatroom research.

The Information Society, *20*(2), 127–139. http://www.tandfonline.com/doi/abs/10.1080/01972240490423030

Humphreys, M. (2015). Reflections on the ethics of social experimentation. *Journal of Globalization and Development*, *6*(1), 87–112. https://www.degruyter.com/view/journals/jgd/6/1/article-p87.xml

Hyatt, J. (2005). The birth of the ethics industry. *Business Ethics: The Magazine of Corporate Responsibility*, *19*(2), 20–27. https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=bemag&id=bemag20050019000200200027

Im, J., Dimond, J., Berton, M., Lee, U., Mustelier, K., Ackerman, M. S., & Gilbert, E. (2021, May). Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). Association for Computing Machinery. http://doi.org/10.1145/3411764.3445778

Jacob, M.-A. (2007). Form-made persons: Consent forms as consent's blind spot. *Polar: Political and Legal Anthropology Review*, *30*(2), 249–268.

Keegan, B., & Matias, J. N. (2016, March). *Actually, it's about ethics in computational social science: A multiparty risk-benefit framework for online community research (Conference session)*. AAAI Spring Symp. on Observational Studies through Social Media and Other Human- Generated Content, Stanford, CA, United States.

Kim, H., Bell, E., Kim, J., Sitapati, A., Ramsdell, J., Farcas, C., & Ohno-Machado, L. (2017). iCONCUR: Informed consent for clinical data and bio-sample use for research. *Journal of the American Medical Informatics Association*, *24*(2), 380–387.

Koenig, B. A. (2014). Have we asked too much of consent? *The Hastings Center Report*, 44(4), 33–34. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4249719/

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, USA*, *111*(24), 8788–8790. https://www.pnas.org/content/111/24/8788

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of board of scientific affairs' advisory group on the conduct of research on the Internet. *American Psychologist*, *59*(2), 105–117. https://doi.org/10.1037/0003-066X.59.2.105

Liat Racin. (2016, February). *Research in the community lightning talk: The community IRB Project, Liat Racin*. The Engagement Lab, Emerson College. https://vimeo.com/154358905

Mackenzie, C., & Stoljar, N. (2000). Introduction: Autonomy refigured. In C. Mackenzie & N. Stoljar (eds) *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self* (pp. 3–31). Oxford University Press.

Marwick, A. E., Blackwell, L., & Lo, K. (2016, October). *Best practices for conducting risky research and protecting yourself from online harassment* [Technical report]. Data & Society.

Massanari, A. L. (2018). Rethinking research ethics, power, and the risk of visibility in the era of the "Alt-Right" gaze. *Social Media + Society*, *4*(2), 1–9. https://doi.org/10.1177/2056305118768302

Matias, J. N. (2016). Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1138–1151). ACM Press.

Matias, J. N., & Mou, M. (2018). CivilServant: Community- led experiments in platform governance. In *Proceedings of the 20 CHI Conference on Human Factors in Computing Systems—CHI' 18* (pp. 1–13). ACM Press. http://dl.acm.org/citation.cfm?doid=3173574.3173583

Matias, J. N., Mou, M. E., Penney, J., & Klein, M. (2020, September). Do automated legal threats reduce freedom of expression online? Preliminary results from a natural experiment. https://osf.io/nc7e2/

Melville, H. (1853, July). Bartleby, the Scrivener: A story of wall street. *Putnam's Monthly. A Magazine of Literature, Science, and Art*. https://www.kelmscottbookshop.com/pages/books/35182/herman-melville/putnams-monthly-magazine-of-american-literature-science-and-art-volume-ii-bartleby-the-scrivener

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649. http://link.springer.com/10.1007/s11109-016-9373-5

Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020, April). Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. http://doi.org/10.1145/3313831.3376321

Olivia Solon. (2019, March). Facial recognition's 'dirty little secret': Social media photos used without consent. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

Platform manipulation and spam policy. (2019, September). https://help.twitter.com/en/rules-and-policies/platform-manipulation

Puneet Chawla Sahota. (2009). *Research regulation in American Indian/Alaska Native communities: Policy and practice considerations*. https://www.ncai.org/policy-research-center/initiatives/Research_Regulation_in_AI_AN_Communities_-_Policy_and_Practice.pdf

Rawls, J. (1971). *A theory of justice*. Harvard University Press. http://www.jstor.org/stable/j.ctvkjb25m

Rothman, D. J. (2017). *Strangers at the bedside: A history of how law and bioethics transformed medical decision making*. Routledge.

Sasha Costanza-Chock. (2020). *Design justice: Community-Led practices to build the worlds we need*. MIT Press. https://mitpress.mit.edu/books/design-justice

Shilton, K. (2018). Values and ethics in human-computer interaction. *Foundations and Trends® in Human–computer Interaction*, *12*(2), 107–171. https://www.nowpublishers.com/article/Details/HCI-073

Solove, D. J. (2012). Privacy self-management and the consent dilemma. *Harvard Law Review*, *126*, Article 1880.

Solove, D. J. (2020, February). *The myth of the privacy paradox* (SSRN Scholarly Paper No. ID 3536265). Social Science Research Network. https://papers.ssrn.com/abstract=3536265

Sommers, R. (2020, August). *Commonsense consent* (SSRN Scholarly Paper No. ID 2761801). Social Science Research Network. https://papers.ssrn.com/abstract=2761801

Stein, J.-P., & Appel, M. (2021). How to deal with researcher harassment in the social sciences. *Nature Human Behaviour*, *5*(2), 178–180. http://www.nature.com/articles/s41562-020-01011-6

Stoller, D. (2020, January). IBM hit with lawsuit claiming image use for facial recognition. *Bloomberg Law*. https://news.

bloomberglaw.com/privacy-and-data-security/ibm-hit-with-lawsuit-claiming-image-use-for-facial-recognition

Tait, A. R., Voepel-Lewis, T., Malviya, S., & Philipson, S. J. (2005). Improving the readability and processability of a pediatric informed consent document: Effects on parents' understanding. *Archives of Pediatrics & Adolescent Medicine*, *159*(4), 347–352.

Terblanche, M., & Burgess, L. (2010). Examining the readability of patient-informed consent forms. *Open Access Journal of Clinical Trials*, *2*, 157–162.

United States, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

US Department of Health and Human Services. (2017, January). *Revised common rule*. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html

Vitak, J., Proferes, N., Shilton, K., & Ashktorab, Z. (2017). Ethics regulation in social computing research: Examining the role of institutional review boards—Jessica Vitak, Nicholas Proferes, Katie Shilton, Zahra Ashktorab, 2017. *Journal of Empirical Research on Human Research Ethics 12*, 372–382. http://journals.sagepub.com/doi/10.1177/1556264617725200

Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 941–953). ACM Press.

Wilbanks, J. T. (2018). Design issues in e-consent. *The Journal of Law, Medicine & Ethics*, *46*(1), 110–118.

Wilbanks, J. T. (2020). Electronic informed consent in mobile applications research. *The Journal of Law, Medicine & Ethics*, *48*(1 suppl), 147–153.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*, 121–136.

Wright, D. (2012). Redesigning informed consent tools for specific research. *Technical Communication Quarterly*, *21*(2), 145–167.

Yeshimabeit Milner. (2019, July). *Abolish big data*. https://medium.com/@YESHICAN/abolish-big-data-ad0871579a41

Zhang, J. J. (2017). Research ethics and ethical research: Some observations from the Global South. *Journal of Geography in Higher Education*, *41*(1), 147–154. https://doi.org/10.1080/03098265

Author Biographies

Jonathan Zong (SM, Massachusetts Institute of Technology) is a doctoral candidate in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. His research interests include human–computer interaction, research ethics, and data visualization.

J. Nathan Matias (PhD, Massachusetts Institute of Technology) is an Assistant Professor of Communication at Cornell University. His research interests include digital governance and behavior change in groups and networks shaped by algorithms.