Ethical pitfalls for natural languages processing in psychology

Ethical pitfalls for natural language processing in psychology[1]

Mark Alfano

Emily Sullivan, Eindhoven Technical University

Amir Ebrahimi Fard, Delft University of Technology

Abstract

Knowledge is power. Knowledge about human psychology is increasingly being produced using natural language processing (NLP) and related techniques. The power that accompanies and harnesses this knowledge should be subject to ethical controls and oversight. In this chapter, we address the ethical pitfalls that are likely to be encountered in the context of such research. These pitfalls occur at various stages of the NLP pipeline, including data acquisition, enrichment, analysis, storage, and sharing. We also address secondary uses of the results and tools developed through psychometric NLP, such as profit-driven targeted advertising, political campaigns, and domestic and international psyops. Along the way, we reflect on potential ethical guidelines and considerations that may help researchers navigate these pitfalls.

Keywords: natural language processing, psychometrics, big data, ethics, privacy, Cambridge Analytica, psyops, dual use

[1] The research leading to this chapter was supported by GRANTS.

> "This has caused me the greatest trouble and still does always cause me the greatest trouble: to realize that *what things are called* is unspeakably more important than what they are. The reputation, name, and appearance, the worth, the usual measure and weight of a thing -- originally almost always something mistaken and arbitrary, thrown over things like a dress and quite foreign to their nature and even to their skin -- has, through the belief in it and its growth from generation to generation, slowly grown onto and into the thing and has become its very body"
>
> ~ Friedrich Nietzsche, *The Gay Science* section 58, translated by Josefine Nauckhoff. Cambridge University Press.

Introduction

Psychologists study patterns of thought, feeling, motivation, emotion, and behavior in the human animal. Traditionally, such research has been done by examining how people speak and act under controlled conditions, guided by hypotheses about what situational factors (e.g., the DIAMONDS of social psychology -- see Rauthmann et al. 2014) and personality factors (e.g., the Big Five or Big Six of personality psychology -- see Peabody & Goldberg 1989; Ashton et al. 2004; Saucier 1997) are likely to explain the variance observed in these patterns of thought, feeling, motivation, emotion, and behavior. Typical experiments involve a few dozen to a few thousand participants, with a recent trend towards larger, multi-lab studies that aim to produce more robust and trustworthy findings (e.g., Klein et al. 2014; Klein et al. 2017; Ebersole et al. 2016).

In most of this work, the data does not exist prior to the beginning of the research project: data is collected to help test a hypothesis, though it may later be re-analyzed with a different hypothesis in mind. However, in the last decade or so, some psychologists (e.g., Hoover et al. 2018, 2019; Pennebaker 2011), computational linguists (e.g., Waseem & Hovy 2016), computer scientists (e.g., Caliskan, Bryson, & Narayanan 2017), and experimental philosophers (e.g., Christen, Alfano, & Robinson 2017; Alfano, Carter, & Cheong 2018; Alfano, Higgins, & Levernier 2018) have flipped the script. Under the new paradigm, large, extant textual (and visual) corpora are curated, enriched, and mined using natural language processing (NLP) methods for insights and evidence that would be difficult or impossible to gather using traditional methods. In this chapter we canvass the ethical and political pitfalls that researchers may encounter in the context of such research.

A brief history of natural language processing in psychology

The new NLP paradigm in some ways resembles the much older tradition of psycholexical analysis that dates back to Francis Galton (1884). The basic idea behind that approach was that a natural language is more likely to include a term for a property or phenomenon to the extent that the property or phenomenon is important to those who speak the language. For example, English has the word ‘defenestrate’ because it’s been important to be able to talk about events in which someone is thrown out a window (often resulting in their death). By contrast, English lacks a word for someone being thrown *in* a window, presumably because such events were sufficiently rare and not-worth-talking-about that no one bothered to coin the term ‘infenestrate’. This is not to say that every phrase or term refers. There are no unicorns despite the existence of the word

'unicorn'. Nor is it to say that everything worth talking about is already denoted by a phrase or term. Words and phrases are sometimes coined because new phenomena come into existence or in order to make it easier and more interpretable to refer to phenomena that -- for various reasons -- had hitherto evaded our linguistic resources (e.g., 'sexual harassment' -- on which see Fricker 2007). In any event, it is hard to deny the rough generalization that there is a positive correlation between the importance and prevalence of phenomena in the lives of the speakers of a language and the existence of a term in the language that refers to those phenomena. Given these considerations, researchers generally agree that it would be foolhardy to "ignore such a storehouse of accumulated wisdom as a natural starting-point for the study of behavioral attributes" (Wiggins 1973, p. 329). Studying the psychological language of a society is thus an indirect way of studying the psychological properties that members of that society care about.

Researchers in the psycholexical tradition did not stop there, though. They also argued that the semantic structure of a language reflects to some extent the structure of the phenomena the language describes. In personality psychology, this insight was used by Allport & Odbert (1936) to create a taxonomy of thousands of personality-relevant terms, which they argued represents the popular conception of personality. The step from language to conceptions of personality is not identical to the step from conceptions of personality to actual personality, but it is tempting to think that there will at least be a positive correlation between how people think about personality and how personality actually is. This two-step connection (from language to conceptions, and then from conceptions to actual personality) has been empirically validated by personality models such as the Big Five (Peabody & Goldberg 1989) and the Big Six (Ashton et al. 2004; Saucier 1997).

More recent work uses NLP to examine much larger corpora and attempts, among other things, to measure individual differences in people's attitudes and personalities based on what is sometimes called their digital footprint: the text, emoji, memes, and videos they post online, as well as the content that they "like," "share," "reply to," and so on. For example, Kosinski, Stillwell, & Graepel (2013) showed that it is possible to predict someone's psychological traits (e.g., Big Five, intelligence) and a range of sensitive attributes (e.g., sexual orientation, ethnicity, religious views, political views, use of addictive substances, parental separation, age, gender) using their pattern of "likes" on Facebook. This is an exciting development, and one that has been borne out by subsequent studies and meta-analyses (Azucar, Marengo, & Settani 2018). It takes a non-trivial amount of time for participants to respond to the Big Five personality questionnaire, which consists of dozens of items. Typically, participants must be paid to provide this data, which must be funded by research grants or other resources. If participants know that their personality is being measured, they may try to game the questionnaire. And the questionable ecological validity of lab studies is always a challenge to be overcome. By contrast, researchers using NLP can gather or access immense datasets with relatively low costs in terms of time and participant payments. If they use historical data (e.g., the record of someone's tweets), then gaming the study is impossible. And if they examine patterns in people's real-world behavior, the challenge of ecological validity evaporates. In addition, NLP can be applied to corpora produced by past generations, making it possible to study the psychology of people who are now dead and thus could not possibly respond to a questionnaire. And NLP can be applied to corpora of underrepresented and hard-to-reach populations, which promises to help psychology

overcome its WEIRD problems (Henrich, Heine, & Norenzayan 2010). For these reasons and more, the use of NLP in psychological research has much to recommend it.

The bloom comes off the rose

If you've been paying attention to the news over the last few years, you probably came away with a more dismal picture of NLP methods in psychology. Why might that be?

As Hirsh et al. (2012) showed, the psychological profile that can be built using someone's digital footprint can be used to help craft personalized persuasive messages. For instance, people high in openness might respond more positively to a message that appeals to novel experiences ("See the world!"), whereas people high in conscientiousness might respond more positively to a message that appeals to the importance of planning ("Know where you're going at every step of your journey!"). This kind of personalization can be done rather cheaply and at scale (Matz et al. 2017). The prospect of more persuasive power raises at least two ethical concerns. First, what are people going to be persuaded to think or do? If they are persuaded to hate immigrants and vote for racist politicians, that would obviously be bad. Second, is it objectionably manipulative to use informational asymmetries to more effectively persuade people? If they don't know that the messages used to target them were chosen based on psychological profiling, we might wonder whether their autonomy or consent has been undermined.

In the vast majority of cases, personalized persuasive messages that rely on NLP-enabled psychological profiling involve targeted online advertising with the goal of selling products and services. In other words, most of this activity is driven by the profit motive. However, such

persuasive messaging can also be put to use in the service of political campaigns, including activities ranging from above-board messaging that discloses its methods to domestic psyops and international electoral interference. For example, during the 2016 United States Presidential election campaign, Aleksandr Kogan, a Russian-American who at the time worked at the University of Cambridge, wrote a Facebook app, thisismydigitallife, that was distributed as a "quiz" that users could take to learn more about their own personalities (Chang 2018). The app exploited a loophole in the Facebook application programing interface (API) to collect data not just from quiz-takers but also from their "friends." Approximately 270,000 users took the quiz, which led to the exposure of data from 87 million total users.[2] Kogan then passed this data to Cambridge Analytica, an American company founded by Steve Bannon, which provided services to Donald Trump's presidential campaign. Cambridge Analytica functioned as a shell corporation surrounding the British SCL Group, a PR firm with clients around the globe; in violation of US electoral law, British and other foreign national employees of SCL Group in turn used Kogan's data to serve personalized political ads to potential American voters in the run-up to the 2016 election, in hopes of persuading them to vote for Trump or -- if they were not likely Trump supporters -- for a third-party candidate or not at all. The same data, or models based on it, may also have been passed (via Paul Manafort and Konstantin Kilimnik) to Russian state counterintelligence officers and to one or more Israeli intelligence firms, such as Inspiration or Psy-Group, for use in politically persuasive messaging (Abramson 2019). Although Trump lost

---

[2] Facebook has been fined a record-setting five billion dollars for this data breach. URL = < https://www.theguardian.com/technology/2019/jul/24/facebook-to-pay-5bn-fine-as-regulator-files-cambridge-analytica-complaint >, accessed 2 October 2019.

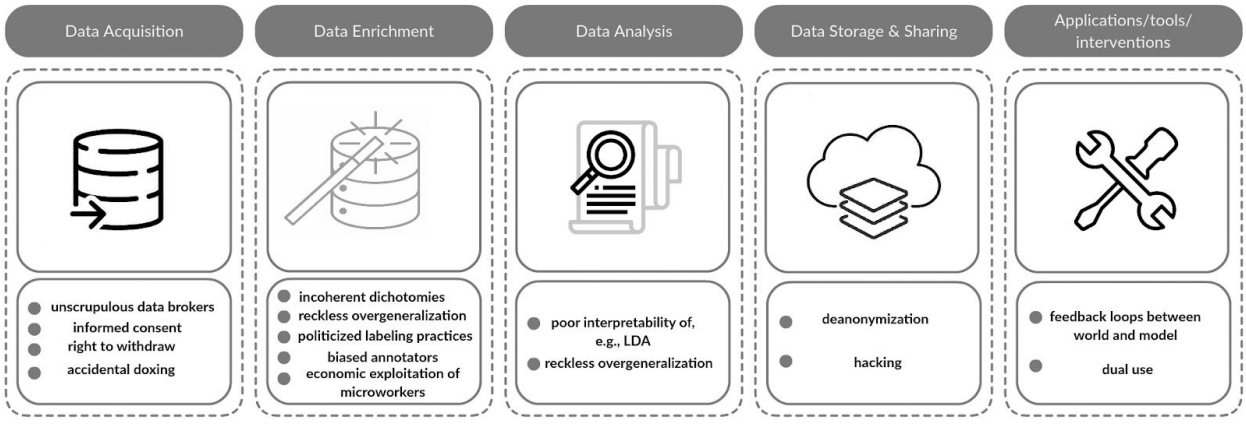
the popular vote by nearly three million ballots, he managed to cobble together a victory in the byzantine and anti-democratic Electoral College. Had just 78,000 voters in the crucial states of Pennsylvania, Wisconsin, and Michigan voted for Clinton rather than Trump, she would have won the Electoral College.

This razor-thin margin suggests that the outcome of the election was overdetermined: any number of factors could have made the difference. Kathleen Hall Jamieson (2018), a professor at the University of Pennsylvania, argues that the psyops campaign run by Cambridge Analytica, Russian state hackers, and potentially other actors was enough to tip the balance. For obvious reasons, it is unknown how effectively-targeted the advertisements associated with this illicit campaign were, and on the basis of what characteristics they were targeted. A postmortem assessment indicates that the ads were mainly focused on hot-button issues such as the Black Lives Matter (BLM) movement, immigration, Christianist and Islamophobic sentiment, gay rights, incel activism, anti-feminism, and gun control (DiResta et al. 2018). Knowing whether someone is extroverted or conscientious may not help much when targeting ads associated with these topics. However, as mentioned above, the same NLP techniques that make it possible to estimate someone's Big Five attributes can also be used to infer characteristics such as sexual orientation, ethnicity, religious views, political views, use of addictive substances, parental separation, age, and gender. It's not hard to imagine that targeting only straight men with incel-related ads would be more efficient than targeting both straight men, gay men, and women. It's not hard to imagine that targeting only Christians with Islamophobic ads would be more effective than targeting both Christians and Muslims. Likewise, it's not hard to imagine that targeting only African Americans with BLM-related ads that discourage voting would be more

efficient than targeting all users regardless of their race. And it's not hard to imagine that targeting only gay people with ads that discourage voting based on Clinton's ambivalent record on gay rights would be more effective than targeting all users regardless of their sexual orientation. While it remains unclear how effective these covert campaigns were (Gibney 2018), they nevertheless ring ethical and political alarm bells. After all, even if Cambridge Analytica's data and models were not up to the task in 2016, there's no in-principle reason why a more sophisticated operation would fail in future electoral campaigns.

The Cambridge Analytica-Trump scandal is a high-profile dramatization of one of the ways things can go ethically (and politically) wrong when people use NLP to do psychological research or design tools and interventions that draw on such research. However, such psyops are far from the only ethical pitfall that NLP researchers face. In the remainder of this chapter, we attempt to taxonomize these pitfalls based on the stage of the pipeline they are most associated with: from data acquisition to enrichment, analysis, storage, and sharing (Figure 1).

Figure 1: The NLP pipeline



We also address dual use of the results and tools developed for NLP research, such as profit-driven targeted advertising, political campaigns, domestic and international psyops, and the automated curation of enemies lists by governments. While there have been efforts to address ethical issues that arise in connection with social data more broadly (Daly, Devitt, & Mann 2019), to our knowledge this chapter represents the first attempt to think through the ethical pitfalls of the entire NLP pipeline in a focused way.

Pitfalls of data acquisition

NLP researchers work with corpora, which need to be bought, borrowed from open-source repositories, or built. All three methods of acquiring corpora face ethical pitfalls.

In the case of bought and borrowed corpora, what's ready-to-hand may already embody ethically problematic features. For instance, if researchers were to purchase the psychological data acquired by Aleksandr Kogan's Facebook app mentioned above, they would be receiving data that was not collected from informed, consenting adults. This is perhaps an obvious problem, but it is related to broader concerns regarding informed consent. When the NLP researcher knows with certainty that a corpus was acquired in a way that violated informed consent protocols or principles, they should refuse to buy or borrow it. However, in many cases, they will not be able to verify one way or the other the actual circumstances or process of data acquisition. They will be forced to accept the data broker's assurance on trust. Contemporary data brokerage, however, is the Wild West: there are many unscrupulous actors, and few consensus norms of conduct. Unless the broker in question is a well-established actor with a

track-record and some form of external oversight or auditing in place, it would be foolhardy to just take their word for it.

Even then, there remains the problem of ensuring that participants who consented when their data was first acquired have not withdrawn their consent. Consent is arguably best understood not as one-off agreement or acquiescence but as ongoing, affirmative acceptance (Helgesson & Johnsson 2005; McConnell 2010). In other words, research participants are ordinarily presumed to have the right to withdraw from a study; this right covers both cases in which participants pull out of the study while data is still being collected and cases in which participants indicate later that they would like their data removed from any analysis and deleted. Because NLP data is typically anonymized or pseudonymized[3], it may be impossible to know which text and other data to remove from a corpus should a participant express the desire to withdraw. Additionally, if a participant's data is used to train a supervised learning algorithm, it will generally be impossible to cleanly extract that data's contribution from the algorithm without retraining the whole model from scratch.

These concerns are less pressing when the corpus is derived from publicly accessible sources, such as newspapers, open government APIs, and the APIs of publicly-facing social platforms such as Twitter. However, even in these cases, researchers should be cautious. Someone who provides a quote to their neighborhood newspaper on the record is of course speaking publicly, but their primary audience is small and local. Likewise, an unverified tweep with a few dozen followers is speaking publicly when publishing a tweet, but their primary audience is small. Such actors may have an expectation of obscurity, if not outright privacy or

[3] Though see below for concerns about the security of pseudonymization and anonymization.

pseudonymity. Whether such an expectation of obscurity is warranted or not is, in a way, irrelevant. NLP researchers who draw the attention of thousands of strangers (some of whom may engage in targeted harassment) to otherwise-obscure individuals may inadvertently destroy their obscurity, which might even amount to a form of doxing (Douglas 2016). NLP researchers should exercise due care in anonymizing or pseudonymizing even public corpora when obscurity is at stake in this way.

A separate ethical pitfall, which also relates to built corpora, has to do with representation. Above, we pointed out that NLP promises to help psychology overcome its WEIRD problems. However, this promise will only be fulfilled if researchers make a point of acquiring or building corpora with an eye to including textual sources from non-Western, less educated, non-industrialized, impoverished, and/or non-democratic groups. The cheapest, best documented, and most accessible corpora will in many cases be just as WEIRD as the American undergraduates on whom so much extant psychological research is based. There is thus a risk of erasure in this sort of research. To mitigate this risk, psychologists using NLP should carefully document how much diversity (including linguistic diversity, as most NLP is still done only for Western European languages, Arabic, and Mandarin) is embodied in their corpora -- both for their own studies and for any subsequent studies that may draw upon and unthinkingly generalize from these corpora. This documentation could be provided in the form of a short summary "card" attached to the corpus and any results and tools grounded in it, much like the "model cards" recommended by Mitchell et al. (2019) for machine-learning models.

One challenge to creating genuinely representative corpora is that underprivileged members of society are less likely to have published written text that can be entered into a

corpus. A stark example of this problem is the Corpus of Founding Era American English (COFEA 2019). Built by legal scholars at Brigham Young University as a supplement to the Corpus of Historical American English (COHA) and the Corpus of Contemporary American English, COFEA is a tool for "originalist" statutory and constitutional interpretation. According to originalists, the meaning of a law is fixed by the public meaning of its constituent language at the time the law was written (Scalia 1997). On this view, legal interpretation is a matter of inferring psychological states from natural language: what would have been intended and understood by competent speakers of the language at the time and in the place where the law was promulgated? In order to shed light on this question, legal scholars have begun to make use of rather primitive NLP techniques, such as frequency analysis, n-gram analysis, and collocation analysis (e.g., Lee & Phillips 2019). Indeed, a recent decision by the Utah State Supreme Court (on which Lee sits), *Richards vs. Cox*, made use of NLP and both COCA and COHA to infer the alleged original meaning of the bigram 'employment in'.

Whereas both COCA and COHA seem to live up to their promise of diversity and representativeness, COFEA is ethically troubling. It consists of nearly 120,000 documents comprising 133 million words. About half of these words are contributed by a sub-corpus known as Evans Early American Imprints, which includes "books, pamphlets and periodical publications printed in the United States of America" between the years of 1639 and 1820. About a quarter of the words come from HeinOnline; these are primarily laws, executive department reports, and legal treatises. Presumably all or almost all of them were written by propertied white men. An additional 30% of the words in COFEA come from the correspondence of just six propertied white men (George Washington, Benjamin Franklin, John Adams, Thomas Jefferson,

Alexander Hamilton, and James Madison), four of whom held property in the form of slaves and one of whom (Madison) proposed the notorious three-fifths compromise. It is not clear to us what percentage of the words in COFEA are attributable to women, nor what percentage are attributable to enslaved Blacks, free Blacks, or Native Americans. Almost certainly, though, these groups are massively underrepresented in COFEA. For this reason, we find it ethically problematic that Lee & Phillips (2019) recommend using COFEA as the sole corpus from which to infer the meaning of phrases such as 'domestic violence' and 'commerce'. Indeed, their own analysis shows that two of the twenty most frequent collocates of 'commerce' in COFEA are 'shackle' and 'shackled'.

While the example of COFEA may seem tangential to the *psychological* use of NLP, we think it is instructive. Most psychologists are now familiar with the fact that experimenter degrees of freedom can be used to manipulate a study in order to increase the chance that its results are statistically significant or consistent with the experimenter's expectations or preferences (Simmons, Nelson, & Simonsohn 2011). What we want to emphasize here is that researcher degrees of freedom in assembling a corpus can similarly be used to manipulate it in order to increase the chance that NLP analysis of the corpus will produce results that are consistent with the researcher's expectations and preferences. One particularly egregious form of this practice involves the erasure of the voices (or writings) of underrepresented and underprivileged members of a community, and the resultant amplification of dominant members of that community. As we discuss further below, the biases of the authors of a corpus are guaranteed to be embodied in the corpus itself, as well as in any analyses and tools dependent on that corpus (Caliskan, Bryson, & Narayanan 2017). And it is not just legal corpora that are at

stake here, as Pfeffer, Mayer, & Morstatter (2018) recently demonstrated that similar manipulative influences can be exerted on the Twitter API.

Pitfalls of data enrichment

A raw text corpus is a valuable resource, but in many research projects it is necessary to enrich the corpus in some way in order to make fuller use of it. Enrichment as we understand it here is the process of adding inferred data to a corpus. Enriching can be either manual or automated, and it can be done at the word level or the document level. For instance, researchers might use automatic part-of-speech tagging or named entity recognition to enrich the words contained in each document in a corpus of newspaper articles. Alternatively, they could try to infer the race, gender, and ethnicity of each author of each document from the authors' names. Or they could try to identify by hand (perhaps using crowd-sourcing) instances of sarcasm, toxic speech, or incitement to violence within documents. Or they could try to classify news stories as falling into one or another class of themes using topic modeling. The range of possible enrichments is quite wide.

Classification is a dominant supervised learning approach for NLP problems. In this approach, a model is built by training a classifier with a dataset comprising samples of predetermined classes. This means the raw dataset has to be annotated with desired classes before being fed to the model. In many cases, such data annotation is a smooth procedure. For instance, if we would like to build a news classifier to determine whether a news item belongs to politics, culture, sport, or some other category, labelling the training data is not controversial because each of the politics, culture, and sport categories has a relatively clear conceptualization.

However, classification can be less straightforward for application domains such as rumor detection and hate speech. In such cases, the model conventionally is trained by two classes: rumor / non-rumor or hate speech / non-hate speech. In the case of rumor, researchers most often refer to definitions with similar elements, thanks to years of work by rumor scholars that ultimately led to relative convergence on rumor conceptualization (DiFonzo & Bordia 2007). This consensus on what constitutes or reliably signals the presence of a rumor allows researchers to annotate rumor-related documents consistently. But what about non-rumors? How exactly is the notion of non-rumor defined? Are rumor and non-rumor complementary concepts, in such a way that by specifying one of them, the other will be automatically specified? Or they are not complementary, and some data points fall into neither rumor nor non-rumor categories? The same questions can be raised about the hate speech / non-hate speech dichotomy. In both cases, the choice of conceptualization and operationalization may have ethical (or unethical) influences and implications.

Non-rumor is a fuzzy concept that was coined by computer scientists who used binary classification for rumor detection. In the literature on computational rumor detection, there are two major interpretations regarding non-rumor. One of them takes non-rumor to be news items that are extracted from credible news sources (Kwon et al, 2017), while the other treats non-rumors as any related information that cannot be given the rumor label (Zubiaga et al, 2016). The lack of consensus regarding the conceptualization of labels is a serious pitfall in data annotation. It allows the researchers to formulate their own ad hoc conceptualization, which leads to several issues that are at once methodological and moral.

First, the models used by different research groups (or by the same research group on different projects or at different times) become inconsistent and unreliable as we cannot be sure about their functionalities and what they separate. This undermines reproducibility and may even make the theoretical claims associated with such research unfalsifiable. Making policy or decisions informed by these theoretical claims would be (and, indeed, already sometimes is) reckless.

Second, when different researchers use different conceptualizations and operationalizations of the same phenomenon, we cannot compare the outcomes of their models because they measure different things. This is a challenging problem to tackle; however, there is a practical approach called one-class classification which can ignore the controversial class (which is non-rumor in rumor detection problem, non-hate speech in the hate speech detection problem) and is trained by the class on which we have achieved consensus (Fard et al. 2019).

Third, exogenous and illicit pressure may influence how a class is conceptualized, operationalized, and labeled. For example, Twitter uses NLP classification tools to identify accounts affiliated with or supportive of the Islamic State (ISIS or ISIL) and automatically suspends or bans such accounts for violating rules against "abuse and hateful conduct." In so doing, they inevitably sweep up some accounts (including anti-ISIS activists) that are not actually affiliated with or supportive of ISIS. These false positives are widely agreed to be a price worth paying for the expungement of ISIS propaganda and snuff films. However, Twitter pointedly does *not* use the same tools to classify and suspend accounts affiliated with or supportive of white supremacy because doing so in an automated way is guaranteed to sweep up

some accounts that Jack Dorsey and his colleagues are reluctant to suspend. Presumably there would be some false positives for anti-racist and anti-fascist accounts, as there are in the parallel case of ISIS. However, recent journalism by Cox & Koebler (2019) indicates that the main consideration holding back Twitter from automatically banning white supremacist accounts is that the accounts of many Republican politicians, such as Steve King and Donald Trump, as well as their most vocal supporters, would be swept up in such a purge. Perhaps understandably, Dorsey is reluctant to provoke the ire and attacks that would certainly follow such a purge. In this case, the conceptualization, operationalization, and use of a hate speech classifier are being shaped by exogenous influences.

Perhaps Dorsey would respond to this criticism by asking, "Who am I to judge whether prominent politicians with millions of supporters are engaging in hate speech?" Such an abdication of responsibility might be followed by a call for crowd-sourcing the flagging of hate speech and other forms of toxic speech. Such crowd-sourcing, however, introduces new methodological and moral problems. For example, Davidson et al. (2017) found that some forms of hate speech (especially homophobic and racist hate speech) can be annotated successfully via crowd-sourcing but that other forms of hate speech (especially misogynist hate speech) were harder to address. It is unclear what causes this divergence, but one plausible explanation is that misogyny is so widespread and unstigmatized that crowd-sourcing the enrichment process simply invites prejudiced participants to encode their own biases in the allegedly "gold standard" or "ground truth" of the enriched corpus. Waseem (2016) has shown that expert annotators outperform lay annotators, which is consistent with this line of reasoning.

These considerations raise broader questions about the ethics of enriching corpora and other datasets using crowd-sourced methods. There are two quite distinct additional problems that arise in this connection. First, if the crowd of annotators embodies morally objectionable attitudes such as racism, sexism, homophobia, classism, and so on, then they are liable to incorporate those attitudes into the enriched dataset. A striking recent dramatization of this phenomenon was the ImageNet Roulette app released by Kate Crawford and Trevor Paglen 2019.[4] ImageNet is a vast repository of images that have been enriched with labels for the objects pictured in them by mTurk workers and others.[5] While many of the labels in this impressive dataset (which involves both NLP and image-recognition elements) are entirely unproblematic, the ones associated with people are ethically troubling. These include anodyne descriptors like 'tennis player' but also slurs like 'jigaboo', 'yid', 'fatso', 'faggot', and 'whore'. Via the ImageNet Roulette app, Crawford and Paglen made it possible for people to upload selfies or other images, which were then automatically labeled by a convolutional neural network trained on the ImageNet dataset. In an opinion piece published recently by the *Guardian*, Julia Carrie Wong reflects on what it felt like to have an AI call her a "gook, slant-eye." This art project / demonstration helped to dramatise how human prejudice can end up encoded in enriched datasets that are subsequently available for allegedly ethically-neutral use.

---

[4] The website for this app has since been taken down, but a description of it is available at url = < https://www.excavating.ai/ >.

[5] The database is available at url = < http://www.image-net.org/ >.

We cannot help but wonder whether the kind of problem described here would have cropped up in such a virulent form had the research teams that initially developed so many of the tools and datasets that are now foundational to NLP, AI, and the semantic web been more diverse along the dimensions of race, ethnicity, gender, class, sexual orientation, national origin, and so on. This is not to suggest that only women care about or are sensitive to misogynistic slurs or that only people of color care about or are sensitive to racist slurs, but a moderate form of standpoint epistemology suggests that they are likely to care more and be more sensitive, and perhaps induce their collaborators to care more and be more sensitive (Intemann 2010).

A very different concern about crowd-sourcing relates to the working conditions of the enrichers of datasets on microworking platforms such as Amazon Mechanical Turk (mTurk) and Crowdflower. Not all microworking tasks are NLP-enrichment tasks, but such tasks are not especially distinctive on these platforms in terms of worker compensation (Hara et al. 2018). Recent analyses indicate the about half of the microworkers on these platforms are located in the United States, and that a large plurality of the remainder are located in India (Berg 2016). American microworkers tend to gravitate towards these employment platforms to supplement other sources of income or because they have reason to work from home -- in some cases because they simply prefer to, but in other cases because they have care obligations or disabilities that prevent them from working outside the home. Indian microworkers tend to use mTurk or Crowdflower not to supplement their incomes but as a primary source of income; they too often list the desire or need to work from home as one of their main reasons for engaging in microwork.

The fact that many microworkers have few alternative employment options suggests that they are a vulnerable workforce subject to various forms of exploitation. And indeed, this concern is borne out by recent surveys. For example, Boyd (2016) found that nearly one quarter of the time microworkers spend on these platforms is essentially unpaid labor, such as searching for tasks, corresponding with task-requesters, and performing qualifying tasks in order to be allowed to perform (and be paid for) more lucrative or more interesting tasks. American mTurkers' adjusted median wage was $4.65 per hour. The federal minimum wage in the United States at the time Boyd's paper was published was $7.25 per hour, meaning that the median mTurker earned 36% below the minimum wage. In the same study, the median Indian mTurker earned $1.65 per hour, and the median Crowdflower worker earned just $1.00 per hour. Wages can presumably be kept low on these platforms because they are highly deregulated and because they enjoy a labor surplus. Over 90% of the microworkers in Boyd's survey indicated that they would like to be doing more work. In addition, workers are disempowered on these platforms, with over half reporting that at least some of their work product went uncompensated without justification. These findings have recently been reproduced by Hara et al. (2018), who found that only 4% of mTurkers earn above the US federal minimum wage.

It would not be unreasonable to compare the crowd-sourced work done as part of the enrichment of NLP datasets to sweatshop labor, with the main difference being that sweatshop workers have traditionally been physically co-present in the same location and legal jurisdiction as one another, which facilitates labor organizing. By contrast, microworkers are isolated from one another both physically and across national and legal lines (though there are some efforts to

share information and even organize via Turkopticon[6]), making it more difficult for them to organize. NLP researchers seeking to enrich their datasets via crowd-sourced microwork need to acknowledge the asymmetric power relations involved in such employment and take responsibility for the ways in which they treat and interact with the laborers on whom their research and tools essentially depend.

Pitfalls of data analysis

In recent years, computer scientists have introduced a plethora of statistical models to capture semantic layer of documents. Topic models are among the most promising category of such modelling approaches. They are "algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes" (Blei et al. 2010). The most prominent variation of topic models is called latent dirichlet allocation (LDA). The paper that introduced this method, by Blei et al. (2003) has so far received more than 28,000 citations. Although the promise of LDA is to discover the main themes in a corpus of documents, what LDA delivers is far from actual topics of a text corpus. In fact, it is difficult to draw insights from LDA outputs. The "topics" are a set of words that can only be interpreted by researchers with deep domain knowledge and expertise (Barhate 2018). Additionally, even in the case of topic interpretation, there is no guarantee that the topics unearthed by LDA are the only topics discussed in the corpus. In recent years, many studies have been done to solidify the outcomes of LDA and make it interpretable. Several evaluation methods (Chang et al. 2009) and visualization techniques

[6] See url = < https://turkopticon.ucsd.edu/ >.

(Sievert & Shirley 2014) have been proposed; however, LDA results still suffer from poor interpretability and lack of comprehensiveness.

This methodological challenge has ethical implications. As we mentioned in the context of rumor classification above, in the near future legislators, regulators, and others may start to base policy and decision making on the theoretical claims attached to the outputs of NLP models. Indeed, a recent paper in *Nature* recommended that materials science research programmes should be directed by unsupervised learning on the corpus of published papers (Tshitoyan et al, 2019). It's not hard to imagine that some NLP enthusiast will soon make similar calls for the fields of psychology, political science, and so on. When the theoretical claims associated with NLP research are not robustly supported, basing policy and decision making on them would be reckless.

In addition, there is a danger that ham-fisted employment of NLP methods may elide the descriptive-normative distinction. NLP is capable of estimating the presence and strength of various attitudes and traits. But knowing what attitudes and traits someone embodies tells us nothing about whether those attitudes and traits are good or bad, virtuous or vicious. To put a finer point on it, Caliskan, Bryson, and Narayanan (2017) showed that semantics derived from natural language corpora are guaranteed to contain the same biases as the humans who produced the documents that make up those corpora. This suggests that NLP is well-suited to studying and portraying humanity as it is, warts and all, but that it is not particularly well-suited to studying and portraying humanity as it should be. For that, we would need a corpus of documents written by saints.

For example, if an algorithm is trained to select the CVs of job applicants who are likely to make good leaders in a corporate setting, and if the corpus on which that algorithm is trained is the set of CVs of all employees in the company ordered by rank and by frequency of promotion, then that algorithm is guaranteed to be biased in precisely the same way that previous managers and executives at that company were biased. Indeed, Amazon recently decided to shut down its automated vetting of job applicants because the algorithm they were using was systematically biased against women.[7] And an audit of a different hiring algorithm recently found that the two factors that it considered most predictive of success on the job were being named 'Jared' and having played lacrosse in high school.[8] While these examples might seem shocking, they should be completely unsurprising. If a hiring algorithm is trained on historical data, it will end up recommending CVs that are similar to the most highly-rated CVs in the corpus. If those CVs are in turn associated with mediocre men who were hired and promoted by implicitly or explicitly sexist and racist bosses, then the algorithm will end up recommending CVs that are similar to the CVs of the sort of people who tend to get selected and elevated by implicitly or explicitly sexist and racist bosses. This problem mirrors to some extent the

---

[7] See url = < https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G >, accessed 2 October 2019.

[8] See url = < https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased/ >, accessed 2 October 2019.

methodological problem with LDA. A corpus can be analyzed into topics, but those topics need a label; a set of CVs can be divided into groups, but those groups need a label. As NLP is currently practiced, there are almost no constraints on how labels are chosen. In the hiring case, the same set of CVs might be labeled by some as 'leadership material' and by others as 'beneficiaries of the racist patriarchy'. What one ends up *doing* with the classifier will presumably be very different depending on which label is chosen, which means that the seemingly neutral choice of a label has significant moral implications -- implications that simply cannot be ignored as "someone else's department" by psychometric NLP researchers.

A distinct ethical pitfall at the stage of data analysis is the potential to infer protected characteristics from unprotected characteristics. What counts as a protected characteristic differs from country to country. In the United Kingdom, there are nine: age, disability status, gender reassignment status, marriage, pregnancy and maternity (though oddly not paternity), race, religion or belief, sex, and sexual orientation. In the European Union, protected characteristics include sex, race, color, ethnic or social origin, genetic features, language(s) spoken, religion or belief, political opinion, membership in a national minority, property, birth, disability, age, and sexual orientation. In the United States, protected characteristics include race, religion, national origin, age, sex, pregnancy, familial satus, disability, veteran status, and genetic information. Many other countries specific similar, if non-overlapping, lists of protected characteristics. These agent-level characteristics are "protected" in the sense that it is illegal to discriminate against someone on their basis. Such discrimination is possible only when the person or agency that intends to discriminate is capable of classifying people on the basis of their protected

characteristics, so one natural way to protect against discrimination is to erase or mask datapoints that explicitly refer to protected characteristics in datasets that inform decision making.

However, NLP often makes it possible to infer or use proxies for protected characteristics from unprotected characteristics. For instance, it may be possible to infer someone's race from the way they spell certain words or their forename and surname. Likewise, it may be possible to infer someone's age from their forename. Again, it may be possible to infer whether someone has had their gender reassigned from their longitudinal digital footprint, which would include any name change from a masculine forename to a feminine forename (or vice versa). In addition, more subtle cues in people's public use of language may be used to make inferences about their mental health.

These concerns are not purely speculative. In a recent review article, Hinds & Joinson (2018) analyzed 327 studies that attempted to infer at least one demographic characteristic. Fourteen different characteristics were successfully inferred from digital footprints, including many protected characteristics such as gender, age, location, political orientation, ethnicity, race, familial relationships, language(s) spoken, health, religion, and sexual orientation.

Beyond these demographic characteristics, mental health characteristics have become a major focus of research (and presumably industrial targeted advertising). For example, Cook et al. (2016) found that responses to unstructured text prompts given to adults who had recently been discharged from psychiatric inpatient or emergency room settings predicted suicidal ideation. In a similar vein, Coppersmith, Dredze, & Harman (2014) used Twitter data to infer psychiatric disorders such as post-traumatic stress disorder (PTSD), depression, bipolar disorder,

and seasonal affective disorder (SAD). And Calvo et al. (2017) explore a range of NLP techniques that have already been used not only to make automated mental health diagnoses but also to target personalized, automated psychiatric interventions.

While we do not doubt the sincerity or good intentions of the researchers behind these and similar initiatives, we worry that the radical implications of this work are not fully appreciated even by those who do it. If this kind of analysis proves accurate, reliable, and precise, then all personal records are essentially demographic, health, and mental health records. Psychometric NLP researchers need to bear this in mind as they proceed with their work.

Pitfalls of data storage and sharing

Above, we emphasized the importance of anonymizing or pseudonymizing names and other identifying features. This is obviously important when the corpus includes text that was never publicly available. However, we also pointed out that even when speaking publicly, people may have an expectation of obscurity. To this end, it is useful to follow Schwartz & Solove (2014) in distinguishing three categories of datasets: identified, identifiable, and non-identifiable. In an identified dataset, the identities of the speakers or authors of the text are explicitly included (e.g., in a column labeled “author name” or “tweep”). In a non-identifiable dataset, these details are not included or are replaced with pseudonyms. Moreover, non-identifiable datasets do not provide sufficient information to infer back to the identities of individuals represented in the data.

The tricky middle case is the identifiable dataset. In such a dataset, identifying details are not included or are replaced with pseudonyms, but it is possible to infer the identity of the speaker or author using the rest of the dataset. Verbatim quotations are typically sufficient for this purpose, especially if they are longer than a single word or bigram. So too are sufficiently rich *patterns*. For example, there may be only one user of Facebook who lists Des Moines as their residence, tends to post at least three times per day, posts about both capybaras and the incel movement, and never posts between the hours of 2 PM and 3 PM local time. Even if no verbatim quotations or directly identifying information is provided about such an individual, knowing enough about them may make it possible to rule out enough other potential candidates that only a singleton remains. The rows in such a dataset are definite descriptions in disguise, essentially saying, "There exists one and only individual x, such that x is F_1, F_2, …, and F_n."

This sort of example represents an irresoluble tension in NLP and related approaches to big social data: rich datasets involving complex patterns are, from a pure research perspective, almost always better and more interesting, but the richer the dataset, the more likely it is to be an identifiable dataset *even when researchers make serious efforts to render it non-identifiable*. Open science (especially open data) is thus in tension with privacy and security concerns.

This is not an idle or hypothetical concern. There have already been multiple cases in which datasets that had been presumed non-identifiable turned out to be identifiable. For instance, Narayanan & Shmatikov (2008) showed that it was possible to deanonymize the dataset associated with a prize competition sponsored by Netflix. Likewise, de Montjoye et al. (2015) showed that it is possible to deanonymize an allegedly non-identifiable dataset of credit card metadata.

Researchers must keep this fact in mind when designing their studies and disseminating their results and materials. As no strict and clear rule seems feasible, the tradeoff should be carefully thought through each time a study is planned. In some cases, the researchers may decide that it is too risky to follow what would otherwise be best practices in open data. In such cases, they may still share a subset of their data and keep identifying and identifiable data to themselves on a secure server. However, even then there is always the risk that the supposedly secure server could be hacked, thereby exposing data subjects' identities to the hackers. In addition, a dataset that is, on its own, non-identifiable may become identifiable when combined with other publicly available datasets. In fact, this was how Sweeney et al. (2013) managed to deanonymize the Personal Genome Project dataset at Harvard University.

Pitfalls of applications, tools, and interventions: Feedback loops and the spectre of dual use

Because they can be automated and operated at scale, applications and tools that rely on psychometric NLP are often deployed in interactive online interfaces, such as recommender systems and chatbots. This creates a feedback loop in which *what is measured* by psychometric NLP (people's traits and attitudes) is the very same thing as *what is affected* by the applications and tools into which psychometric NLP is fed. This feedback loop has the potential to cause Heisenberg effects, where the object of study changes in virtue of the fact that it is being studied. As we have seen in some of the other pitfalls canvassed above, this leads to both methodological and moral challenges.

Methodologically, researchers tend to assume that the object of study is static, or at least that its dynamics are not influenced by the researcher themselves. Morally, researchers need to

be wary of causing transformative experiences that the user has not consented to and which may even make them morally worse. A transformative experience is an experience that deeply changes a person's values, preferences, or worldview (Paul 2014). Standard examples include getting married, having a baby, going to college, and immigrating to a new country. But transformative experiences can also change us for the worse. Alfano, Carter, & Cheong (2018) explore radicalization and self-radicalization under the rubric of transformative experience. They argue that ongoing interactive feedback loops between people and the systems that both measure and influence their attitudes, such as the YouTube recommender system, may lead to self-radicalization, which they call "technological seduction." In a nutshell, the idea is that someone's digital footprint might be used to estimate with probability, say, .72 that they are sympathetic to the alt right. They would then be recommended videos that people who support the alt right tend to watch, which could influence their political attitudes, making them more extreme. This in turn could lead to their receiving more recommendations for alt right videos and even for neo-Nazi propaganda. This example refers to the YouTube recommender system, which uses techniques and datasets beyond (but also including) NLP, the structure of the transformative experience would be the same in a pure NLP case. Researchers who build or make available models that inform applications and tools that have the potential to cause such transformative experience need to be wary of this problem.

Whereas the transformative experiences described above involve unintentional harms caused by the negligent or reckless use of psychometric NLP, there is also the potential for intentional harms and abuses that arise from dual use. A piece of scientific research or a technological artifact is considered to have "dual use" or to be "dual-use" when it can be used to

pursue at least two seemingly contrary purposes. Scholars disagree about exactly how to characterize the opposition, but the contrast is usually thought to involve either military and civilian purposes or beneficial and harmful purposes where the potential for harm is severe and large-scale, as in the case of nuclear or biological research informing the construction and use of weapons of mass destruction (Miller 2018). For example, Albert Einstein's research on atomic physics made possible both civilian (peaceful) nuclear energy production and the atom bomb. There continues to be controversy about whether dual-use research should face additional oversight, regulation, or even censorship (Miller & Selgelid 2008).

Historically, debates about dual use have been primarily concerned with kinetic attacks such as nuclear bombs, chemical weapons, and weaponized bacteria or viruses. While kinetic attacks remain a pressing problem, in the twenty-first century we must also consider cases involving cyber attacks. As the Cambridge Analytica-Trump scandal shows, NLP methods and tools developed for innocent, civilian research purposes (personality assessment) can be repurposed by state and non-state actors to bolster domestic psyops and international electoral interference. While more covert and less dramatic than a nuclear attack, a cyber attack can do a great deal of harm, as the last several years have amply demonstrated. Perhaps psychologists and computer scientists have assumed that their fields are not implicated in and partially responsible for the dual use of their research, but the time has come to admit that that was wishful thinking (Hovy & Spruit 2016).

Covert psyops and electoral interference are not the only potential dual use of NLP research in psychology. Perhaps even more worrisome are overt dual use by corrupt and authoritarian regimes, as well as powerful non-state actors. For example, it's not hard to imagine

that these methods and tools could be (indeed, probably already are) used to create and curate lists of enemies and targets. As we mentioned above, NLP and related methods can be used to infer a range of sensitive characteristics, including sexual orientation, ethnicity, religious views, political views, and use of addictive substances. If governments have backdoor access to the profiles of users of social platforms (as China currently does with WeChat and other platforms[9]) or require visitors or immigrants to turn over their social media profiles when applying for a visa or entering the country (as the United States currently does[10]), it would be a trivial exercise to harness NLP and related tools to classify individuals as, in essence, enemies of the state. In a recent case, a Palestinian student who was about to start his studies at Harvard University was deported after immigration officials examined *his friends'* social media footprints and found posts with political content critical of the United States.[11] During the second half of the twentieth century, the East German Stasi employed nearly 100,000 full-time domestic spies, who sought and analyzed reports from hundreds of thousands of informants, as well as various analog and digital surveillance tools (Borneman 1991). The Stasi's annual budget is estimated to have been

---

[9] See url = < https://www.theepochtimes.com/security-risks-exist-in-chinese-mobile-apps_2811727.html > for details.

[10] See url = < https://www.brennancenter.org/analysis/timeline-social-media-monitoring-vetting-department-homeland-security-and-state-department > for an up-to-date timeline of ICE and USCIS policies.

[11] See url = < https://www.thecrimson.com/article/2019/8/27/incoming-freshman-deported/ > for details.

the equivalent of two billion dollars per year in today's money. Using NLP, this kind of ongoing surveillance and enemies list-curation could be automated relatively cheaply and at scale. WeChat alone has over a billion users whose communications can be analyzed and classified automatically. Saudi Arabia, which has purchased sophisticated cyber-surveillance tools from firms staffed by former Israeli Defense Force personnel (Abramson 2019), may have the capability to classify its citizens (whether resident or expat) as, among other things, dissidents or homosexuals. The Saudi domestic torture campaign recently instigated by Mohammad Bin Salman Al Saud (colloquially known as 'MBS') may very well be informed and directed by these tools. To clarify the stakes: homosexual acts are currently criminalized as sodomy in Saudi Arabia, a violation that is punishable by stoning.

Conclusion

In this chapter, we canvassed a range of ethical and political pitfalls that psychological researchers using NLP are liable to encounter. These pitfalls occur at all stages of the research pipeline, from data acquisition to data enrichment, analysis, storage and sharing. In addition, the results obtained by such research can be put to legitimate civilian uses, but they can also be harnessed for dual use by authoritarian governments and non-state actors. Because these issues are so controversial and complex, and because the field is developing so rapidly, it is difficult to establish a binding code of conduct that could and should legitimately govern research in this domain. Instead, we have focused on elaborating the many problems that can arise and pointing to generic concerns that researchers should bear in mind -- if only in the backs of their heads -- as they plan, execute, and disseminate their research. We hope that further attention to this topic

will lead to a more systematic understanding of all of the ethical problems that might be encountered, as well as a set of unified normative principles suitable for the governance of this domain.

References

Abramson, S. (2019). *Proof of Conspiracy: How Trump's International Collusion is Threatening American Democracy*. Simon & Schuster.

Alfano, M., Carter, J. A., & Cheong, M. (2018). Technological seduction and self-radicalization. *Journal of the American Philosophical Association*, 4(3): 298-322.

Alfano, M., Higgins, A., & Levernier, J. (2018). Identifying virtues and values through obituary data-mining. *Journal of Value Inquiry*, 52(1): 59-71.

Allport, G. & Odbert, H. (1936). Trait-names: a psycho-lexical study. *Psychological Monographs*, 47(1): i-171.

Ashton, M., Lee, K., Perugini, M., Szarota, P., de Vries, R., Di Blas, L., Boies, K., De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology*, 86(2): 456-66.

Azucar, D., Marengo, D., & Settani, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, 124: 150-59.

Barhate, P. (2018). Latent Dirichlet Allocation for Beginners: A high level intuition. Medium. Url = <https://medium.com/@pratikbarhate/latent-dirichlet-allocation-for-beginners-a-high-level-intuition-23f8a5cbad71 >, accessed 29 September 2019.

Berg, J. (2016). Income security in the on-demand economy: Findings and policy lessons from survey of crowdworkers. *Comparative Labor Law Policy Journal*, 37(3): 543-76.

Blei, D.M., Carin, L & Dunson, D. (2010). Probabilistic Topic Models. *IEEE Signal Processing Magazine*, *27*(6), 55-65.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Borneman, J. (1991). *After the Wall: East Meets West in the New Berlin*. Basic Books.

Caliskan, A., Bryson, J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183-6.

Calvo, R., Milne, D., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5): 649-85.

Chang, A. (2018, May 2). The Facebook and Cambridge Analytica scandal, explained with a simple diagram. *Vox*. url = < https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram >. Accessed 22 September 2019.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, 288-296.

Christen, M., Alfano, M., & Robinson, B. (2017). A cross-cultural assessment of the semantic dimensions of intellectual humility. *AI & Society*.

Cook, B., Progovac, A., Chen, P., Mullin, B., Hou, S., & Baca-Garcia, E. (2016). Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms

in a text-based mental health intervention in Madrid. *Computational and Mathematical Methods in Medicine*. Article ID 8708434.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. URL = < https://www.aclweb.org/anthology/W14-3207/ >.

Corpus of Contemporary American English. (2019). URL = < https://www.english-corpora.org/coca/ >, accessed 28 September 2019.

Corpus of Founding Era American English, v. 3.00. (2019). URL = < https://lcl.byu.edu/projects/cofea/ >, accessed 28 September 2019.

Corpus of Historical American English. (2019). URL = < https://www.english-corpora.org/coha/ >, accessed 28 September 2019.

Cox, J. & Koebler, J. (2019, April 25). Why won't Twitter treat white supremacy like ISIS? Because it would mean banning some Republican politicians too. *Vice*. URL = < https://www.vice.com/en_us/article/a3xgq5/why-wont-twitter-treat-white-supremacy-like-isis-because-it-would-mean-banning-some-republican-politicians-too >, accessed 1 October 2019.

Daly, A., Devitt, K., & Mann, M. (2019). *Good Data*. Institute of Network Cultures. Url = < https://eprints.qut.edu.au/125605/1/Good_Data_book.pdf >. Accessed 22 September 2019.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, 512-15.

de Montjoye, Y.-A., Radaelli, L., Singh, V. K., & Pentland, A. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221): 536-9.

DiFonzo, N., & Bordia, P. (2007). *Rumor psychology: Social and organizational approaches*. American Psychological Association.

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J., & Johnson, B. (2018). The tactics & tropes of the Internet Research Agency. Url = < https://disinformationreport.blob.core.windows.net/disinformation-report/NewKnowledge-Disinformation-Report-Whitepaper.pdf >. Accessed 24 September 2019.

Douglas, D. (2016). Doxing: A conceptual analysis. *Ethics and Information Technology*, 18(3): 199-210.

Ebersole, C., Atherton, O., Belanger, A., Skulborstad, H., Allen, J., Banks, J., Baranski, E., Bernstein, M., Bonfiglio, D., Boucher, L., Brown, E., Budiman, N., Cairo, A., Capaldi, C., Chartier, C., Chung, J., Cicero, D., Coleman, J., Conway, J., Davis, W., Devos, T., Fletcher, German, K., Grahe, J., Hermann, A., Hicks, J., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D., Joy-Gaba, J., Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, J., Klein, R., Lucas, R., Lustgraaf, C., Martin, D., Menon, M., Metzger, M., Moloney, J., Morse, P., Prislin, R., Razza, T., Re, D., Rule, N., Sacco, D., Sauerberger, K., Shrider, E., Shultz, M, Siemsen, C., Sobocko, K., Sternglanz, R. W., Summerville, A., Tskhay, K., van Allen, Z., Vaughn, L., Walker, R., Weinberg, A., Wilson, J., Wirth,

J., Wortman, J., Nosek, B. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67: 68-82.

Fard, A. E., Mohammadi, M., Chen, Y., & Van de Walle, B. (2019). *Computational Rumor Detection Without Non-Rumor: A One-Class Classification Approach*. IEEE Transactions on Computational Social Systems.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Galton, F. (1884). Measurement of character. *Fortnightly Review*, 36: 179-85.

Gibney, E. (2018). The scant science behind Cambridge Analytica's controversial marketing techniques. *Nature*.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, paper #499.

Hall Jamieson, K. (2018). *Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What we Don't, Can't, and Do Know.* Oxford University Press.

Helgesson, G. & Johnsson, L. (2005). The right to withdraw consent to research on biobank samples. *Medicine, Health Care and Philosophy*, 8(3): 315-321.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3): 61-83.

Hinds, J. & Joinson, A. (2018). What demographic attributes do our digital footprints reveal? A systematic review. *PLoS ONE*. url = < https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207112 >.

Hirsh, J., Kang, S., & Bodenhausen, G. (2012). Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological Science*, 23(6): 578-81.

Hoover, J., Dehghani, M., Johnson, K., Iliev, R., & Graham, J. (2018). Into the wild: Big data analytics in moral psychology. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 525-536). New York, NY, US: The Guilford Press.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T., Chin, J., Leong, C., Leung, J. Y., Mirinjian, A., & Dehghani, M. (2019, April 10). Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. https://doi.org/10.31234/osf.io/w4f72

Hovy, D. & Spruit, S. (2016). The social impact of natural language processing. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591-8.

Intemann, K. (2010). 25 years of feminist empiricism and standpoint theory: Where are we now? *Hypatia*, 25(4): 778-96.

Klein, R., Ratliff, K., Vianello, M., Adams, R., Bahník, Š, Bernstein, M., Bocian, K., Brandt, M., Brooks, B., Brumagh, C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E., Hasselman, F., Hicks, J., Hovermale, J., Hunt, S. J., Huntsinger, J., IJzerman, H., John, M.-S., Joy-Gaba, J.,

Kappes, H., Krueger, L., Kurtz, J., Levitan, C., Mallett, R., Morris, W., Nelson, A., Nier, J., Packard, G., Pilati, R., Rutchick, A., Schmidt, K., Skorinko, J., Smith, R., Steiner, T., Storbeck, J., van Swol, L., Thompson, D., van 't Veer, A., Vaughn, L., Vranka, M., Wichman, A., Woodzicka, J., & Nosek, B. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, 45(3): 142-52.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J., Bahník, Š., Berkics, M., Bernstein, M., Bialobrzeska, O., Bocian, K., Brandt, M., Cantarero, K., Cemalcilar, Z., Cicero, D., Chandler, J., Chatard, A., Chen, E., Cheong, W., Coen, S., Collisson, B., Conway, J., Corker, K., Curran, P., Cushman, F., Dalla Rose, A., Davis, W., Devos, T., Dogulu, C., Dunham, Y., Eller, A., Finck, C., Friedman, M., Giessner, S., Gnambs, T., Gómez, Á., Graham, J., Grahe, J., Green, E., Haigh, M., Haines, E., Heffernan, M., Hicks, J., Houdek, P., Huntsinger, J., IJzerman, H., Inbar, Y., Kende, A., Innes-Ker, Å., Jiménez-Leal, W., Kappes, H., Karabati, S., Keller, V., Kervyn, N., Krueger, L., Lakens, D., Lazarević, L., Levitan, C., Lins, S., John, M.-S., Mallett, R., Milfont, T., Morris, W., Myachykov, A., Neave, N., Nichols, A., O'Donnell, S., Orosz, G., Pérez-Sánchez, R., Petrovic, B., Pilati, R., Pollmann, M., Salomon, E., Schmidt, K., Sekerdej, M., Smith, M., Smith-Castro, V., Sobkow, A., Stouten, J., Street, C., Traczyk, J., Torres, D., Theriault, J., Ujhelyi, A., van Aert, R., van Assen, M., Vaughn, L., Vázquez, A., Verniers, C., Verschoor, M., Vranka, M., Wichman, A., Williams, L., Young, L., Zelenski, J., Nosek, B. (2017). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychology Science* (RRR).

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15): 5802-5.

Kwon, S., Cha, M., & Jung, K. (2017). *Rumor detection over varying time windows*. PloS one, 12(1).

Lee, T. & Phillips, J. (2019). Data-driven originalism. *University of Pennsylvania Law Review*, 167(2): 261-335.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. (2017): Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America*, 114(48): 12714-9.

McConnell, T. (2010). The inalienable right to withdraw from research. *Journal of Law, Medicine and Ethics*, 38(4): 840-6.

Miller, S. (2018). *Dual Use Science and Technology, Ethics and Weapons of Mass Destruction*. Springer.

Miller, S. & Selgelid, M. (2008). *Ethical and Philosophical Consideration of the Dual-Use Dilemma in the Biological Sciences*. Springer.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Url = < https://arxiv.org/abs/1810.03993 >, accessed 28 September 2019.

Paul, L. A. (2014). *Transformative Experience*. Oxford University Press.

Peabody, D. & Goldberg, L. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of Personality and Social Psychology*, 57(3): 552-67.

Pennebaker, J. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury.

Pfeffer, J., Mayer, K., & Morstatter, F. (2018). Tampering with Twitter's sample API. *EPJ Data Science*, 7(50): 1-21.

Rauthmann, J., Gallardo-Pujol, D., Guillaume, E., Todd, E., Nave, C., Sherman, R., Ziegler, M., Jones, A., & Funder, D. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situational characteristics. *Journal of Personality and Social Psychology*, 107(4): 677-718.

Richards vs. Cox. UT 57 (2019).

Saucier, G. (1997). Effects of variable selection on the factor structure of person descriptors. *Journal of Personality and Social Psychology*, 73(6): 1298-1312.

Scalia, A. (1997). *A Matter of Interpretation: Federal Courts and the Law*. Princeton University Press.

Schwartz, P. & Solove, D. (2014). Reconciling personal information in the United States and the European Union. *California Law Review*, 102: 877-916.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 63-70.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11): 1359-66.

Sweeney, L., Abu, A. & Winn, J. (2013). Identifying participants in the personal genome project by name. Harvard University. Data Privacy Lab. White paper 1021-1. Url = < https://dataprivacylab.org/projects/pgp/ >, accessed 28 September 2019.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571: 95-99.

Waseem, Z. (2016). Are you racist or am I seeing things? Annotator influence on hate speech detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138-42.

Waseem, Z. & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the NAACL student research workshop*, 88-93.

Wong, J. C. (2019, 18 September). The viral selfie app ImageNet Roulette seemed fun -- until it called me a racist slur. *The Guardian*. URL = < https://www.theguardian.com/technology/2019/sep/17/imagenet-roulette-asian-racist-slur-selfie >, accessed 26 September 2019.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3).