Article

Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics

Nicholas Proferes[1] iD **, Naiyan Jones[2]** iD **, Sarah Gilbert[3]** iD **,
Casey Fiesler[4], and Michael Zimmer[5]** iD

Abstract
This article offers a systematic analysis of 727 manuscripts that used Reddit as a data source, published between 2010 and 2020. Our analysis reveals the increasing growth in use of Reddit as a data source, the range of disciplines this research is occurring in, how researchers are getting access to Reddit data, the characteristics of the datasets researchers are using, the subreddits and topics being studied, the kinds of analysis and methods researchers are engaging in, and the emerging ethical questions of research in this space. We discuss how researchers need to consider the impact of Reddit's algorithms, affordances, and generalizability of the scientific knowledge produced using Reddit data, as well as the potential ethical dimensions of research that draws data from subreddits with potentially sensitive populations.

Introduction

Reddit has become one of the most prominent social platforms on the web with 52 million daily active users (Reddit. com, 2020a) and over 138,000 active topical communities called "subreddits" (Marotti, 2018). Reddit has also been home to a number of prominent and controversial events, such as attempts to identify the Boston-city bombing terrorists (Starbird et al., 2014); a massive leak of hacked celebrity photos (Marwick, 2017); the coordinated attempt to take on short-sellers of the GameStop stock (Roose, 2021); as well as sometimes racist (Mittos et al., 2020), sexist (Farrell et al., 2019), and vitriolic political discourse (Mills, 2018). In part because of its prominence, influence, and history of controversy, it has also become a data source for researchers.

Tufekci (2014) once called Twitter the "model organism" for academic study because tweets are considered to be "public," because Twitter has open APIs which fosters easy data collection, and because Twitter users often respond to world events as they unfold, making it a useful location to gather observational data. Reddit has started fulfilling many of these same criteria, while offering additional advantages for researchers. For example, Reddit's subreddit structure means that finding relevant research data can be easier than on Twitter, and in contrast to the character limits of Twitter, Reddit offers researchers a qualitative and quantitatively more expansive dataset.

However, working with Reddit data may also present complications. Because of the myriad of media forms on Reddit,

researchers may find that they need multiple methodological approaches in their analysis. Subreddits have their own individual norms and cultures, as well as moderation practices, meaning insights from social phenomena in one subreddit may not translate across contexts. The site also offers a large degree of anonymity and one-time use accounts are not uncommon. Because users may feel as though they can speak freely on Reddit as a result of fairly permissive content policies and the anonymity afforded, researchers may be collecting sensitive discussions.

These properties raise questions about how researchers engage in scientific practice when it comes to using data from Reddit. To date, there is no systematic work detailing how researchers are studying Reddit, the phenomena they are studying and approaches they are using, what aspects of Reddit are being studied, and how researchers are engaging the potentially thorny ethical questions of research in this space. Modeled after Zimmer and Proferes's (2014) systematic topology of research on Twitter, this study

[1]Arizona State University, USA
[2]UK Office for National Statistics (ONS), UK
[3]University of Maryland, USA
[4]University of Colorado Boulder, USA
[5]Marquette University, USA

Corresponding Author:
Nicholas Proferes, Arizona State University - West campus 4701 W. Thunderbird Road Glendale, AZ 85306, USA.
Email: nprofere@asu.edu

presents a systematic overview of 727 research studies that used Reddit data published between 2010 and May of 2020. The analysis offers insights into the growth in the use of Reddit as a data source, the range of disciplines in which this research is occurring, how researchers are accessing Reddit data, characteristics of the datasets researchers are using, the subreddits and topics frequently studied, the kinds of analysis and methods researchers are engaging in, and emerging ethical questions of research in this space.

Review of Relevant Literature

Public Data Use

Researchers have used social media data for a wide range of purposes—from predicting postpartum depression from Facebook posts (De Choudhury et al., 2014) to trying to predict movements in the stock market based on the sentiment of Tweets (Mittal & Goel, 2010). However, not all uses of social media data have been welcomed by users or seen as acceptable in the research community. For example, personally identifiable information from more than 87 million Facebook users was collected in an academic study, but then the data were used by Cambridge Analytica to micro-target political advertisements (Isaak & Hanna, 2018). Transgender YouTubers have had their images collected without consent to train facial recognition software and as a part of automatic-gender recognition research and development (Vincent, 2017). And a group of Danish researchers were criticized after they publicly released a data set of nearly 70,000 users of the online dating site OkCupid, which included "usernames, age, gender, location, what kind of relationship (or sex) they are interested in, personality traits, and answers to thousands of personal profiling questions used by the site" (Zimmer, 2018).

Within research communities and among Institutional Review Boards (IRBs), there is disagreement about the ethical practices that should follow from the use of public data for research purposes and if, or when, using social media constitutes human subject research (Metcalf & Crawford, 2016; Vitak et al., 2016). In the United States, institutions that house research with human subjects and who receive federal funds are required to have an IRB. However, the use of publicly available data from social media platforms often does not meet the threshold criteria of "research involving human subjects" according to many IRBs. Thus, some IRBs may exempt these kinds of studies from ongoing compliance review and informed consent practices, though others may not (Vitak et al., 2017).

Ongoing questions around using public social media data have led researchers to question how users' feel about their data being used for research. Fiesler and Proferes (2018) conducted a survey of Twitter users to assess how they felt about their data being used. Their findings showed that users were largely unaware that their data were used for research purposes and that perceptions of data use varied by contextual factors, such as who the researchers are and the topic of study, a finding echoing that of Beninger (2017).

This prior work reveals inconsistencies in the way "human subjects research" is defined and applied by ethics bodies and researchers, as well as potential discomfort among many social media users in being research subjects. Increasingly, researchers are using Reddit data as a source; however, there are no systematic reviews of the contexts on Reddit that researchers are studying, nor the ethics practices they are engaging in relation to their work.

Reddit

Discussions on Reddit are primarily public in that anyone, with or without a Reddit account, can view content (with the exception of private subreddits). Both original shared content and discussion comments are "voted" on by users, which determine their visibility. To become a Reddit user, all users need is to select a unique username and a password—email verification is not required. The terms of service dictate users must be at least 13 years of age to sign up. Site-wide norms discourage participation with one's real name as a privacy-protecting measure. Participation history on the site is also public, meaning that anyone can see all of a user's public comments and posts by clicking on their username. The ease with which users can create accounts means that it is possible, and not uncommon, for one person to have multiple accounts. "Throwaway" accounts, or single-purpose accounts created for limited time use, are commonly used when users do not want a post or comment associated with their main or primary account, such as sharing sensitive or personal information (Ammari et al., 2019; Leavitt, 2015). Because participation on Reddit is pseudonymous, demographic information is somewhat difficult to obtain. According to Reddit's site administrators (Reddit.com, 2021) a majority (58%) of users are between 18 and 34 years old and are male (57%).

Subreddits are both user-created and user-moderated. While there are a few overarching Reddit rules about content, subreddits vary considerably regarding what they allow, and in their specific cultures and norms (Chandrasekharan et al., 2018; Fiesler et al., 2018). As part of their subreddit specific-rules, some subreddits carry warnings to researchers about data collection in the communities. For example, r/depression and r/SuicideWatch state all research-related posts and surveys must be approved by the moderator team, and r/IndianCountry prohibits unauthorized research and requests that anyone interested in using the subreddit for research purposes must complete a form for review by moderators.

In addition to individual subreddit rules, Reddit also has a site-wide user agreement. Reddit.com (2020b) user agreement includes the following prohibition related to collecting data:

> Access, search, or collect data from the Services by any means [automated or otherwise] except as permitted in these Terms or in a separate agreement with Reddit. We conditionally grant permission to crawl the Services in accordance with the parameters set forth in our robots.txt file, but scraping the Services without Reddit's prior consent is prohibited.

These terms are fairly standard in their ambiguity (Fiesler et al., 2020), but do suggest that data collected outside the confines of specific allowances—for example, using their API—*may* be a violation of this user agreement. However, Reddit's API is freely available and can be used to access content on the site.

Reddit posts, comments, and metadata can be accessed via the site itself, or via its APIs. Reddit's official API is free and publicly available and provides an array of functions. For these reasons, Reddit has an ecosystem of bots created by its user base to help in several ways, such as content moderation (Jhaver et al., 2019), adding functionality through summarizing information and linking to other websites, or providing humor through parody bot accounts (Long et al., 2017). There are additional ways of accessing Reddit data outside of means provided directly by the platform. One of the largest is known as Pushshift, a social media data collection, analysis, and archiving platform founded in 2015 by Jason Baumgartner. Pushshift ingests data from Reddit's official API and collates the data into public data dumps and a livestream of new comment and post data that can be accessed by Pushshift's own unique API. The Pushshift dataset contains submissions and comments posted on Reddit since June 2005, and has been popular for researchers due to its ease of use and larger querying limits (Baumgartner et al., 2020). However, PushShift is not an exact mirror of data from Reddit (see: Gaffney & Matias, 2018 and the rejoinder from Baumgartner, 2018 for more). After posts, comments, and metadata from Reddit's API are ingested by PushShift they are functionally distinct. So, for example, once a person deletes their user history on Reddit those public comments and posts may still exist on Pushshift.

Method

Data Collection

We built our initial corpus of Reddit studies by systematically searching the ACM, EBSCOhost, EconLit, PLOS One, JStor, SCOPUS, and Web of Science databases for manuscripts that have the term "Reddit" in their title, abstract, or keywords. Our initial search was completed on 30 April 2020 and resulted in 857 studies. We stored bibliographic data for each study in the reference software Zotero. We removed duplicates, inaccessible materials (even through interlibrary loan), materials not written in English, student theses and dissertations, and books (but not book chapters). This resulted in a total corpus of 727 manuscripts (see Figure 1).
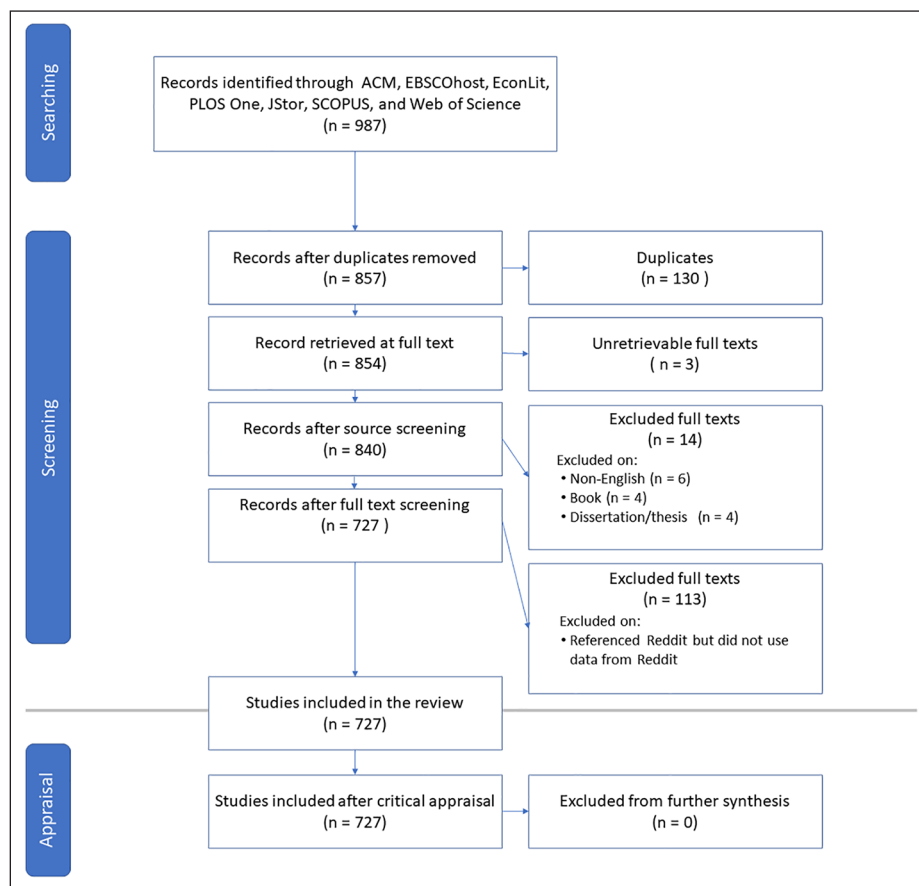

Figure 1. ROSES flow diagram for Reddit corpus construction, model adapted from Haddaway et al. (2018).
Note. We share the complete bibliography of the 727 studies in the corpus here: http://www.nicholasproferes.org/reddit-studies/

Coding

For each manuscript, we recorded the information shown in Table 1:

Table 1. Information Recorded from Each Study.

Bibliographic Information	*Author Information*
• Title of manuscript	• Lead author's institutional home
• Year of publication	○ Whether that institution is "Higher
• Authors	Ed," "A business," or "Other"
• Publication title (journal name, conference name)	○ Country of institution
• Publication type (journal, conference proceeding, book chapter, White Paper)	• Lead author's institutional unit
○ Ulrich's classification of journal discipline	○ Recategorization of unit by discipline
○ Publisher for each conference proceeding	
• Author-supplied keywords	
Study Information	*Subsequent Distribution*
• List of subreddits used as data source, if specified	• Whether or not that study has ever been shared on Reddit
○ Count of subreddits studied	• Whether or not it was shared on the subreddit(s) from which data had been collected
○ Author's description of the data collected	
○ Count of original posts collected within the study	• Whether or not it was the original researcher who was sharing the study on Reddit
○ Count of comments collected within the study	
○ Count of users represented in data	
○ Count of surveys	
○ Count of interviews	
○ Notes on other data collected	
○ Methods or tools used to collect Reddit data.	
• Data analysis method.	
• Research paradigm (Quant/Qual/Mixed/Other)	
• Ethics specific information	
○ Whether or not the manuscript explicitly mentioned an IRB or ethics review (even if the mention was "we did not seek IRB review").	
○ Whether or not the manuscript explicitly states that it sought consent for collecting data.	
○ Whether or not the manuscript used direct quotes from Reddit users.	
• Whether or not the manuscript used specific Reddit usernames in the paper	
• Whether or not the manuscript discussed ethics in any capacity in the paper	

IRB = Institutional Review Boards.

We recorded the data after reviewing each publication. The analysis we provide reflects the level of transparency and precision presented by the publication's authors. For instance, while many manuscripts noted that they analyzed certain volumes of content from Reddit, not all authors included the exact number of comments they analyzed or from which subreddits they collected data.

After reviewing the initial corpus database, we sought additional contextual information about the explicitly mentioned subreddits used as data sources. During the first week of November 2020, we collected subscriber numbers for each subreddit, as well as information about whether the subreddit was marked as 18+, private, quarantined, or banned. We note that subreddit information is not necessarily the same as when a study collected data from the subreddit. For example, a 2015 study may have collected data from a subreddit that was much smaller than today, or which has now been banned by Reddit's administrators.

Findings

Bibliographic Overview

Of the 727 manuscripts in the corpus, 338 (46.5%) are journal articles, 382 (52.5%) are conference proceedings, 6 are book chapters (0.8%), and 1 is a White Paper (0.1%). As shown in Table 2, there has been growth on a year-over-year basis of the number of publications using data from Reddit.

Table 3 categories the 338 journal publications by discipline based on Ulrich's disciplinary classification. We then created a second-level sorting based on Wikipedia's major categories of academic disciplines ("Outline of Academic Disciplines," 2021). Computer Science, Engineering and Math constituted 33% of published journal articles in this space, Medicine and Health 23%, Social Science 22%, the Humanities 17%, and Natural Sciences 5%, respectively.

Unlike existing categorization schema for journal disciplines, there is no widely accepted disciplinary categorization

of conferences. Therefore, we grouped conference proceedings by publisher (Table 4). Findings reveal that computer science-related conferences account for a majority of conference publications. This is perhaps unsurprising given that conferences proceedings are often considered high-impact by the computer science discipline (Patterson et al., 1999).

A total of 665 manuscripts provided author-supplied keywords, and Table 5 shows the top 20 most commonly

occurring. Keyword choices spanned different descriptions, from methods used in the paper, to specific topics being studied, to other social media platforms whose data were also being used. Perhaps not surprisingly, "Reddit" and "Social Media" appear as the two most consistently occurring keywords. Interestingly, several health-related topics, such as "mental health," "depression," and "eating disorders" appear in the top 20 most common keywords. Furthermore, "gender" appears, though race, class, disability, and other demographic characteristics do not.

Author Information

We captured information about the lead author's institutional home, their type of institution, the country that institution resides in, and their home department. A total of 692 of the 727 manuscripts had a lead author in a higher education institution (95.2%); 12 of the lead authors reside at businesses (1.7%), 2 of the papers were led by independent researchers (0.3%), and 21 (2.9%) manuscripts were authored by individuals at institutions we labeled as "Other" (which includes, for example, the Max Planck Institute, the Chinese Academy of Sciences, and the Pacific Northwest National Laboratory).

Table 6 provides a breakdown of the 727 publications by country of the institution of the first author. Although

Table 2. Year of Publication.

Year	Count
2010	2
2011	4
2012	1
2013	17
2014	20
2015	48
2016	66
2017	102
2018	146
2019	230
2020[a]	91

[a]Data cover only Jan–Apr.

Table 3. Count of Journal Articles Using Reddit Data by Discipline of Journal Publisher.

CS, MATH, ENG (112)	MEDICINE & HEALTH (78)	SOCIAL SCIENCE (76)
Computers (91)	Medical Sciences (55)	Sociology (27)
Library And Information Sciences (13)	Drug Abuse And Alcoholism (9)	Psychology (17)
Mathematics (5)	Nutrition And Dietetics (6)	Education (9)
Electronics (1)	Public Health And Safety (5)	Business And Economics (6)
Engineering (1)	Handicapped (1)	Social Sciences: Comprehensive Works (5)
Statistics (1)	Tobacco (1)	Social Services And Welfare (4)
	Women's Health (1)	Political Science (3)
		Consumer Education And Protection (2)
		Criminology And Law Enforcement (1)
		Children And Youth (1)
		Public Administration (1)
HUMANITIES (56)	NATURAL SCIENCES (16)	
Communications (8)	Sciences: Comprehensive Works (11)	
Linguistics (8)	Environmental Studies (4)	
Literature (6)	Biology (1)	
Journalism (5)		
Philosophy (5)		
Women's Studies (4)		
Humanities: Comprehensive Works (3)		
Anthropology (2)		
Law (2)		
Leisure And Recreation (2)		
Lifestyle (2)		
Motion Pictures (2)		
Music (2)		
Sports And Games (2)		
Architecture (1)		
Art (1)		
Religions And Theology (1)		

Table 4. Count of Conference Proceedings Using Reddit Data by Publisher.

Proceedings publisher	Count
Association of Computing Machinery (ACM)	161
Institute of Electrical and Electronics Engineers (IEEE)	77
Association for the Advancement of Artificial Intelligence (AAAI)	35
Association for Computational Linguistics (ACL)	33
CEUR Workshop Proceedings	22
Springer	12
Other[a]	42

[a]For reasons of space, all titles with less than five items have been collapsed into the "other" category.

Table 5. Top 20 Most Commonly Occurring Keywords Across Corpus.

Keyword	Count of papers using keyword (author-supplied)
Reddit	192
Social media	147
Online communities	36
Mental health	24
Machine learning	23
Social networks	23
Natural language processing	16
Text mining	16
Twitter	15
Moderation	14
Depression	13
Sentiment analysis	13
Classification	11
Anonymity	10
Deep learning	10
Gender	10
Online community	10
Content analysis	9
Crowdsourcing	9
Eating disorders	8

Table 6. Count of Publications by Country of the Institution of First Author.

Country	Count
North America	427
Europe	153
Asia	85
Australia	44
South America	14
Africa	4

North American institutions constitute a majority, research using data from Reddit is occurring in many different locations.

Table 7. Count of Institutional Units of First Author.

Recoded institutional unit discipline	Count	Percent ($n = 727$)
CS, Math, ENG	417	57.4
Humanities	132	18.2
Medicine & Health	70	9.6
Social Science	57	7.8
Natural Science	5	0.7
Other or N/A	46	6.3

Finally, we captured the name of the disciplinary unit of the first author. Some authors reported department affiliation, others college or school. We stemmed all unit names to only focus on the discipline, not the level. When the author was employed at a for-profit business or there was not a unit that could be reasonably identified, we coded the entry as "N/A." Units were recoded using the same disciplinary condensation strategy used in the journal discipline recoding. The results appear in Table 7.

Computer Science, Math, and Engineering units are the institutional homes of a majority of the lead authors of the works in the corpus.

Study Information

Subreddits. Specific subreddits were named as data sources 1,773 times within the corpus, and of these 832 were unique. Two studies generated "fake names" for subreddits they studied as a mechanism to protect the privacy of the communities. We provide a list of the most commonly studied subreddits in Table 8 (limited to 20 for reasons of space).

Within this list, there are a few noteworthy trends. First, the prominence of subreddits focusing on politics and news discussion, such as r/politics, r/worldnews, and r/The_Donald (a community that was banned by Reddit in 2020 for inciting harassment). Second, the prominence of mental health and drug subreddits, such as r/depression, r/SuicideWatch, r/bipolarreddit, and r/opiates, which may include content generated by potentially vulnerable populations. Finally, subreddits and unique communities, topics, and phenomena that are specific to Reddit, such as r/changemyview, r/IAmA, and r/ExplainLikeImFive. This list also speaks of the diversity of content being studied in relation to Reddit. While r/politics is the most frequently cited subreddit data is drawn from, it is only explicitly named as a data source in 36 manuscripts.

We also checked each subreddit for and whether there were any special meta-flags on the subreddit as of the first week of November 2020 (such as 18+, Banned, Private, Quarantined, or Restricted). Of the 832 subreddits mentioned by name, 30 (3.6%) have been banned, 17 (2.0%) were private, 16 (1.9%) were marked as being "18+," (used as a marker for pornography rather than other kinds of mature content), 4 (0.5%) were quarantined by Reddit, 1 (0.1%) was

Table 8. Top 20 Most Commonly Subreddits Used as Data Source.

Subreddit	Count
r/politics	36
r/worldnews	33
r/AskReddit	32
r/News	28
r/depression	26
r/science	23
r/changemyview	20
r/SuicideWatch	20
r/funny	18
r/IAmA	18
r/anxiety	16
r/Movies	16
r/The_Donald	16
r/askscience	15
r/pics	15
r/gaming	13
r/todayilearned	13
r/bipolarreddit	12
r/opiates	12
r/ExplainLikeImFive	11

Table 9. Subscriber Size of Studied Subreddits in Corpus.

Number of subscribers	Count of unique occurrences in corpus
Over 20 million	19
10–20 million	31
1–10 million	147
500k–1 million	50
250k–500k	60
100k–250k	105
50k–100k	59
10k–50k	141
1k–10k	99
Less than 1k	69
Data unavailable	52

Table 10. Count of Number of Subreddits Included in Study.

Number of SubReddits used as data source in manuscript	Count
Subreddits not named in paper, nor count of subreddits provided	282
1	220
2–5	109
6–10	41
11–20	30
21–50	13
51–100	4
100+	28

both private and quarantined, and 1 (0.1%) was restricted. We note that just because a subreddit currently has a meta-flag on it, does not mean it necessarily did at the time that the researchers were getting data from these sources.

We checked the subscriber count of each subreddit in early November of 2020 to map the relative size of these communities. Subscriber counts were available for 780 of the 832 unique subreddits listed. Table 9 provides a grouping of the number of subscribers in each of the 832 unique subreddits.

There is a wide range in the sizes of the communities being studied. While just shy of a quarter of the uniquely mentioned subreddits studied have a subscriber base of over 1 million, nearly 40% of the unique subreddits functioning as a data source in the corpus mentioned have a subscriber base of fewer than 50k users.

However, we observed that many studies did not explicitly list the subreddits they used as a data source. In some cases, the authors indicated a count of subreddits implicated without naming them (such as one study which indicated it had pulled data from 200+ subreddits, but did not list them), while others provided no counts nor subreddit names. Table 10 provides a breakdown of a count of subreddits included in a study, either by being explicitly named or through a quantitative measure listed in the study.

Data. Researchers used many different types of data in their work, including original posts from Reddit, comments and the comment threads on posts, meta-data about posts or comment threads, links or media from posts or comments, upvoting/downvoting information, information about sub-reddits themselves (such as rules, subscriber counts), as well as surveys and interviews with Reddit users and Reddit moderators. However, how data were described varied from paper to paper. In 204 papers, authors did not provide a description of their dataset with enough specificity to parse out, for example, the number of posts, comments, or users impacted. Furthermore, in some cases, terms were used interchangeably or inconsistently. For example, some authors indicated they collected "posts" from Reddit, but their figures would show both original posts and comment threads that followed rather than only content uploaded by the thread's originator.

We provide histograms in Figures 2 and 3 that list the number of studies using (*n*) sized datasets of posts and comments (where information was provided). Figure 3 in particular suggests that much of the research using Reddit as a data source tends toward larger datasets, particularly when comments are being analyzed.

Research methodology and data collection practices—and the degree to which they were explained—varies throughout the corpus. We found that 217 studies of the 727
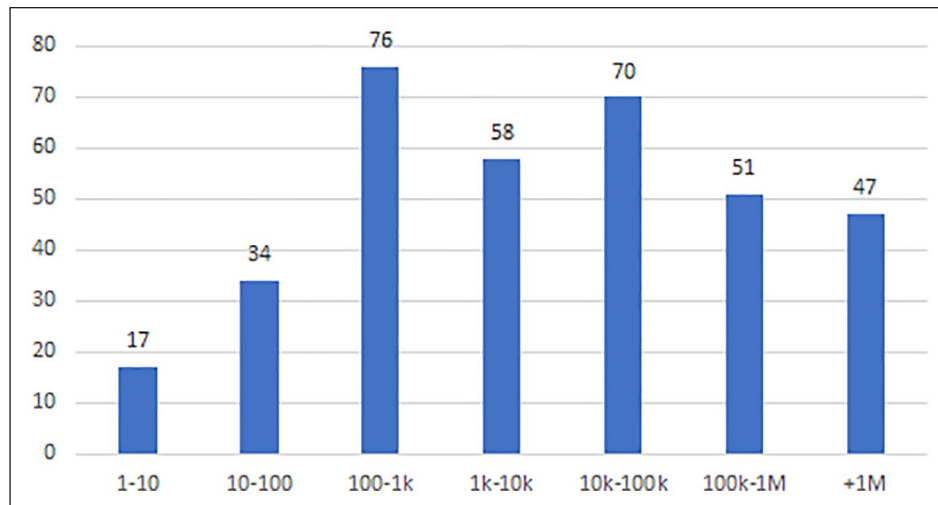
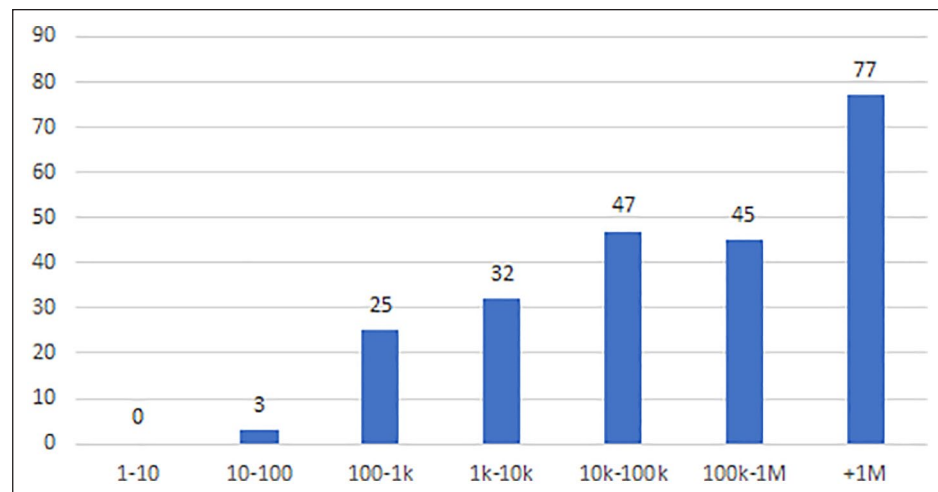**Figure 2.** Studies using *X* number of posts.

Figure 3. Studies using *X* number of comments.

in the corpus (29.8%) had no explicit description of how they collected their data. Of the remaining 510 studies, we found 597 distinct accounts of collection methods listed, with some studies listing multiple collection techniques. Table 11 provides a tabular breakdown of the most commonly occurring data collection tools below. We note that, for reasons of space, methods with less than five mentions have been grouped into the "other" category.

We also captured the analysis methods the authors used in their work. We identified 482 quantitative, 183 qualitative, and 56 mixed-method studies, suggesting a majority of the papers in the corpus are primarily using quantitative approaches to Reddit data. Table 12 provides a categorization of the analytical methods. The categories of analysis detailed in Table 12 are non-exclusive, meaning that a study can contain more than one type of analysis. For example, a study might include both content analysis and network

analysis. For reasons of space, we only include methods with at least five appearances in the corpus.

Computational analysis methods are a clear plurality. This is perhaps unsurprising given the relative ease with which researchers can collect large corpora of textual data from Reddit.

Ethics

We looked at each manuscript for mentions of an IRB or similar ethical review process (even if the mention was "we did not seek IRB review"). We found that 101 studies (13.9%) mentioned the term "IRB" or for example, "ethics review" and 626 (86.1%) did not. Of the 101, 23 were papers using interviews or surveys, methods more regularly requiring ethics-body approval. The vast majority of the remaining 78 papers mentioned ethics review while noting an "exempt"

Table 11. Most Commonly Used Data Collection Tools in Corpus.

Data collection method	Count
Reddit Application Programming Interface (API)	149
PushShift API	92
Reddit Website Directly (i.e., screengrabs, copy/pasting)	59
Other	55
Another Researcher	44
Survey	40
Google BigQuery	37
Python PRAW Package (which relies on Reddit's API)	32
Distributed 3rd Party Dataset (e.g., eRisk Dataset)	23
Interviews	20
Ethnographic Observation	16
Dataset Generated by Reddit.com itself (such as Reddit Donated Voting Datasets or Datasets Produced by SubReddit Moderators)	10
Python (Non-PRAW)	8
Experiment	7
Nvivo NCapture	5

Table 12. Count of Analysis Methods in Corpus.

Analysis type	Count of occurrences in corpus
Computational[a]	321
Content	143
Statistical	126
Discourse Analysis	55
Network	45
Survey Methods	40
Ethnographic	24
Narrative	23
Forensic	18
Grounded Theory	18
Predictive Modeling	18
Experimental	8
Case Study	7
Bibliometric	5

NLP = natural language processing; LDA = Latent Dirichlet Allocation.
[a]Computational analysis includes analysis methods such as machine learning approaches, NLP, and LDA-based topic modeling.

review status (e.g., "was exempt from ethics approval" or "approved under exempt review"). However, it is impossible to know in many cases whether "exempt" was an official designation given by a review board or whether the authors made this judgment themselves. Disclosure of ethics review is also not always the standard practice even for human subjects research, so we do not assume that papers that do not mention it did not go through some ethics review.

We examined each manuscript for whether or not the authors indicated they had some kind of consent seeking process. Many (though not all) ethics review bodies would likely view most Reddit data as "public" and therefore not require researchers to seek consent (see Vitak et al., 2017). However, ethics bodies would be likely to require consent for surveys, interviews, or the use of data from closed communities. Furthermore, as part of their own practices, some researchers seek consent for the collection of public data, particularly if they seek to build connections with a community. In all, 44 of the papers (6.1%) mention seeking consent as part of their data collection process, while 683 (94.0%) did not; 31 of those mentioning consent utilized user surveys or interview methodologies, leaving 11 which sought consent for other reasons.

We analyzed the corpus to determine whether or not researchers used specific and identifiable Reddit usernames or direct quotes from Reddit users in their publications. A total of 68 manuscripts (9.4%) explicitly mentioned identifiable Reddit usernames in their paper and 659 (90.7%) did not; 207 papers (28.5%) used direct quotes from users as part of their publications, 18 papers used paraphrased quotes, noting they were paraphrased (2.5%), and 502 (69.1%) did not include direct quotes.

As open-data and sharing research can be ethical issues, we examined whether or not authors were sharing their datasets, and whether or not each research paper has ever been shared on Reddit. In 54 papers (7.4%), the authors explicitly mentioned sharing their datasets with other researchers (often providing links within the paper), with the remaining 673 (92.6%) making no statement. We found 201 (27.6%) manuscripts from the corpus shared on Reddit, however, often not by the original authors or on the subreddits from which data was collected; 25 manuscripts had been shared on the subreddit from which data had been initially collected (with 176 being shared on other subreddits), 24 manuscripts were shared by (what we believe to be) the original researchers (with 177 being shared by other Redditors), and 8 were shared on the subreddit from which data had been initially collected by the original researchers. This suggests some research is finding its way back to Reddit, but rarely via the original authors sharing the content nor is it often being shared with the data-originating subreddit.

Discussion

What Researchers Are Studying and How They Are Studying It

The kinds of subreddits researchers are drawing data from vary considerably. However, political subreddits (such as r/politics, r/worldnews, and r/The_Donald), mental health subreddits (such as r/depression, r/suicidewatch, r/anxiety, r/bipolarreddit), and drug use subreddits (such as r/opiates) are some of the more prominent data sources in the corpus. This finding raises questions about why researchers are choosing

these specific venues as data sources. Are researchers studying Reddit for the purpose of studying Reddit-specific phenomena, or are they studying social phenomena and the fact the data are from Reddit incidental? From our review of this work, the answer appears to be both.

There are two potential problems that stem from researchers using Reddit as a vehicle for gathering data potentially without considering Reddit's context. First, there is some inherent entanglement between Reddit's site, structure, subreddit norms and conventions, and content. Models built from Reddit data may carry traces of that structure. For example, as many Reddit users see conversation sorted by its popularity, content that is more broadly agreeable, clever, funny, or even biting is more likely to be responded to. There is also sometimes gaming of Reddit's sorting algorithm which can also drive conversational patterns (Shepherd, 2020). Thus, if a researcher were to scrape every comment from a particular thread, they may end up with a larger volume of data that interact with those "top posts." The kinds of conversational patterns seen on Reddit may not mimic conversations that happen in other media with different organizational structures or affordances. Second, participation on Reddit is by and large pseudonymous and demographic information about Redditors is limited; however, we know it is majority male and skews young. Hargittai (2020) argues that those of higher socioeconomic status are more likely to be on social media, and therefore, their views oversampled in big-data research. Massanari (2017) has also observed the development of toxic technocultures on Reddit that have led to the proliferation and amplification of misogynist movements such as #GamerGate. Studies such as these suggest that researchers may need to consider the generalizability and representativeness of models built using Reddit data, particularly in the context of ongoing conversations regarding language modeling (see Bender et al., 2021).

Finally, researchers from an incredible diversity of disciplines are making use of Reddit data. However, the volume of journal articles published in Computer Science, Engineering, and Math outlets; conference papers published in Computer Science-related conference proceedings; and first authors coming from Computer Science-related academic units stand out as notable. Computational-driven textual analysis (often achieved through machine learning, natural language processing [NLP], and topic modeling using Latent Dirichlet Allocation [LDA]) stand out as major vehicles in the generation of new knowledge built on Reddit data. We note the importance of contextualizing these large-N, computational approaches with the qualitative and mixed-methods research that frequently comes from other disciplines.

The Limitations of Studying Reddit

In recent years, major social media platforms like Facebook (Freelon, 2018), Instagram, and Twitter (Bruns, 2019) have begun restricting API access (Tromble, 2021). Currently, Reddit's API is open and free but whether this wider trend eventually applies to Reddit remains to be seen. Given the reliance on Reddit's APIs for accessing research data (particularly large N data), researchers relying on Reddit APIs should take heed. Pushshift offers a compelling alternative for researchers, as shown by its prominence in the corpus. However, the mapping between Reddit data and Pushshift data is not one-to-one. It is difficult to say how researchers are confronting these challenges when relying on PushShift data, and whether or not the differences impact the validity of their insights in any meaningful way. We suggest further exploration of this issue.

Research Ethics and Reddit

Ethical norms for research conducted on online platforms and using public data are not only highly variable in different disciplines and for different methodological traditions but are also contested within research communities (Vitak et al., 2016). Although there have been calls for more open discussion of ethical issues within these communities to help establish these kinds of norms (Bruckman et al., 2017), explicit discussion of ethical considerations in this dataset of papers is uncommon. Less than 15% of papers in our dataset mention some form of ethics review; however, the disclosure of ethics review within publications is also not always the standard practice even for human subjects research. When mentions of ethics did occur, most authors were making note of their "exempt" status. However, particularly given the potentially sensitive nature of some of the data sources, we suggest that researchers do not simply rely on the adage that just because the data are public, there aren't harms that may stem from the use of the data.

Similarly, compliance with Terms of Service is not a proxy for ethical research or privacy protection (Fiesler et al., 2020). Some papers in our dataset did make note of complying with Reddit's TOS, which indeed does not explicitly prohibit data collection. However, it is worth noting that individual subreddits also have their own community guidelines, and often these include rules to protect the privacy or safety of their members (Fiesler et al., 2018). As Fiesler et al. (2020) note in their analysis of data scraping provisions in a large number of social media TOS, these policies largely lack the context that would be relevant to an ethical decision (e.g., what kind of data or what it is being used for).

Privacy, Anonymity, and Discoverability. One of the major concerns of research using public data is privacy, including what constitutes "public," whether measures should be taken to prevent the discoverability of data sources, and to what extent research subjects should be disguised (Bruckman, 2002; Markham, 2012; Zimmer, 2018). As Markham (2012) noted in her paper on "ethical fabrication," though researchers often conceptualize "public" and "private" as a binary with a clear line, people interacting online are making more

fine-tuned distinctions in reality, not just about whether something is "public" but also about the use or flow of that information. Accordingly, as many scholars have subsequently pointed out in the context of research ethics, whether something is "public" is not the only relevant question for whether data collection and use are ethical (Fiesler & Proferes, 2018; Zimmer, 2018).

We found smaller subreddits are being studied in addition to larger ones; 20% of the subreddits mentioned in our corpus have less than 10,000 subscribers, which may have implications for "participant" comfort level. Hudson and Bruckman (2004) found in their study of online chatrooms that the smaller the group being observed, the less comfortable they were with researcher presence. Fiesler and Proferes (2018) also found that Twitter users were more comfortable with their tweets being analyzed as part of larger datasets than smaller ones. Moreover, the smaller the community being studied in any context, the more difficult it is to maintain the anonymity of research subjects (Saunders et al., 2015), even without the additional complication of direct quotes being discoverable by search engines (Bruckman, 2002; Markham, 2012).

About 10% of research in the corpus used identifiable Reddit usernames in their publications, and just under 30% used direct quotes from users. While this is a fairly common practice in research papers (Ayers et al., 2018), where subreddit content is potentially sensitive (such as when the quote involves mental health, drug use, sexual activity, and is potentially from a minor), there may be outsized safety or privacy risks to those data subjects if their content is shared beyond its intended context (Dym & Fiesler, 2020). We suggest that researchers should carefully consider the risks presented to data subjects by direct quotation or username inclusion.

Dataset sharing also raises a number of thorny ethical questions and values-tensions in this research space. Open science is a laudable goal, particularly in the wake of concerns over a reproducibility crisis in social science research (Baker, 2016), though concerns about ethics are one potential barrier to sharing datasets and other research artifacts (Wacharamanotham et al., 2020). For example, redistributing datasets can deny agency to individuals who have subsequently deleted their Reddit posts. Pushshift has dealt with this problem by allowing users to request having their content removed from that service. However, users may be entirely unaware that their data are still circulating in third-party datasets shared among researchers. Some have suggested that dataset sharing upon request is a good compromise (Fiesler, 2019), and Twitter has dealt with this issue by changing its terms of service so that full JSON data is not allowed to be redistributed, instead, only Tweet IDs can be shared which must then be "rehydrated" (see Summers, 2016/2021). Deleted tweets are not rehydrated. However, this introduces a separate problem of having incomplete archives, and thus bringing the reproducibility of that work

into question. We do not have a solution to this challenge, but instead note that researchers using Reddit data should carefully consider how and why they are sharing their data.

Sharing Back With the Community. Research ethics can extend beyond the scope of notice and consent, and reducing harm. Indeed, sharing research outputs with data subjects can achieve the ethical principles of autonomy, non-maleficence, and beneficence (Ferris & Sass, 2011). In the case of the Reddit corpus, we found almost 30% of the corpus shared on Reddit, but very little of it shared back to the originating community, and little shared by the authors who had conducted research. This suggests that research that draws on Reddit makes its way back to the platform, but there may be key opportunities being missed by the researchers to actually engage with their data subjects. There are, of course, situations in which it may not make sense for a researcher to share their research; for example, when doing so may put the researchers in some kind of jeopardy (for more, see Suomela et al. [2019]). However, ethical considerations for research sharing should be in part about harm and benefit. For research that should benefit a community, it is unlikely to do so if the community does not know about it.

Ambiguous or Missing Details in Published Articles

Finally, many of the manuscripts we examined provided incomplete or ambiguous descriptions of their datasets. Nearly 30% did not describe in any significant detail how the authors collected their data. Although a handful of studies in our corpus obfuscated their data collection methods for stated ethical reasons, most research with ambiguous or missing details did not state a reason for doing so. Furthermore, the language used to describe Reddit data was found to be inconsistent across the corpus, with authors using terms such as "comment" and "post" interchangeably or inconsistently. This raises potential issues for comparisons between studies, replicating studies, and synthesizing studies in meta-analyses.

Conclusion

This article set out to provide an account of how researchers are using Reddit as a data source. First, we find growth in the volume of research using Reddit data in the past decade. Much of this work is occurring in computer science disciplines and using computational methods. However, Reddit's unique structure and demographic characteristics suggest challenges to the generalizability of knowledge and models produced using Reddit sourced data, particularly if they are to be applied to new contexts outside of Reddit. Further exploration of the limits of the generalizability of science sourced strictly from Reddit data is needed. For example, research focusing on depression or drug use may need to

consider how Reddit's user base trends toward particular demographics, and how the structure and affordances of Reddit shape the kinds of conversations that happen there and that are most visible.

The topics being studied using Reddit data vary widely, and at times, researchers are drawing data from communities on Reddit that may include vulnerable populations. While we report on the practices as described in manuscripts, researchers may be engaging in ethical practices beyond what appears on the printed page. This suggests that research into the ethical practices of researchers beyond the printed page is in order. Relatedly, while Reddit has become an important data source for researchers, there are serious questions regarding the degree to which this prominence matches users' expectations for their data. Many subreddits position themselves as small communities rather than public fora, setting up a potential mismatch between IRB interpretation of Reddit as a public space and users' understandings of the communities they are participating in. Further study is needed into Reddit users' expectations for the content they create, and their contextually driven understandings of what happens to their data.

Finally, few researchers are sharing the science they produce on Reddit, yet almost 30% of the Reddit research in our corpus appears on Reddit. This suggests an interest on Reddit broadly for research about Reddit. However, further exploration is needed to better understand the value that is (or is not) created by this kind of engagement and knowledge sharing.

ORCID iDs

Nicholas Proferes https://orcid.org/0000-0002-0295-9616

Naiyan Jones https://orcid.org/0000-0002-2567-1949

Sarah Gilbert https://orcid.org/0000-0003-2718-4121

Michael Zimmer https://orcid.org/0000-0003-4229-4847

References

Ammari, T., Schoenebeck, S., & Romero, D. (2019). Self-declared throwaway accounts on Reddit: How platform affordances and shared norms enable parenting disclosure and support. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–30.

Ayers, J. W., Caputi, T. L., Nebeker, C., & Dredze, M. (2018). Don't quote me: Reverse identification of research participants in social media studies. *NPJ Digital Medicine*, *1*(1), 30.

Baker, M. (2016). Reproducibility crisis. *Nature*, *533*(26), 353–366.

Baumgartner, J. (2018, April 7). *Update for the Reddit corpus* [Reddit Post]. R/Datasets. www.reddit.com/r/datasets/comments/8aen5g/update_for_the_reddit_corpus/

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 830–839.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf

Beninger, K. (2017). Social media users' views on the ethics of social media research. In L. Sloan & A. Quan-Haase (Eds.), *The Sage handbook of social media research methods* (pp. 57–73). SAGE.

Bruckman, A. S. (2002). Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology*, *4*(3), 217–231.

Bruckman, A. S., Fiesler, C., Hancock, J., & Munteanu, C. (2017). CSCW research ethics town hall: Working towards community norms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 113–115). https://dl.acm.org/doi/abs/10.1145/3022198.3022199

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566.

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, *2*, 32.

De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). Characterizing and predicting postpartum depression from shared Facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 626–638). http://www.munmund.net/pubs/cscw_14_1.pdf

Dym, B., & Fiesler, C. (2020). Ethical and privacy considerations for research using online fandom data. *Transformative Works and Cultures*, *33*. https://journal.transformativeworks.org/index.php/twc/article/download/1733/2445?inline=1

Farrell, T., Fernandez, M., Novotny, J., & Alani, H. (2019). Exploring misogyny across the manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 87–96). https://dl.acm.org/doi/abs/10.1145/3292522.3326045

Ferris, L., & Sass, K. (2011). Sharing research findings with research participants and communities. *International Journal of Occupational and Environmental Medicine*, *2*(3), 172–181.

Fiesler, C. (2019). Ethical considerations for research involving (speculative) public data. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1–13.

Fiesler, C., Beard, N., & Keegan, B. C. (2020). No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods

in social media terms of service. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 187–196.

Fiesler, C., Jiang, J., McCann, J., Frye, K., & Brubaker, J. (2018). Reddit rules! Characterizing an ecosystem of governance. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1), Article 1.

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of twitter research ethics. *Social Media + Society*, *4*(1), 2056305118763366.

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668.

Gaffney, D., & Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLOS ONE*, *13*(7), Article e0200162.

Haddaway, N. R., Macura, B., Whaley, P., & Pullin, A. S. (2018). ROSES RepOrting standards for Systematic Evidence Syntheses: Pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environmental Evidence*, *7*(1), 7.

Hargittai, E. (2020). Potential biases in Big Data: Omitted voices on social media. *Social Science Computer Review*, *38*(1), 10–24.

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.

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge analytica, and privacy protection. *Computer*, 51(8), 56–59.

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of Reddit automoderator. *ACM Transactions on Computer-Human Interaction*, *26*(5), 31.

Leavitt, A. (2015). "This is a throwaway account": Temporary technical identities and perceptions of anonymity in a massive online community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 317–327). https://dl.acm.org/doi/abs/10.1145/2675133.2675175

Long, K., Vines, J., Sutton, S., Brooker, P., Feltwell, T., Kirman, B., Barnett, J., & Lawson, S. (2017). "Could you define that in bot terms"? Requesting, creating and using bots on Reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3488–3500). Association for Computing Machinery.

Markham, A. (2012). Fabrication as ethical practice. *Information, Communication & Society*, *15*(3), 334–353.

Marotti, A. (2018, April 23). Reddit to open Chicago office as part of advertising push. *Chicagotribune.Com*. https://www.chicagotribune.com/business/ct-biz-reddit-chicago-office-20180418-story.html

Marwick, A. E. (2017). Scandal or sex crime? Gendered privacy and the celebrity nude photo leaks. *Ethics and Information Technology*, *19*(3), 177–191.

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346.

Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, *3*(1), 2053951716650211.

Mills, R. A. (2018). Pop-up political advocacy communities on reddit.com: SandersForPresident and The Donald. *AI & Society*, *33*(1), 39–54.

Mittal, A., & Goel, A. (2010). *Stock prediction using Twitter sentiment analysis*. http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf.

Mittos, A., Zannettou, S., Blackburn, J., & De Cristofaro, E. (2020). "And we will fight for our race!" A measurement study of genetic testing conversations on Reddit and 4chan. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 452–463.

Outline of academic disciplines. (2021). *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Outline_of_academic_disciplines&oldid=1006927472

Patterson, D., Snyder, L., & Ullman, J. (1999, September). Evaluating computer scientists and engineers for promotion and tenure. *Computing Research News*. https://cra.org/resources/best-practice-memos/evaluating-computer-scientists-and-engineers-for-promotion-and-tenure/

Reddit.com. (2020a). *Reddit's 2020 year in review*. https://reddit-blog.com/2020/12/08/reddits-2020-year-in-review/

Reddit.com. (2020b). *User agreement—October 15, 2020—Reddit*. Reddit User Agreement. https://www.redditinc.com/policies/user-agreement-october-15-2020

Reddit.com. (2021, January 17). *Advertising—Audience—Reddit. Discover what makes Reddit ads unique*. https://web.archive.org/web/20210117184818/https://www.redditinc.com/advertising/audience

Roose, K. (2021, January 28). The GameStop reckoning was a long time coming. *The New York Times*. https://www.nytimes.com/2021/01/28/technology/gamestop-stock.html

Saunders, B., Kitzinger, J., & Kitzinger, C. (2015). Anonymising interview data: Challenges and compromise in practice. *Qualitative Research*, *15*(5), 616–632.

Shepherd, R. P. (2020). Gaming Reddit's algorithm: R/the_donald, amplification, and the rhetoric of sorting. *Computers and Composition*, *56*, 102572.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 Boston Marathon bombing. In M. Kindling & E. Greifeneder (Eds.), *iConference 2014 Proceedings* (pp. 654–662). iSchools.

Summers, E. (2021). *DocNow/hydrator* [JavaScript]. Documenting the Now. https://github.com/DocNow/hydrator (Original work published 2016)

Suomela, T., Chee, F., Berendt, B., & Rockwell, G. (2019). Applying an ethics of care to internet research: Gamergate and digital humanities. *Digital Studies/Le Champ Numérique*, *9*(1), 4.

Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media + Society*, *7*(1), 2056305121988929.

Tufekci, Z. (2014). Big questions for social media Big Data: Representativeness, validity and other methodological pitfalls. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), Article 1.

Vincent, J. (2017, August 22). Transgender YouTubers had their videos grabbed to train facial recognition software. *The Verge*. https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai-facial-recognition-dataset

Vitak, J., Proferes, N., Shilton, K., & Ashktorab, Z. (2017). Ethics regulation in social computing research: Examining the role of institutional review boards. *Journal of Empirical Research on Human Research Ethics*, *12*(5), 372–382.

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). https://dl.acm.org/doi/abs/10.1145/2818048.2820078

Wacharamanotham, C., Eisenring, L., Haroz, S., & Echtler, F. (2020, April). Transparency of CHI research artifacts: Results of a self-reported survey [Conference session]. CHI '20: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, United States.

Zimmer, M. (2018). Addressing conceptual gaps in Big Data research ethics: An application of contextual integrity. *Social Media + Society*, *4*(2), 2056305118768300.

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250–261.

Author Biographies

Nicholas Proferes (Ph.D., University of Wisconsin-Milwaukee) is an Assistant Professor at Arizona State University's School of Social and Behavioral Sciences. His research interests include users' understandings of social media, technological discourse, and issues of power and ethics in the digital landscape.

Naiyan Jones (PGDip., University of Birmingham) is a researcher in the UK's Office for National Statistics. His scholarly work focuses on online communities, social media platforms, and political communications.

Sarah Gilbert (Ph.D., University of British Columbia) is a postdoctoral researcher in the College of Information Studies at the University of Maryland College Park. She studies content moderation, online communities, and research ethics.

Casey Fiesler (Ph.D., Georgia Institute of Technology/J.D. Vanderbilt University Law School) is an Assistant Professor in the Department of Information Science at the University of Colorado Boulder. She studies social computing and governance, including the ethical and legal implications of researching and designing technology.

Michael Zimmer (Ph.D., New York University) is an Associate Professor in the Department of Computer Science at Marquette University. His work focuses on data ethics, digital privacy, internet research ethics, and the broader social and ethical dimensions of emerging technologies.