CONDUCTING SOCIALLY RESPONSIBLE AND ETHICAL COUNTER INFLUENCE OPERATIONS RESEARCH

A PRACTICAL GUIDE FOR RESEARCHERS AND PRACTITIONERS

Andrew Maynard[*], Cassian Corey[†], Amna Greaves[†], Mark Kozar[†], K. Hazel Kwon[*], Marissa Scragg[*]

[†]MIT Lincoln Laboratory

[*]Arizona State University

A collaboration between MIT/Lincoln Laboratory, the ASU Global Security Initiative, and the ASU Risk Innovation Laboratory

January 29, 2022

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

ASU Risk Innovation Lab
Arizona State University

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In today's world, domestic and foreign Influence Operations (IO) campaigns of misinformation and disinformation are increasingly being channeled through digital and social media platforms to divide, deceive, and create unrest. Effective approaches to responding to IO campaigns depend on targeted research into novel and agile methods and techniques for countering them. Yet research into Counter Influence Operations (CIO) requires a sophisticated understanding of the associated landscape around ethical and responsible research if it is to avoid being stymied by ethical mis-steps, and to be effective and impactful. The US Government recognizes the challenges of Grey Zone Warfare (GZW) and the need to counter the effects of IO. However, there are also institutional challenges to conducting research in this domain. With the appropriate guidance on socially responsible and ethical research, academic and government research institutions are uniquely positioned to conduct unbiased, innovative and actionable research and development in support of effective CIO. This position paper specifically addresses the need for such guidelines.

Researchers in academic and government research institutions in particular will find the ethics discussion herein useful in the novel application of emerging capabilities such as Artificial Intelligence and Machine Learning (AI/ML) to online CIO. For instance, countering adversarial IO using the same tactics and approaches as adversaries may come with social and ethical risks that undermine the utility of such approaches. And well-meaning but naïve research may risk losing the trust of key stakeholders, and thus its ability to be impactful. Core to the approaches taken here is the Risk Innovation Framework developed in the Arizona State University Risk Innovation Lab, and the careful and continued consideration of stakeholders, risks, and ethical principles at all stages of the design, development, deployment, test, and evaluation of a countermeasure.

To examine the ethical application of CIO methods in-the-wild, this position paper begins by describing *The Need for Strategic Counter Influence Operations Research* in chapter 1. We then map out the stakeholder and CIO landscapes in chapter 2, and explore the ethics and risk landscape around CIO research in chapter 3 – especially with respect to social media-based research. Chapter 4 addresses how learning from the ethical application of AI/ML applies to CIO research, while chapter 5 provides a broad overview of Risk Innovation and its application to responsible and ethical CIO research. Chapter 6 is devoted to examining institutional risk associated with CIO research, while chapter 7 develops a framework for the practical application of Risk Innovation to CIO research. The remaining chapters (8-11) provide a unique set of resources for practitioners designing, conducting and applying CIO research.

Intended Audience

Disinformation as a means to influence public opinion touches all our lives. The art of developing capabilities to detect, measure, and counter the deliberate and malign manipulation of public opinion depends heavily on both the social and technical sciences. Research has demonstrated that exposure to contrary beliefs can lead to further radicalization. This means that just the presentation of logical arguments and fact checked data may not be the most effective way to counter a popular and widely held false fact. So how can countermeasures be developed? For technologists, potential solutions require three things: access to data, access to compute resources, and a vector by which to approach the problem. For social scientists, solutions must be practical, ethical, and most of all, remediable when the unforeseen harm occurs. Researchers, policy makers, and technology directors need to consider risks and unintended biases which can lead to harm when deploying Artificial Intelligence in the fight against influence operations. While not comprehensive, this paper provides a structured way to understand, evaluate, and build a foundation to address the risks involved in fighting disinformation.

1. The Need for Strategic Counter Influence Operations Research

There are distinct differences between peacetime and open warfare. In peacetime, alliances exist and competition is guided by agreed upon rules. In contrast, open warfare can range from non-state and state-based violent conflict, to total war. In between these two extremes, Grey Zone (GZ) warfare operates “below the threshold of open military conflict and at the edge of international law” (The White House 2017). GZ activities are deliberate, planned, and meant to achieve political interests over a period of time, rather than decisively as in open warfare. Such activities can be intended to project strength or dominate a geographical area through political and military actions that utilize proxy militias, military exercises, territorial annexation, space and cyber aggression, physical barriers, and Influence Operations (IO) (Hicks, Friend et al. 2019).

Cyber election meddling, for example, can utilize on several techniques, including cyber tampering operations access a State’s election infrastructure, altering vote tallies and registration databases to change election outcomes, and preventing voters from casting their vote. The goal of cyber IO is to influence attitudes, behaviors, and decisions of a targeted audience. Conceptually, cyber IO can be divided into doxing and IO (e.g., “malinformation” and disinformation)[1]. In doxing operations an adversary gains access to a computer system or digital service to exfiltrate non-public data with the intention of leaking it to the public. Malinformation operations may rely on trolling where an adversary conducts threatening, abusive, discriminatory, harassing and disruptive online behavior with malicious intent, or they may involve the intentional use of fake information to make a position or false “truth” more believable. Disinformation operations are the spread of partially or completely false information for economic gain or to intentionally deceive. These techniques are often employed together to dismiss, distort, distract, and dismay believers through the use of data dumps, mass media, and social media platforms. More broadly, these techniques erode trust in institutions, delegitimize democracy and political systems, and incite unrest (Sanders 2019).

The U.S. and its Allies deter these adversarial challenges with positive narratives of policy, aims, and objectives with counter information as a national defense strategy. These Counter Influence Operations (CIO) can assure or solidify relationships with allies, promote the legitimacy of political systems, and prevent or contain unrest. Although IO is well documented, today’s battlespace increasingly occurs in the digital information environment, where adversarial information can be spread and amplified at the speed of cyber infrastructure and social networks.

Online mobilization and calls to arms as part of an emerging landscape around IOs present an acute law enforcement and national security challenge, and one that must be addressed. In the past, domestic terrorists have often been lone actors whose ideologies intersect with conspiracy theories, misinformation, and disinformation. Today, this information is often channeled through

[1] Malinformation is popularly defined as being based on factual events but presented in a manner to deliberately inflame social divisions; this differs from misinformation and disinformation which are based on false information (in the latter case to cause intentional harm, versus the former which causes unintentional harm.)

social media platforms to a massive audience with the intention to undermine and erode the fabric of society (Executive Office of the President National Security Council 2021). These same platforms, whose business goal is to increase usership and online time, have little motivation to implement solutions (Lima 2021).

The potential impacts of IO on substantial populations requires innovative approaches to CIO and necessitates new research into methods and mechanisms that enable IO – and especially social media-based IO – to be effectively countered. However, the very nature of IO and CIO raises complex questions around ethical research into CIO. For example, adversarial IO may be amplified through divisive messaging using Artificial Intelligence (AI) and Machine Learning (ML) technologies, such as online Bots (Marcellino, Magnuson et al. 2020). AI/ML is an active field of study with much promise, but generated models and applications are only as good as the data they are trained on, and may have or introduce biases of their own. This becomes important when IO and CIO research involves human subjects and has the potential to adversely impact them. These challenges are amplified when using social media as a research platform, where the norms and expectations around ethical and responsible research practices are still being developed. Here, there are dangers of not only inadvertently causing harm to research subjects, but placing researchers and their institutions in jeopardy if key stakeholders perceive ethical lines to have been overstepped.

This paper is developed primarily for US-based academic and research community audiences, but may be of use to commercial and non-profit organizations looking to “fight the good fight” against harmful false information. Readers should consider the special circumstances under which they themselves conduct their research, including how to harmonize the guidance and use cases outlined for their own institutional, state, and federal or legal guidelines. The authors’ intent is to enable and support the development of countermeasures and counter-technologies to combat false information, regardless of source or intent. Very specifically we hope that, by sharing the outputs from our exploration of this emerging field of study, the pace of collaboration and development of solutions to the threat of misinformation and disinformation may be accelerated globally.

2. Mapping out the Counter IO Stakeholder Landscape

IO actors usually have political motivations at all levels of domestic and foreign government, and may include special interest groups, individuals funded at the grass-roots level, and unofficial extensions of State entities. Foreign governments have increasingly launched efforts to discredit democracy and sow division that play on a myriad of biases as examples of government mistrust or overreach (Powers and Kounalakis 2017), including migration, arms ownership rights, and vaccine hesitancy. Unfortunately, these examples have also proven to be lightning rods for daily media discourse and lucrative funding opportunities for political gain that domestic government leaders continue to echo. With little incentive for change, even social media platforms benefit from heavy user communications traffic without intervention (Bauder, Liedtke et al. 2021). Whether actors are domestic or foreign, citizens are covered by rights and guided by social norms and responsibilities.

Clearly, the range of unethical and criminal behavior is wide, far reaching, global, and at all levels of power where the reward continues to outweigh the risk. When planning interventions, CIO researchers need to clearly map out all stakeholders for a complete understanding of goals, motivations, benefits, opportunities, and challenges from multiple points-of-view. The following subsections offer a discussion of example stakeholders and ethical concerns.

Federal, State, and Local Government and Entities

State and federal entities, and by extrapolation most governmental organizations, generally have rules and regulations for assessing both risk and ethics of research enterprises, especially when they involve human-facing technologies and activities. A challenge for these entities is often perceptions around how rules are applied, or even interpreted. For research organizations, there are specific pitfalls in the belief that these overarching regulations cover the researcher and thus obviate the need for institutional and practitioner due diligence. As with most social and political structures, rules evolve based on popular sentiment, but vary widely in implementation (for example, the minimum drinking age in the United States is set by the federal government, but implementation at the state and even local level is governed by a wide range of exceptions). In addition, the time needed and standard of evidence required to modify or create new regulations are substantial. State and Federal rules for ethical behavior for CIO research, and research in general, should be viewed as a minimum baseline, the violation of which would likely have immediate legal and social repercussions.

Media Entities

The ethical standards and risks for media entities, especially social media, are currently being fought in the theater of public and political opinion. The complexity of the issues involved are far beyond the scope of this paper. However, several key points are worth highlighting here. First, the majority of media companies are for-profit. As such the desire to reduce harm is potentially in conflict with the potential to benefit financially from content which could cause harm, but at the same time draws (or keeps engaged) a larger audience. For the researcher, the same rules apply for Media ethics as for State and Federal Entities, meaning these are baseline rules, beyond which an organization might find itself at risk legally. End User License Agreements (EULA’s) and other License Agreements are meant to provide guardrails for the majority of users, and can force limitations in how experiments are devised or implemented. As such there are two quick solutions if your experiment (for example, the deployment of bots as part of an online study) may be unduly constrained by these licenses: utilizing professional survey firms, or creating a closed, social-media simulator. In the former, there is the trade between the cost to engage the firm, and often to pay for audience interaction and the ability to conduct an experiment in situ (people know they're being paid/surveyed and that can bias their responses). In the latter, there is the level of effort to create a user experience which engenders the same visceral reactions as actual social media platforms. It also provides a reusable framework for further experimentation and development. The benefits in both cases, however, are that the researcher is then free to observe effects and collect data (such as specific demographic, or biometric sensor data) that would be otherwise too challenging to attempt.

Second, there are ethical and risk concerns regarding the use of data created or provided by media corporations. For example, from early on, scientific research projects that either used Facebook user data or partnered with the company for large-scale online experiments have raised multifaceted ethical concerns, such as privacy (e.g. (Lewis, Kaufman et al. 2008)), disrespect of observed individuals' autonomy (e.g. (Bond, Fariss et al. 2012, Kramer, Guillory et al. 2014)) and data-driven political manipulation (e.g. (Berghel 2018)). In response to the recurrent outcries over (un)ethical research, Facebook has gradually shifted its data access policy from the Publicly Available Information (PAI) framework in early days (data made open to anyone through API programming) to a restrictive and controlled data access framework (e.g., Facebook's FORT system with prerequisites of administrative permission, legal paperwork, and private VPN setup to access the database). That being said, such a controlled access framework has posed other ethical challenges on the transparency and reproducibility of scientific findings, because researchers receive not raw data but processed data –which has been manipulated to some degree on the backstage by Facebook's in-house research team. Recent scandals surrounding Facebook's provision of flawed data to scientists (New York Times 2021) as well as the company's hindrance of some academic misinformation research (which may be subject to further discussions though) Edelson and McCoy (Edelson and McCoy 2021) demonstrate how researchers who use corporate-owned data can find themselves at odds with scientific integrity. CIO researchers need to be aware of these ethical dilemmas regarding the use of data provided by media companies.

Science and Education Entities

Government research laboratories and academic institutions have in common the fact that oversight and review boards are often part and parcel of research development. Unlike commercial institutions which are often free to act rapidly on new ideas and may only be constrained by directors or shareholders, the reasons for the acquisition of funding for an initiative within Government and academic institutions is often written down, discussed, examined by peers or sponsors, and challenged to identify their merit. However, with innovation comes new and sometimes unidentified risks. Often scientists may decline to fully examine the risks associated with a new concept because the impacts of the use (or misuse) of the innovation cannot be fully predicted, or because there are practical limitations to doing so.

A substantial challenge here is that recognition of individuals in this category is often tied to publications, presentations, and demonstrations as metrics of success and primary means of technology transfer. Unlike private (and some Government) organizations which may quietly patent their intellectual property, or set a classification level for its access, the greatest asset of public research institutions may also be their greatest risk in some cases.

Internet and Data Storage and Management Entities

Infrastructure and data management often go hand in hand. Large corporations who dominate the market in offering internet services are often the customer-facing organizations who sell space, servers, and even the customization of both. Although these entities devote considerable resources to security, the endless discovery of vulnerabilities in underlying infrastructure, together with insider-threats, mean that the risks of data breaches or loss are always looming. Ethical considerations are often considered the domain of top-level management, where decisions to do business with or otherwise work for governments or other organizations that are themselves ethically fraught raise visible concern. The potential harm (and resulting liability) from malicious attacks which result in data breaches are generally more of a financial and public image concern. Thus it falls to the data owner to ensure that data security goes beyond just encryption, that obfuscation of private information is baked into the data architecture.

Social and Tribal Entities

Tribal entities – defined here as communities and groups held together with a strong bond of common identity, belief and purpose – are perhaps some of the most impactful entities in our lives. They constitute our families, friends, colleagues, and peers, and often have a direct influence on our perspectives, beliefs, and actions. It is very difficult to separate the sense of whom we associate with and with whom we are – or would like to be – perceived as. While there may be a certain amount of protection within online communities, there is also additional risk as exposure can be very quick and uncontrolled. Posting or sharing something that might seem inconsequential could have overwhelming real-life results, catapulting the unfortunate into a wave of unwanted public intrusion and judgment.

While many of us attempt to separate our work and private lives, we cannot, and should not, forget the “tribes” that bind us with social and behavioral constraints. While it may appear to our employers that these are distractions which could result in the unnecessary delay of business, the understanding of the need for diversity and inclusion as part of both business and research models have been on the rise. This means that the risks (as perceived by your tribes) should be assessed, and the gaps in tribal and organizational ethical standards should be evaluated (because if you don’t, then who will?)

The Individual

As an individual with agency, the buck stops here. Ultimately, regardless of external influences and organizations, you generally have the choice to decide if the rules and frameworks which you are required to abide by are sufficient or not. Most people have an inbuilt set of benchmarks and metrics which prevent them from ethical lapses or unhealthy risk-taking. For instance, avoiding enticing vaporware, or knowing that what a sponsor intends is actually different than what is written down, are not uncommon circumstances – especially in an age where whistleblowing is increasingly common.

While the risks associated with a specific project or program might not have direct personal impact, if you can extrapolate how your invention could cause harm, you should spend time to at least raise awareness, within the bounds of what is safe for you. It's worth reflecting that, while it is highly unlikely that Alfred Nobel would have decided not to invent dynamite (his initial motivation being to improve the safety of miners who used highly volatile and dangerous substances such as nitroglycerine) he was sufficiently concerned with the impact his invention had on warfare to make the Nobel Prize his more lasting legacy.

Science will continue to move forward, and the argument is often "well if I don't invent it someone else, maybe someone with less inclination to evaluate the risks and ethics will" is nevertheless a valid one. The fundamental admonition here is to be aware, keep track, practice due diligence, take notes, and do your best to build in safety, security, and indications as to when something might be misused. This includes incorporating measures for transparency, planning for potential harm, and providing ways for this to be communicated and documented. As a rule, it is expedient to be open. Act responsibly and ethically, and most of all, don't be deterred.

3. Research Ethics and Risks: Understanding the Counter IO Landscape

The ethics landscape that IO and CIO research sits within is complex and evolving. There are relatively few requirements for conducting research in this domain. The landscape becomes even more complex when CIO uses the same platforms and similar strategies as IO.

There is a lack of clarity defining the ethical, responsible, and appropriate use of publicly available information. To further complicate matters, this uncertainty places researchers and research institutions in a vulnerable position where they risk overstepping what their stakeholders and other constituencies consider to be ethical boundaries.

To understand and navigate this landscape, it is first necessary to understand the broader landscape around ethical social media research with no underlying IO or CIO component, and how this informs IO and CIO research. Reviewing ethical social media research is particularly important as the scope and impact of IO and CIO have become unprecedentedly far-reaching through interactions with social media audiences. Some CIO researchers have specified this domain of activities by calling it ”hostile social manipulation” to underscore the importance of countermeasures against the spread of adversarial influence in social media spaces (Mazarr, Casey et al. 2019). Considering that such countermeasure effort often necessitates collection and use of social media data or human contacts with social media users, understanding ethical concerns surrounding social media research is necessary for IO and CIO researchers.

Using social media in research

The past two decades have seen growing interest in using publicly available online data for research – especially from social media platforms such as Twitter, Facebook, and Instagram. This easily accessible data has led to novel, highly informative research on emerging techniques in machine learning and big data analytics. However, it has also raised complex questions about research ethics and the fine line between ethically appropriate and ethically inappropriate research – irrespective of the accessibility of data or the legality of planned studies (Markham and Buchanan 2012, Metcalf and Crawford 2016, Williams, Burnap et al. 2017, Hesse, Glenna et al. 2018). These questions have largely focused on the probability of harm to individuals or communities, and how this consequence may be ethically navigated. Research that is perceived to be unethical by different stakeholders may lead to reputational and operational risks to researchers and institutions conducting studies – especially when the research touches on socially and politically sensitive areas or is designed to have an impact on subjects without their consent (Metcalf and Crawford 2016, Benigni, Joseph et al. 2017, Zimmer 2018).

Many researchers across the field of ethical social media research have raised similar questions regarding the right to use publicly accessible information (or PAI), including: What is the nature of privacy on social media? What is meant by “harm” in the context of social media research? What constitutes non-consensual and unwitting inclusion of individuals in studies? What are the responsibilities of researchers in understanding and addressing the ethics of the work they are involved in?

There are no clear answers to the complex questions regarding ethical versus unethical research in social media. Rather, there is general acknowledgement that decisions on *what* research is conducted and *how* it is conducted need to be made on a case-by-case basis guided by an ethical framework or set of principles. The Association of Internet Researchers (AoIR) publishes recommendations on ethical decision-making and internet research (now in their third edition) which provide context and guidance for individuals as well as research institutions (Ess 2002, Markham and Buchanan 2012, franzke, Bechmann et al. 2020).

The 2012 AoIR guidelines provide a highly valuable framework within which researchers are encouraged to ask critical questions around the potential impacts of their research on individuals and communities, and to develop context-specific ethical procedures and expectations. Beyond these though, there are a number of themes that occur across the growing literature on ethical social media research, including navigating between what it's possible to do and what it's ok to do; understanding privacy in the context of social media research; and understanding the many forms that "harm" can take.

Navigating between "can do" and "okay to do"

One of the prevailing justifications for using social media data without rigorous ethical checks and balances is the "because it's there" justification. As Zimmer (Zimmer 2018) and many others point out however, simply because a social media post is in the public domain, doesn't mean that it is ethical for a researcher to use it in any way they see fit. This is emphasized by Hesse and colleagues writing about qualitative research ethics and big data where they write:

"The assertion 'just because we can do something doesn't mean that we should do it' serves as a foundational ethical statement in bioethics and in the social studies of science and technology" (Hesse, Glenna et al. 2018).

The case that Zimmer cites is the Kirkegaard and Bjerrekær OkCupid Study (Kirkegaard and Bjerrekær 2016). In addressing the ethics of their work, Kirkegaard and Bjerrekær state "Some may object to the ethics of gathering and re-leasing this data. However, all the data found in the dataset are or were already publicly available, so re-leasing this dataset merely presents it in a more useful form." (re-leasing here refers to making previously-collected and curated PAI available to other researchers).

This, according to Zimmer, led to "[n]umerous news outlets reported on the controversial release of the data set, and experts in research ethics were quick to point out how Kirkegaard brazenly violated a fundamental principle of obtaining consent prior to releasing sensitive or personally identifiable information about research subjects, taking issue his claim that the data were already public and free for the taking" (Zimmer 2018).

Despite a number of social media researchers taking the attitude that public data is free to use "because it's there," there has been a growing consensus amongst scholars studying research ethics that this is an overstretch. Arguments draw in part on the nature of "informed consent" in human subjects research, and whether this applies to someone who has made a public statement on social media. However, they also draw on questions around user expectations of how their data will be used, the intent behind posting, understanding of what private and public mean in the context of social media, and how vulnerable the individual and their community are to harm resulting in the use of material they have posted. All of these questions raise ethical concerns that transcend legal considerations, and that underline the strong ethical principle of not doing something simply because it is possible and legal.

Rather, they require researchers to consider how use of social media data may impact the person associated with the data, and how this may threaten them in ways that, while not intended, nevertheless overstep expectations of ethical behavior. And within this framing, how privacy is understood is paramount.

Understanding privacy in the context of social media research

It is tempting to think of there being a black and white divide between public and private information, with the latter being off-limits without consent, and the former being fair game for researchers to use. However, as a number of researchers have pointed out, privacy on social media is more nuanced, and depends on the intent and understanding of the person generating content, and the context within which they are generating it (Ess 2002, Williams, Burnap et al. 2017, Fiesler and Proferes 2018).

Here, research has shown that many social media users are unclear about what privacy means in the context of what they post (Fiesler and Proferes 2018). While many realize that others can see their content, they are unaware to what extent others can use that content, and potentially use it in ways that disadvantage or harm them.

This becomes especially relevant when considering the context within which people post on social media. They may be responding to a very specific set of circumstances that lead to their content being potentially misconstrued when taken out of context. They may have been under emotional stress, or vulnerable in other ways. They may have not had a full grasp of how their content might be propagated and used in myriad ways around the world and over the coming decades. They could have been posting under conditions of diminished responsibility, or be a minor, or have violated someone else’s privacy in the process. And they could have had an expectation that their content was seen and used by a specific community at a specific point in time; no more.

Things get even more complex when the question of whom might potentially use their data for what is asked. Fiesler et al. (Fiesler and Proferes 2018) found that many Twitter users for instance aren’t aware that researchers can use their tweets without permission, and that some users may want to opt out of their data being used in certain ways. For instance, they found that a significant number of respondents in a survey of Twitter users would be uncomfortable with some uses of their tweets – especially if their name was used in publications. One survey respondent noted:

> “I would want to know how it was to be used, who would see it, whether my information would be kept anonymous and how long the tweet would be kept.”

And another:

> “If it’s personal, has identifying information, or embarrassing/ offensive/private I don’t want my tweets used.”

These and other studies suggest that, in the mind of social media users, privacy is not a legal concept, but one that depends on context and intent – and that a legalistic interpretation of privacy risks overstepping ethical boundaries. It is also one that taps deeply into the ideas of autonomy and dignity, where actions are modulated by potential impacts on a person's basic human rights, and not through a naïve demarcation between "public" and "private."

A useful analogy here is how privacy is considered in public spaces. Where someone is clearly grappling with embarrassing and challenging circumstances in public – especially where documenting and distributing details could threaten their autonomy and dignity – social and ethical norms typically dictate that the person's privacy is respected. And exploiting such public discomfort is generally seen as socially inappropriate.

Translating this to social media, to what extent are users entitled to their privacy when in danger of engaging in actions which others could use to threaten their autonomy and dignity, and in doing so cause considerable harm? And to what extent should researchers be aware of how their work might lead to such threats, irrespective of whether they were intended or not?

These are challenging but important questions to ask throughout the research process (franzke, Bechmann et al. 2020). They also require a sophisticated understanding of the nature of privacy. Reflecting this, Zimmer explores the conceptualization and execution of social media research in terms of "Information flows" – a term established by Helen Nissenbaum (Nissenbaum 2004) – and outlines a nine-step process to determine if planned research represents a potential violation of privacy (Zimmer 2018):

1. Describe the new practice in terms of its information flows.
2. Identify the prevailing context in which the practice takes place at a familiar level of generality, which should be suitably broad such that the impacts of any nested contexts might also be considered.
3. Identify the information subjects, senders, and recipients.
4. Identify the transmission principles: the conditions under which information ought (or ought not) to be shared between parties. These might be social or regulatory constraints, such as the expectation of reciprocity when friends share news, or the obligation for someone with a duty to report illegal activity.
5. Detail the applicable entrenched informational norms within the context, and identify any points of departure the new practice introduces.
6. Making a prima facie assessment: there may be a violation of contextual integrity if there are discrepancies in the above norms or practices, or if there are incomplete normative structures in the context to support the new practice.
7. Evaluation I: Consider the moral and political factors affected by the new practice. How might there be harms or threats to personal freedom or autonomy? Are there impacts on power structures, fairness, justice, or democracy? In some cases, the results might overwhelmingly favor accepting or rejecting the new practice, while in more controversial or difficult cases, further evaluation might be necessary.
8. Evaluation II: How does the new practice directly impinge on values, goals, and ends of the particular context? If there are harms or threats to freedom or autonomy, or fairness, justice, or democracy, what do these threats mean in relation to this context?

Finally, on the basis of this evaluation, a determination can be made as to whether the new process violates contextual integrity in consideration of these wider factors (Nissenbaum 2009).

These steps not only require approaching the concept of privacy from a sophisticated perspective, but thinking broadly about what is meant by "harm."

Understanding the many forms harm can take

The idea of avoiding or minimizing harm is a central tenet of research involving human subjects. However, the nature of that harm is open to interpretation. The 1979 Belmont Report, which is foundational to human subjects research ethics, notes that:

> "avoiding harm requires learning what is harmful; and, in the process of obtaining this information, persons may be exposed to risk of harm" (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1970).

In conventional human subjects research, harm is often taken to mean measurable physical or psychological harm within subjects who have given their consent to be part of a research study. However, harm is harder to pin down in social media studies where public data are used, there is no informed consent, and in most cases no awareness of the "subjects" of how their information may be used and how this in turn may affect them.

Here, the literature on social media research ethics leans toward broad and nuanced understandings of harm. Zimmer, for instance, writes:

> "There are numerous types of harm that participants might be subjected to, including physical harm, psychological distress, social and reputational disadvantages, harm to one's financial status, and breaches of one's expected privacy, confidentiality, or anonymity" (Zimmer 2018).

At this point, notions of dignity come into play, and more specifically, the protection of dignity as a basic right (The United Nations 1948, Bloustein 1964). Again, Zimmer writes:

> "merely having one's personal information stripped from the intended sphere of the social networking profile and amassed into a database for external review becomes an affront to the subjects' human dignity and their ability to control the flow of their personal information."

This emphasis on the dignity of research subjects is further emphasized in the third edition of the AoIR Ethical Guidelines (franzke, Bechmann et al. 2020).

Grappling with how even seemingly innocuous uses of social media data might threaten the dignity of the originators of those data is complex. But it's important – especially where a chain of events or associations may lead to threats to dignity that may not otherwise have occurred. For instance, if research exposes public posts to audiences that may otherwise not have seen them, and who, as a consequence, respond in ways that impact the dignity of the poster. This may take the form of a debilitating awareness of social commentary, online bullying, or social or professional censorship.

Emphasizing this, Metcalf and Crawford note that big data – including social media research – moves “ethical inquiry away from traditional harms such as physical pain or a shortened lifespan to less tangible concepts such as information privacy impact and data discrimination” (Metcalf and Crawford 2016).

From the perspective of social media users, harm begins to take on an aspect of how unanticipated use of information they post could lead to what is important to them being threatened in some way – a concept that reflects thinking around “risk innovation” (Maynard and Scragg 2019). Here, while there may be an expectation of social media users giving some consideration to how their public posts may impact them in the future, considerable responsibility lies with the users of this information to avoid harm where possible – especially in a research setting.

For instance, if the poster is a minor, or vulnerable as a result of their gender, sexual orientation, ethnicity, political or ideological leanings, indigeneity etc. harm may occur where posts are taken out of context, made more accessible to communities that may target the originator, or used to ridicule or discriminate against the originator. This potential for harm may be amplified when posts are used to classify the originator with specific groups or individuals which in turn attract potentially harmful attention – including where classification is inaccurate – and where compiled and inferred information on them is made accessible to organizations and individuals who have influence over their lives. This becomes more complex still in the context of historical power asymmetries, including where without appropriate approaches, “data from qualitative research could be used to further colonize, exploit, surveil, and control indigenous people and knowledge” (Hesse, Glenna et al. 2018). Indeed, Hesse et al. note that many indigenous communities are covered by specific regulatory protocols for research precisely because of this.

In all these cases, using public posts outside the time, locational and cultural context within which they were intended further exacerbates the potential for harm.

The notion of vulnerability is clearly articulated in the second edition of the AoIR guiding principles for ethical research, where it is stated that the “greater the vulnerability of the community / author / participant, the greater the obligation of the researcher to protect the community / author / participant.”

Building on this, the guidelines note that, “because ‘harm’ is defined contextually, ethical principles are more likely to be understood inductively rather than applied universally” and that the best approach to ethical decision-making is through the “practical judgment attentive to the specific context” (Markham and Buchanan 2012).

The resulting complexity around how harm is understood and navigated in social media research places an onus on researchers to think carefully through the potential consequences of their research and how these are balanced by anticipated benefits, as well as who benefits and who bears the risks. And it underlines the need for researchers to be integral to the process of assessing and navigating ethical concerns, rather than simply following a code of conduct (Markham and Buchanan 2012).

Addressing the challenges of uninformed and unwitting participation

As was mentioned previously, the notion of "informed consent" is a cornerstone of human subjects research, but one that raises considerable ambiguity in research using readily accessible social media records. Under the Common Rule in the United States, informed consent is required for human subjects research, except under one of five exceptions (Heath and Human Services 2018).

These include:

1. Research involving normal practices in educational settings;
2. Research that only includes interactions involving educational tests;
3. Research into benign behavioral interventions if consent is prospectively sought;
4. Secondary research on data (within specific limitations);
5. Research conducted or funded by Federal departments or agencies that are designed to "study, evaluate, improve, or otherwise examine public benefit or service programs"; and
6. Taste and food quality evaluation and consumer acceptance studies.

Research using publicly accessible social media data is often considered to be exempt under clause 4 of the Common Rule, and potentially under clause 5 – although within research institutions exemption still needs to be sought and agreed on by an Institutional Review Board in most cases. However, a number of researchers have contested this interpretation of the Common Rule.

In analyzing the OkCupid case study for instance, Zimmer noted that informed consent, while important under conventional human subjects research, is complex with social media research where subjects did not anticipate their material being used for research purposes (Zimmer 2018). More specifically, Benigni, Joseph and Carley note in a Twitter-based study on online extremism that although "many users understand their online behavior is used for marketing purposes, they may not be comfortable with their behavior being used to inform diplomacy or military operations. Indeed, one could assume users would not consent to the use of their information for intelligence collection" (Benigni, Joseph et al. 2017).

Of course, the very nature of social media research means that in all but a few cases, informed consent is not possible to obtain. Recognizing this, most scholars of ethical social media research place the responsibility on researchers to take special care when considering the implications and impacts of their work (Markham and Buchanan 2012, Hesse, Glenna et al. 2018, franzke, Bechmann et al. 2020). But this in turn requires researchers to develop a nuanced understanding of what participation might mean to subjects, should they become aware of it – and especially if participation leads to harm in some form.

Here, concerns around unwitting and uninformed participation are reflected in various ways across the literature. Uninformed participation is, as it implies, the use of public or readily accessible content in a study that is uniquely identified with a particular person, and of which they have no knowledge.

The notion of the uninformed subject covers much social media-based research as, given the nature of large datasets, seeking permission would often throttle research and analysis to the point of making it untenable. As a result, there is broad understanding that most social media research will be conducted without the knowledge of subjects. At the same time, there is growing awareness that this places an extreme level of responsibility on the researchers themselves to ascertain whether the use of data from uninformed subjects is appropriate and, if so, what the bounds of appropriate use are.

Central to this responsibility is consideration of the intent of a subject in posting content and their understanding of acceptable or unacceptable use, and the consideration of the potential of harm occurring to uninformed subjects as a result of their content being used in a study. In other words, a guiding question might be: If a subject did know how their content was going to be used and the potential consequences associated with this, would they have given informed consent if given the opportunity?

Any answer to such a question will inevitably be subjective and open to contestation. However, it can be argued that the process of reflecting on the intent and perspective of subjects is itself an important aspect of due diligence in social media research. It’s also a process that is likely to either provide clear ethical support for proposed research, establish ethical guardrails, or raise ethical red flags, depending on the nature of the research. For instance, it is reasonable to assume that most subjects would agree to support research that seeks to mitigate the impacts of disease or natural disasters. Where research aligns with broadly accepted public good but rubs up against deeply held beliefs or worldviews, such reflection is likely to lead to guidelines and procedures that would be supported by a majority of subjects. On the other hand, research that has the potential to intentionally or unintentionally lead to discrimination against vulnerable communities, or within communities having specific ideology, political leanings or deeply held beliefs, is likely to raise red flags.

In this way, even where subjects are uninformed through the nature of social media research, ethical and responsible research can be guided through consideration of and reflection around whether subjects would give consent if informed. Where research is contentious or the ethics are more complex than usual, this process may be augmented by further engaging with a broad range of experts, including representatives of potentially impacted groups.

Within this category of uninformed subjects, there’s a subcategory of unwitting subjects. These are people who have a limited grasp of who can use their content and how it can be used. For example, this might include users who have posted content without thinking through the consequences: possibly while under the influence of various substances, while suffering from mental or physical illness, while stressed or sleep-deprived, or while engaging with a specific community or group – tweeting from a party or an event for instance. In these and other cases of unwitting participation in social media studies, subjects may not have been aware at the time of the full implications of their public posts, nor have been in a position where they were not able to exercise such awareness.

While it’s easy to dismiss unwitting users as naïve – especially given the public nature of social media platforms – this argument does not hold up to ethical scrutiny. Rather, from an ethics perspective there is a duty not to exploit subjects who inadvertently place themselves in a vulnerable position, and who may subsequently come to regret it.

As with uninformed subjects, this raises an ethical challenge for social media researchers that is not readily resolvable through codes of conduct or research ethics checklists. Rather, it takes a high level of reflexiveness and due diligence to consider the potential risks of including unwitting subjects in studies, and how their inclusion potentially benefits or harms them.

An inductive and deliberative approach to ethical social media research

The responsibility of researchers to test, assess and iteratively examine the risks and benefits of their research to the people and communities they are drawing on, is articulated in much of the literature around ethics and social media research. It is perhaps best captured in the 2012 AoIR Key Guiding Principles, which state:

> “Ethical decision-making is a deliberative process, and researchers should consult as many people and resources as possible in this process, including fellow researchers, people participating in or familiar with contexts/sites being studied, research review boards, ethics guidelines, published scholarship (within one’s discipline but also in other disciplines), and, where applicable, legal precedent.” (Markham and Buchanan 2012).

This emphasis on researcher responsibility is further reinforced in the third edition of the guidelines, with the recognition that researchers are part of a broader community that has shared responsibility for ethical behaviors and practices (franzke, Bechmann et al. 2020). In contrast, it is hard to find experts in this field advocating for ethics boards and processes that are not intimately intertwined with researchers and the research process.

This, of course, places a substantial level of responsibility on social media researchers to think through the consequences of their actions, and to devise methodologies and approaches that respond to the rights, dignity, well-being and safety of the people whose content they are using. Navigating this is not easy – especially in disciplines that have not emphasized an integrated approach to ethical practices. Nevertheless, it is essential if the benefits to subjects and the communities they represent far outweigh any potential risks.

As a first step, deliberative and inductive consideration of ethical issues is important. Researchers should be encouraged to explicitly consider the potential vulnerabilities of the people whose content they are basing their research on, their intent in posting and the likelihood of them giving consent for use if they knew how this content was going to be used, and the potential harm that might arise to individuals from the planned research – including threats to areas of value such as dignity and autonomy. And this should be a process of continual questioning, exploration, and evolving perspectives – both through self-deliberation, and deliberation and discussion with others.

The 2012 AoIR guidelines provides a comprehensive set of questions for researchers to help with this process (Markham and Buchanan 2012). These include (with additional sub-questions not included here):

- How is the context defined and conceptualized?
- How is the context (venue/participants/data) being accessed?
- Who is involved in the study?
- What is the primary object of study?
- How are data being managed, stored, and represented?
- How are texts/persons/data being studied?
- How are findings presented?
- What are the potential harms or risks associated with this study?
- What are potential benefits associated with this study?
- How are we recognizing the autonomy of others and acknowledging that they are of equal worth to ourselves and should be treated so?
- What particular issues might arise around the issue of minors or vulnerable persons?

These questions are not comprehensive and should not be approached as a check-list. However, along with other resources, they are a valuable prompt for researchers as they begin to think about the broader impacts of their research on the people they are drawing information from, and potentially impacting as a result.

They also represent a principle that is deeply embedded in the AoIR guidelines, and is reflected elsewhere in the literature: that ethical issues need to be considered at every step in the research process.

Addressing ethical Issues at every step of the research process

In recent years there has been a growing literature on how the landscape around research affects the appropriateness and impacts of the research – from the social and cultural contexts it exists in, to the institutions and policies that guide and sustain it, to the perspectives, perceptions and biases of the people involved in it. This has long been recognized in areas that are now seen as deeply unethical – eugenics for instance, or public health research conducted on individuals and communities without consent. Building on this, there is a growing scholarship around how the ways research is framed, how research questions are formulated, and how methods are developed and applied, can embed ethically questionable practices into the process. As a result, there is increasing awareness of the need to embed ethical and responsible research practices within every stage of the research process (Jasanoff 2007, Markham and Buchanan 2012, Stilgoe, Owen et al. 2013, Sarawitz 2016, franzke, Bechmann et al. 2020).

This need is heightened where there is considerable uncertainty over the ethical norms and expectations associated with a particular area of research, and where the potential impact on individuals contributing to it – albeit without their knowledge – is unclear yet possibly harmful. This is particularly the case where the formulation of research questions and preliminary research plans run the risk of locking in research directions and practices that potentially overstep ethical boundaries, yet are hard to alter once initiated.

As a result, the AoIR advocates for ethical issues being addressed:

> "during all steps of the research process, from planning, research conduct, publication, and dissemination" (Markham and Buchanan 2012).

This approach once again eschews reliance on separate ethics boards or add-on ethics considerations, and suggests that social media researchers should integrate a process of continuous and accountable consideration of research ethics in every stage of their work. This, through necessity, should include collaboration with research ethics experts, and their integration into research teams. It should also include, as appropriate, consultation with other communities of experts, practitioners, and constituents, to ensure a continued re-evaluation and recalibration of the ethics landscape that research is being conducted within.

Such an approach is intended to help iteratively identify and navigate potential ethical challenges, and specifically threats of harm to people whose content is used in studies – including potential harm related to dignity, equity, justice and inclusion.

However, there is another aspect of threat that is associated with research ethics that is intertwined with harm to subjects, yet extends beyond this, and that is potential threats to institutions involved in social media research – especially when that research touches on sensitive areas.

Darknet-based CIO research ethics

While the main focus of recent IO and CIO research has been in social media, the Darknet sphere that is hidden from the mainstream (or "clear") web and is largely anonymous, remains to be an important cyber-arena for IO. Darknet-based CIO is beyond the scope of this paper. Yet, the current status of ethics discussion surrounding Darknet research is nevertheless worth summarizing.

The Darknet – hidden online networks that are accessible through specialized or filtered routing technologies – is another important space for adversarial influence operations. Thus far, no comprehensive ethical guidelines for Darknet research exist, and thus recommendations have been made for Darknet research projects by borrowing social media research ethics and

Criminology research ethics. (Benjamin, Valacich et al. 2019) is one of the few publications that address the lack of ethical framework for Darknet research, pointing to three issues that are pertinent to Darknet-based IO and CIO research, include (1) data secrecy, (2) circumventing anti-crawling, and (3) direct interaction with Darknet users.

(1) **Data Secrecy:** Social media data are mostly PAI, mainly raising the concerns about “because it’s there” justification (Zimmer 2018). Conversely, the Darknet communities and the data originated from these communities are intended to be covert, hidden, and private. This nature of secrecy can pose a unique dilemma. On the one hand, access to such data could allude to the violation of data privacy because they are meant to be private. On the other hand, the data is inherently anonymous and thus does not contain personally identifiable information. That is, unlike PAI that contains personal cues, most Darknet data do not qualify for the technical definition of human subject research.

(2) **Circumventing anti-crawling:** While web-crawling has been a common data collection practice for researchers, Darknet crawling entails a customization to circumvent the software employed precisely to protect an online community from being crawled. The use of anti-crawling is tied to the covert nature of the Darknet and its users’ pursuit of data secrecy. The use of a customized crawler means some level of “deception” has to be practiced by researchers. That being said, given that there is no hindrance to a manual inspection of the Darknet sites, it is unclear to what extent circumventing anti-crawling should be accepted or deterred. In particular, crawling Darknet data for CIO purpose could bring greater social benefits (i.e., national security, deterrence of cyberattack) while the harm from the crawling may be minimal. Cost-benefit analysis is thus critical in this sense (Benjamin, Valacich et al. 2019).

(3) **Direct interaction with Darknet users:** While the use of Darknet online data does not qualify as human subject research in most cases, some Darknet CIO research may entail direct contact with the Darknet users. Interacting with them is more sensitive than conventional cyber-research contexts that would involve the general public members (Barratt and Maddox 2016). While there is no standard protocol in conducting human subject research in Darknet, recommendation has been made to consult with the IRB protocols, as well as learning insights from human subject-based Criminology research.

Application to CIO research

While CIO research that is focused on social media is only part of the broader CIO research landscape, it is nevertheless a domain of increasing importance, and as such, there is a need to be cognizant of the ethical challenges and pitfalls that infuse it. But understanding the complexities and emerging thinking around ethical social media research, a framework can begin to be built around exploring what constitutes ethical versus unethical CIO research within the context of social media, and how research goals can be reached while operating within ethical and socially responsible boundaries.

Here, there is a close coupling between ethics, values, value creation/protection, and risk. These are addressed later as we look at the application of risk innovation to both understanding and navigating the challenging landscape around CIO research. There are also substantial lessons to be drawn from emerging thinking around AI ethics, and we address this next.

4. Applying Artificial Intelligence Ethics to Counter IO Research

Artificial Intelligence (AI) and Machine Learning (ML) have both direct and indirect relevance to CIO research. Both are at the core of many IO approaches utilizing online platforms. More than this though, AI and ML are becoming increasingly scrutinized for their inherent biases and the ethical and moral challenges they present. And because of this, there is substantial overlap around ethical CIO research and ethical development and use of AI/ML.

There have been multiple efforts to establish a set of core ethical values for AI research to follow (for instance see the AI Ethics Guidelines Global Inventory compiled by Algorithm Watch (Algorithm Watch 2020)). These principles fall into several categories: accountability, bias/fairness, privacy, and understandability. The US Department of Defense, along with dozens of other public and private entities, has issued statements on ethical principles that should guide the development of AI/ML systems. The goal of these principles is to develop unbiased, privacy-protecting algorithms whose decisions are understandable and/or traceable. In this chapter we review the latest standards for ethical AI, frame these standards in the context of Counter IO research, and introduce several tools to mitigate risks associated with using AI for Counter IO work.

Inherent ethical issues with AI

When designed with care and expertise, AI/ML systems have the potential to mitigate biases and ensure more consistent outcomes than human decision makers alone. However, AI/ML systems are also designed to discriminate information which makes them inherently biased. Bias is harmful if it leads to a negative impact on individuals or groups in a way that is not relevant to the intended purpose of the system or otherwise violates societal values. It may take the form of making a decision based on irrelevant sensitive attributes, arbitrarily holding a person or demographic group to a different standard. Recently, numerous examples of algorithms propagating and exacerbating societal biases have come to light across several major industries such as finance, employment, retail, government, and internet services. For example, Twitter's admission that its algorithm favors conservative political themes and amplifies those narratives (BBC 2021).

Bias can emerge from anywhere in the system: the input data, developer assumptions, human interaction, and use in real-world settings. AI/ML systems can learn from historically biased data and create feedback loops that affect future training data and decision-making in operational settings. Human decision makers may override AI systems and impart cognitive biases, others may trust the system even when it makes a mistake.

As an example, systematic errors may result from the following types of bias:

- *Selection bias*, due to imbalanced sampling from certain groups
- *Omitted-variable bias*, when an independent variable is left out
- *Exclusion bias*, due to the systematic exclusion of certain individuals or groups
- *Analytical bias*, due to the way that the results are evaluated
- *Reporting bias*, or skew in the availability of data
- *Attrition bias*, due to loss of participants in a study over time
- *Observer bias*, when the researcher subconsciously influences an experiment due to cognitive bias where judgment may alter how the experiment is carried out
- *Detection bias*, when a phenomenon is more likely to be observed for a particular group
- *Recall bias*, due to differences in the accuracy or completeness of participant recollections of past events

It is critical that any processing, assistance, or decision-making capability by a machine learning or artificial intelligence system operates lawfully and without discrimination against protected classes in accordance with Executive Order 12968 (White House 1995). The US Department of Defense, along with dozens of other public and private entities, has issued statements on ethical principles that should guide the development of AI and ML systems (Algorithm Watch 2020).

Artificial Intelligence for CIO

The definition that an individual, community, or organization chooses for fairness is tangled across many layers, stakeholders, and participants of a system. Definitions of fairness seek to provide equal impact and equal treatment across groups, subgroups, or individuals. The debates surrounding this definition range in scope from the mechanics of a specific algorithm, to the vast social system in which the specific algorithm plays a small role. An emerging approach in the academic literature is to elicit a definition of fairness from non-technical stakeholders by asking questions, examining scenarios, and performing pairwise comparisons of decisions made by an algorithm.

The core principles of ethical AI research can take on different meanings when applied to Counter IO research. Executive Order 12968 requires processing, assistance, or decision-making capability by an AI/ML system to operate lawfully and without discrimination against protected classes. However, during times of war certain executive orders (directives issued by the President of the United States) may not apply and even during peacetime, protected classes, as defined by these executive orders, may not align with the protected classes defined by private entities and assumed by traditional users. For example, citizenship or country-of-origin (which may be protected to prevent bias in a traditional AI/ML setting) is often a necessary discriminator in government applications.

In the past few years, many organizations across the US Government, commercial industry, and public sectors have issued statements on the ethical development and use of artificial intelligence. The 2019 recommendations from the Defense Innovation Board (Defense Innovation Board 2019), which are the foundation for the Department of Defense's approach to Responsible AI (Deputy Secretary of Defence 2021), outline five major principles to guide the design and use of AI: Responsible, Equitable, Traceable, Reliable, and Governable. Furthermore, the DoD "should take deliberate steps to avoid unintended bias in the development and deployment of combat or non-combat AI systems that would inadvertently cause harm to persons."

Within US law, notions of fairness provide the foundation of anti-discrimination laws in employment, labor, education, voting, and housing, including but not limited to: the Civil Rights Act of 1964 and 1968, the Equal Pay Act of 1963, the Age Discrimination in Employment Act, the Rehabilitation Act of 1973, the Civil Rights Act of 1991, and the Americans with Disabilities Act. These laws prevent discrimination on the basis of protected attributes of race, color, religion, sex, national origin, age, disability, sexual orientation, pregnancy, familial status, veteran status, and most recently, genetics.

Discrimination is often defined in terms of disparate treatment and impact. In the US, disparate treatment refers to unequal or unfavorable behavior toward someone because of a protected characteristic. Disparate impact occurs when an individual or group is unequally affected by an otherwise neutral treatment or practices.

Stages of the CIO pipeline and how AI may interact in those stages

The CIO process involves a significant amount of data collection, processing, and human assessment. Artificial intelligence, machine learning, and other forms of software aid can add efficiency, accuracy, and consistency.

Increased Efficiency: AI may expedite the investigation process by handling errors intelligently, routing cases, correlating information, sorting and ranking, structuring data, or converting information (e.g., transcription of verbal interviews).

Capability Enhancement: AI may augment the capabilities of human investigators and adjudicators by discovering societal trends at larger scales, predicting outcomes, recommending investigative leads or follow-on questions, or learning from past decisions.

Process Improvement: AI may help understand and improve the entire business process, such as determining investigative yield of various data sources, assisting adjudicators by providing contextual, statistical, or historical information, performing quality assurance to improve consistency of decisions, or aiding in specializing the workforce (e.g., case routing to experts).

Mitigation Strategies

Many of the inherent biases that affect AI/ML research can be addressed by introducing the appropriate mitigation strategy at the appropriate stage in the research pipeline. In traditional AI/ML applications, there are still a wide variety of biases and privacy needs. Recently, developers of mitigation strategies have begun to step away from reducing bias on narrow classification tasks and instead have begun to offer tools that allow a user to select or define what is appropriate for their given application and context (Shrestha, Kafle et al. 2021). These new frameworks aim to achieve positive trends across several metrics, rather than optimizing for zero bias according to a single measure.

Mitigation strategies typically occur at one or more of the pre-processing phase, the in-processing phase, and the post-processing phase. In terms of the CIO research pipeline, these phases generally align with investigation, review, or adjudication. Mitigation strategies tend to require knowledge of the protected or sensitive attribute. In many real-world systems, this information is unknown or illegal to collect and use, even for the purpose of improving fairness of AI systems making it especially challenging to combat bias.

To make matters more complex, not all instances of bias are necessarily unwanted. For example, if the current empirical data does not reflect a desired statistical distribution, then using intentionally biased training data may be useful. If the bias is not problematic according to the desired fairness goals, it may not be necessary to mitigate it. If it is problematic or impractical to mitigate, bias can be intentionally introduced within other components to balance the system ensuring that it operates fairly in whole.

One such example of a mitigation toolkit that leaves much of the decision-making up to the user is the IBM Fairness 360 framework (Bellamy, Dey et al. 2019).

Applying mitigation strategies to CIO AI/ML

During the CIO process, mitigation strategies can be introduced to reduce bias.

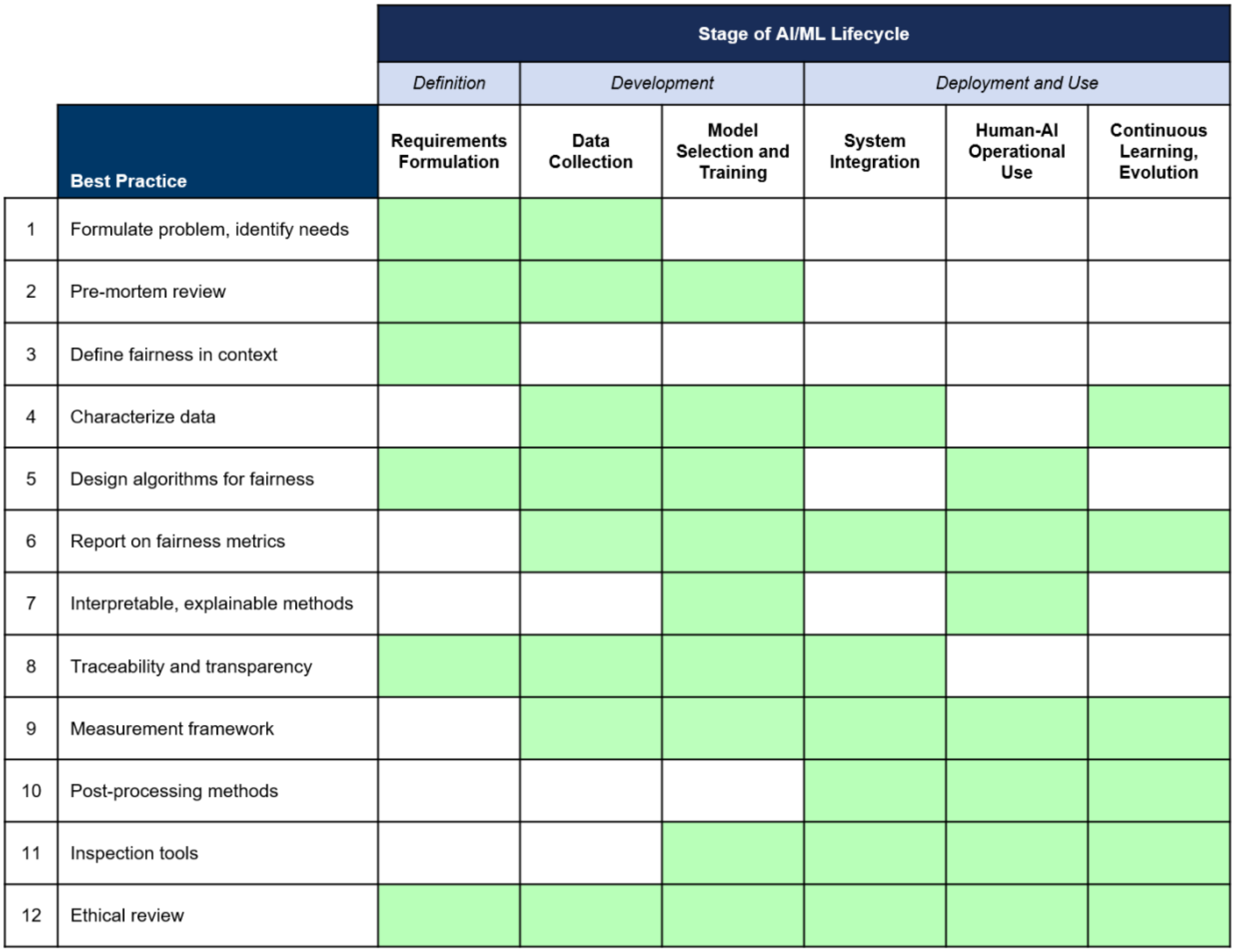
Based on the literature surveyed in this report and our experience developing human-in-the-loop AI/ML systems, we have identified the following best practices for preventing algorithmic bias and related unintended consequences (refer also to figure 1):

1. Carefully formulate the problem to be addressed to understand how an AI/ML solution can transform capabilities beyond the current practice. This process will help identify

what data or resources will be needed, and how the AI/ML solution will fit into the larger system.

2. Before development, acquisition, and use of an AI/ML solution, conduct a “pre-mortem” project review. Similar to a post-mortem review in which causes of project failures are identified, a pre-mortem review will help identify risks of harm from AI/ML solutions, and offer methods to prevent or mitigate such harms. The review should involve a diverse range of stakeholder viewpoints, especially from those who will be affected. Include questions such as:
    a. What is the expected benefit over the current practice or baseline human performance?
    b. What are ways in which the AI/ML can make things worse - through misuse, errors, propagation of societal biases, or other unintended consequences?
    c. What are possible consequences of an evolving system that may involve human-AI interaction? Could humans impart biases on the system?
    d. Could the system become problematic over time as it learns from new data? Is there potential for feedback loops or other long-term, downstream effects?
    e. What is the likelihood, magnitude of impact, and scale of each of these risks?
    f. Can these risks be reduced through safeguards? What is the cost and feasibility of implementing the safeguards?
3. Depending on the use case and who may be affected, define what fair and unfair outcomes would look like. The process of defining these values will provide guidance for assessing the AI/ML components going forward and help identify which measures of fairness/bias fit the particular use case. In some cases, measures tied to legal definitions of discrimination may be appropriate, in others, individual fairness may be most relevant. Consider whether one type of mistake is more costly than another (e.g., incorrectly flagging a trustworthy account vs. failing to detect an untrustworthy account). How does this cost factor into the chosen fairness constraint? Multiple metrics may be needed to characterize the behavior of the system. Fairness elicitation frameworks, such as we are defining here, can help derive appropriate measures from the value judgments of all stakeholders, both technical and non-technical.
4. Knowledge of your data is the best way to anticipate potential sources of bias. Characterize the data with respect to sensitive attributes to understand balance, distribution, and dynamics that may lead to unfair outcomes. Consider whether the data reflects the realities of the intended use case, and how these characteristics may change as the system evolves over time. Explore how the dataset was collected, who annotated it, and whether there are features that may inadvertently act as proxies for sensitive attributes.
5. To safeguard against bias, consider AI/ML learning approaches that are designed with fairness in mind, such as techniques that incorporate fairness as a constraint or joint optimization objective. Avoid optimizing for the majority, especially in cases where datasets are highly imbalanced or not adequately representative of certain groups.
6. To enable developers to assess bias and fairness, provide appropriate datasets with demographic information. Require developers to report on fairness metrics in addition to standard performance measures such as accuracy and speed.

7. When choosing a machine learning method, consider alternatives to “black box systems” that provide greater interpretability and explainability. Though there may be differences in performance, some methods may make it easier to ensure no causal link between a sensitive feature (such as race or gender), and the output of the AI/ML system.
8. Promote traceability and transparency during development to enable enhanced understanding and early identification of sources of bias. Collect documentation of all software requirements, intended use cases, design decisions, known limitations, and performance under varying conditions. Example documents include software engineering artifacts, Datasheets for Datasets, and Model Cards for Model Reporting, which help identify strengths and weaknesses of datasets and models.
9. AI/ML components are often part of much larger socio-technical systems. To manage bias in complicated systems, create a measurement framework by identifying points within the system at which bias and fairness can be assessed at different levels (data and sensing, algorithmic design, human-AI interaction, and system/mission level). Create ‘hooks’ in the software to enable continuous or periodic measurement of bias at various stages, over time.
10. For existing systems in which bias is found, consider updating data, inserting modules, or using post-processing methods to counteract harmful biases
11. To complement automated measurements, create tools or methods to enable inspection of the AI/ML system by various stakeholders to identify issues and build trust. The method employed may differ by stakeholder, such as counterfactual examples for operators and aggregated group statistics for supervisors.
12. Consider having an independent third party perform an ethical review or audit of the AI/ML system. This applies to both future systems as well as existing deployed systems, and safeguards against bias that may come from development team and mission personnel



		Stage of AI/ML Lifecycle					
		Definition	*Development*		*Deployment and Use*		
	Best Practice	**Requirements Formulation**	**Data Collection**	**Model Selection and Training**	**System Integration**	**Human-AI Operational Use**	**Continuous Learning, Evolution**
1	Formulate problem, identify needs	🟩	🟩				
2	Pre-mortem review	🟩	🟩	🟩			
3	Define fairness in context	🟩					
4	Characterize data		🟩	🟩	🟩		🟩
5	Design algorithms for fairness	🟩	🟩	🟩		🟩	
6	Report on fairness metrics		🟩	🟩	🟩	🟩	🟩
7	Interpretable, explainable methods			🟩		🟩	
8	Traceability and transparency	🟩	🟩	🟩	🟩		
9	Measurement framework		🟩	🟩	🟩	🟩	🟩
10	Post-processing methods				🟩	🟩	🟩
11	Inspection tools			🟩	🟩	🟩	🟩
12	Ethical review	🟩	🟩	🟩	🟩	🟩	🟩

Figure 1. Actions can be taken to improve fairness of the system at all stages of the AI/ML lifecycle. Boxes highlighted in green represent when in the process that each of the Best Practices would apply.

In summary, there are close parallels between AI/ML research and CIO research. As a result, developing thinking, literature, and guidelines, on ethical and responsible AI serve to inform and guide the less mature field of ethical and responsible CIO research.

The challenge in both cases comes in moving from identified ethical challenges to practical pathways forward. And here, the risk innovation approach provides a pragmatic framework for decision-making in the face of complex and hard to quantify ethical/social risks.

5. Risk Innovation and Counter IO Research

As can be seen from the preceding chapters on stakeholders, social media research and artificial intelligence ethics, the complex challenges associated with ethics and vulnerabilities around research relevant to IO and CIO leads to potentially novel risks to researchers, research participants, engaged communities, and even the institutions supporting such research. Many of these risks extend beyond well-established risk categories to areas that are often hard to quantify, such as reputation risk, moral/ethical risk, and risk to identity, dignity, and autonomy.

Here, there is a blurring of lines between risk and ethical behavior and actions. Ethics deal with what is considered to be appropriate behavior at a community/societal level. Many aspects of ethics and ethical behavior and norms relate to consequences of actions, or outcomes from decisions. This is not universal across ethics – for instance, deontological ethics holds that there are absolutes to moral behavior that are independent of consequences. However, many ethical frameworks – including consequentialist ethics (which includes utilitarianism) – contend that the outcomes of actions have some bearing on what is ethical versus what is not.

This emphasis on consequences or outcomes provides a direct connection with risk. Risk, in its simplest form, concerns the likelihood of adverse outcomes arising from decisions or actions. And thus, if adverse outcomes are considered to include immoral, inappropriate, or otherwise "bad" consequences as determined by societal norms and consensus, there arises a close relationship between ethics and risk. However, such "ethical" or "societal" risks are often uncertain, ambiguous, and societally complex, meaning that navigating the risk/ethics landscape is challenging. And this is especially the case in CIO research, where many of the risks depend on ambiguous and far from universal assumptions of right and wrong.

Here, the concept and framing of Risk Innovation provides a useful approach to navigating the ethical challenges presented by CIO research. The Risk Innovation framework developed by researchers at ASU is specifically designed to help map out and navigate emerging risk landscapes that are dominated by ambiguity, uncertainty, and socially complex risk (Risk Innovation Nexus 2021). Risk innovation is based on a reframing of risk as a threat to value – both to a principal agent (for instance a researcher, an entrepreneur, or an organization engaging in a specific enterprise) – and to key stakeholders that are impacted by, and in turn have the ability to impact, an enterprise (Maynard 2015). Here, there is often close alignment between "value" and ethics, although care needs to be taken not to inappropriately conflate "value" (which refers to how much or how little worth something is considered to have) with "values" (which refer to beliefs around what is right and what is wrong).

In this context, "value" may constitute something of worth that already exists – the right to pursue certain goals for instance, or the current state of a community or society. "Value" may also be something that is aspired to, such as achieving stated goals, bringing about certain outcomes, or reaching a desired set of conditions. Risk in this context is thus formulated as threats to what an individual, a team, an organization or a community are striving to maintain or aspiring to grow or produce. In the case of CIO research, value may take on a number of forms, including the development of effective CIO measures that lead to a more secure, safer society.

Within this framing of risk, threats to value often take on forms that lie beyond the ability of conventional risk management tools and systems to address. For instance, research that is perceived by investors, publics or political representatives to overstep ethical boundaries may be stymied as these and other stakeholders see it as a threat to something they value (in this case, deeply held ethical principles, or power and influence that is predicated on alignment with specific ethical principles). Such threats are subjective, rooted in human behavior, complex, and near-impossible to quantify. Yet they often play an outsized role in determining success or failure.

To help navigate such risks, the Risk Innovation approach uses eighteen “orphan risks” (Figure 2, Table 1) that span three overarching categories: social and ethical factors, unintended consequences of emerging technologies, and organizations and systems (Risk Innovation Nexus 2021). The term “orphan risk” refers to risks that tend to be ignored or overlooked as they are too subjective or qualitative to easily address, and yet have a tendency to create substantial issues if they are not addressed (Maynard 2018). These eighteen orphan risks are not inclusive, but they do provide a pragmatic framework for making sense of complex and unconventional risk landscapes.

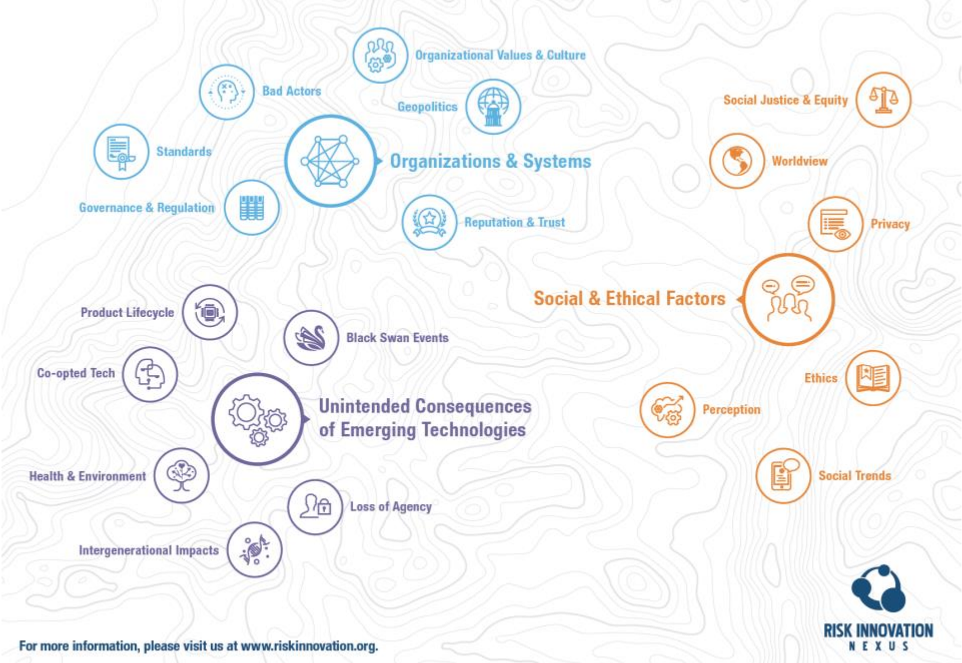


Figure 2. The eighteen orphan risks used within the Risk Innovation framework. (Risk Innovation Nexus 2021)

Table 1. Orphan Risks. (Risk Innovation Nexus 2021)

Social & Ethical Factors		**Unintended Consequences of Emerging Technologies**		**Organizations & Systems**	
Ethics	Risks from business practices overstepping the often-indistinct line between ethical and unethical behavior.	**Black Swan Events**	Risks from very low probability but high impact events.	**Bad Actors**	Risks from enterprises that behave in ways that are ethically questionable or that lead to unacceptable harm.
Perception	Risks created from how people perceive a technology to impact/threaten what they think is important.	**Co-opted Tech**	Risks from technologies and products that are used in ways that undermine the intention of the	**Geopolitics**	Risks from a lack of awareness of or strategies for navigating a shifting geopolitical landscape.

			original business or business owner.		
Privacy	Risks from the social pitfalls associated with the use and misuse of individual’s data.	**Health & Environment**	Risks from new technologies, and the products they are associated with, behaving in sufficiently novel ways that potentially lead to threats to human health and the environment.	**Governance & Regulation**	Risks from often evolving laws, policies, and practices that govern and guide business operations.
Social Justice & Equity	Risks from business practices and technologies that marginalize or disadvantage specific segments within society.	**Inter-generational Impacts**	Risks from technologies that have potential impacts from one generation to another.	**Organizational Values & Culture**	Risks from tensions between business practices, both internal and external, and the set of values that reflect what is important to a business’ founders and members.
Social Trends	Risks from shifts in social norms, changing consumer expectations, or evolving cultural behaviors.	**Loss of Agency**	Risks from products or business practices that reduce the ability of organizations and individuals to make decisions.	**Reputation & Trust**	Risks from a business having only a rudimentary understanding of how their behavior and actions strengthen or weaken reputation and trust.
Worldview	Risks from people’s deeply-held beliefs about how they view the world and how it should function.	**Product Lifecycle**	Risks from unintended impacts of where and how a product’s materials are sourced and manufactured, how it is used, and its disposal and/or reuse.	**Standards**	Risks from a business’ lack of engagement with an often-evolving operational framework for businesses that spans legal requirements, informal guidelines, and norms and codes.

Building on the concepts of risk as a threat to value and orphan risks, the risk innovation framework sets out to provide pragmatic and context-dependent approaches to understanding and navigating novel risk landscapes. The framework and its associated tools were initially developed for entrepreneurs developing novel products, where understanding and navigating unconventional risks could make the difference between success and failure. However, both framework and tools are extendible to any context where informed decisions need to be made in the face of unconventional and unfamiliar risks.

This framing makes the risk innovation especially relevant to researchers, policy makers and organizational directors and leaders where there is a similar tension between time and resources, and effective decision-making – including CIO research and actions.

Building on this, the framework and tools associated with risk innovation embody a philosophy of an informed mindset where the process of exploring and mapping out orphan risks leads to a greater awareness of possible pitfalls and productive pathways forward. This is built on appreciating a complex risk landscape from multiple stakeholder perspectives, while incorporating insights that may not be apparent from narrow disciplinary thinking.

The resulting tools and processes are designed to take individuals, teams and organizations through a process of identifying key stakeholders, highlighting top-level areas of value that are important to these, identifying orphan risks that potentially threaten these areas of value, and then strategizing iteratively around how to simultaneously protect and grow value for the principal agent while avoiding unnecessary threats to stakeholder value (on the assumption that threatening stakeholder value becomes a threat to principal agent value). They are also adaptable to a wide range of challenges.

For instance, in the specific case of CIO research, we were able to extend the tools to identify specific stakeholder groups and personas to help inform discussions and decisions around orphan risks and risk navigation strategies. We explore these further in section 7 below.

This approach to risk leads to unique and powerful ways to consider and navigate ethics in CIO research, and to develop effective approaches to achieving goals that are not blindsided by orphan risks.

It's an approach that is highly effective in helping individual researchers understand both how their work may threaten others--resulting in barriers to its progression and effectiveness--and one of the key tenets of risk innovation is that everyone has both the responsibility and opportunity to understand evolving risk landscapes from their own perspective.

It is also an approach that is highly effective in understanding and addressing institutional risk, including social and ethical risks that may lead to harm to institutions engaged in sensitive/controversial research such as CIO research if not taken seriously.

6. Institutional Risk

Where research is considered to be unethical, the reputations and careers of researchers involved may be put at risk. And where that research is sanctioned by their institution, and by funding and authorizing bodies, these are also potentially placed at risk if practices that are deemed to be unethical come to light.

This type of institutional risk is not well documented – especially around social media research – although there are a growing number of anecdotal cases where individuals and institutions have suffered because of publicly-raised concerns over research ethics. These include instances of research that is overtly funded by agenda-driven parties (such as industry) – especially where funding is not transparent; research that is perceived to cross ethical boundaries; research that incorporates potentially harmful bias; research that is perceived to promote injustice and inequity; and research that is perceived to be driven by a political or ideological agenda.

Supporting or engaging in research that is publicly perceived to be inappropriate can lead to substantial institutional risk, depending on the level of public concern expressed and how this plays out. And this risk becomes substantially elevated if the research may lead to social norms and expectations being overstepped, specific groups being preferentially advantaged over others, or fundamental and foundational shared values being threatened.

Institutional risks are rarely easy to parse out and navigate, which is one reason why work, around risk innovation for example, focuses on pragmatic approaches to helping organizations make informed decisions within a highly uncertain and subjective risk landscape (Risk Innovation Nexus 2021). They are, however, highly important to long term success. And increasingly, social media research is emerging as an area where there is potential institutional risk if care isn't taken.

Here, where research lies within ethically uncertain areas and touches on sensitive issues, it can potentially turn into a liability if not handled appropriately. This is especially the case where stakeholders perceive threats to areas of value outweighing potential benefits; where research is designed to lead to active interventions within communities that have not given consent; where interventions may be construed as causing harm; and where research and research outcomes are perceived as inappropriately disadvantaging some communities over others – especially on political or ideological grounds.

Within this landscape, institutions conducting and supporting social media research need to be especially cognizant of the potential risks they face if studies are publicly questioned – including the risk of defunding, censorship, or legal action.

The Court of Public Opinion

For better or worse, institutional risks are as much predicated on perception as they are on evidence. As a result, research institutions need to be aware of how the stakeholders they engage with – and who have the ability to influence what they do – perceive their actions and their consequences. It is rarely if ever sufficient to claim that you are behaving ethically and in the public interest if the prevailing public opinion suggests otherwise.

Such risks have been referred to as "orphan risks" as they are often overlooked by organizations, simply because they don't fit conveniently into a conventional risk management framework (Maynard 2018). However, they are often critical to an organization's success, or continued ability to operate.

Here, public opinion is everything. If there is widespread concern for instance that a government-funded research project is designed to support a particular political party or agenda, or to marginalize, disadvantage, or place in harm's way, specific groups of citizens, there is a high chance that the institution will be publicly called to account. Even if the intentions are not as they are being represented, there is a reasonable chance of jobs being lost, funding being withdrawn, and policies being put in place that restrict operations – especially if there is widespread public outrage which can be politically leveraged.

Navigating risks associated with public opinion can be challenging – especially from a researcher perspective where there is often a disconnect between research questions, processes and methodologies, and how these fit within a broader institutional context. This is where principles of scientific integrity can be helpful, such as those published by Kretzer et al. (Kretser, Murphy et al. 2019).

Reputational Risk

Perhaps one of the greatest yet hardest to manage risks that institutions face is reputational risk. This is often closely associated with public perception, but is more directly connected with key stakeholders such as funders, investors, customers and collaborators, and how their views and experiences influence their support.

Within the Risk Innovation framework this is represented by the orphan risk of Reputation and Trust – a risk that arises when institutions have only a rudimentary understanding of how their behavior and actions strengthen or weaken reputation and trust (Maynard 2015, Maynard 2018, Maynard and Garbee 2019, Risk Innovation Nexus 2021). As most successful commercial organizations recognize, reputation and trust are hard to build, and easy to lose. And research that is perceived as threatening what is important to key stakeholders and groups can very quickly become a liability that, in turn, threatens to undermine reputation and trust.

This can take on many forms. For instance, research that embarrasses funders or sponsors, or undermines their credibility, is likely to lead to a loss of trust and support. Similarly, research that undermines the credibility of collaborators and partners also runs the risk of strained and broken professional relationships.

These risks associated with trust and reputation may be direct – if a research organization directly threatens the values and plans of a funder through its actions for instance. But they can also be by association. For instance, if a research organization is held publicly accountable for what is perceived to be ethically questionable research in the mainstream media, or in front of congressional committees, funders and partners may back away from their commitments for fear of being implicated by proxy.

This is less likely to be an issue where research is of marginal public interest, or where there are clear social benefits to the work being undertaken. But where the benefits are unclear and the topics of study are socially and politically sensitive, risks may be amplified.

Social and Political Third Rails

In politics, the notion of a “third rail” refers to issues that are so controversial that they become in effect untouchable without severe adverse consequences. The metaphor comes from the third rail in some electric rail systems, where to touch it results in substantial harm or death.

As with other areas of research, social media research comes with its own third rails that have the potential to lead to severe and even terminal institutional risk if touched. Unfortunately, these are not always easy to see. And from the perspective of researchers who are less interested in the political landscape and more interested in the work they are engaged in, they can easily become buried – until touched. Yet when engaging in potentially sensitive research in ethically uncertain areas, institutions and researchers need to remain cognizant of social and political third rails, lest they inadvertently and naively embark on courses of action that can only conclude in significant harm to the organization and its members and associates.

The nature and type of third rail issues in social media research shifts with social norms and political landscapes. However, areas to be especially wary of include those that touch on civil rights and liberties, justice and equity, and the use and abuse of power (especially political). While these do not lie beyond the scope of social media research, extreme care is needed to ensure that naïve or misguided approaches to the formulation, execution and dissemination of research do not become a liability for researchers, institutions and the communities they interact with.

Within this landscape, the distinction between research designed to observe and understand behavior, and research designed to change or otherwise affect behavior, is also worth considering. While the former may be associated with substantial institutional risks when addressing contentious and controversial areas, the latter is likely to significantly elevate any potential risk. This is widely recognized within public health research where research-driven interventions such as vaccines, fluoride in water, and gun control, continue to present challenges to researchers and research institutions – even though the benefit-risk (or reward-risk) calculus is usually clear. It is less widely recognized in areas where interventions are being considered without either the professional ethics or the benefits-justifications of public health research. And within the realm of social media research, this extends to non-consensual interventions.

Non-Consensual Interventions

As has already been noted, the nature of social media research generally precludes informed consent, and instead relies on assumptions of consent. While this creates challenges where research potentially leads to subjects or the communities they represent being placed in harm, harm can often be minimized or eliminated through good study design where no interventions are planned. However, research that is predicated on interventions – whether to study the effect of perturbations on social media users and associated communities, or to develop ways of influencing or otherwise affecting behavior – presents its own unique challenges.

Given growing concerns over adverse personal, societal and political impacts arising from internet-mediated communication, engagement, and information dissemination, there is mounting interest in social media research that is explicitly designed to alter the behavior and beliefs of individuals. There are clear justifications for research in this domain that draw on arguments of public good, and that parallel public health research – for instance, research into harmful social media habits, management of online shaming and bullying, spread of misinformation, user-manipulation, and propagation of ideas and ideologies that are counter to civil and constitutional values. However, as these areas depend on underlying values that may not be universally shared, they may become contested and controversial – especially when subjects are used in research studies without their consent.

Where social media studies set out to actively influence social media users without their consent, extremely high standards of ethical research are needed to avoid crossing ethical lines or substantially increasing institutional risks – and here the two are likely to be tightly coupled. Unlike analysis of existing social media content, proactive engagement and manipulation of subjects lies beyond previous discussions around privacy and harm. Where no content exists until researchers have prompted it, responsibility for the consequences of subsequent actions lies with the researchers and their institutions. Here, a semblance of informed consent may be possible to elicit, with social media users being made aware of the nature of interactions and the intentions behind them. However, in studies that rely on deception, participants may be subject to harm from intentional actions that they have no agency to avoid. And this puts such research in ethically tricky waters.

Non-consensual interventions in social media research raise serious questions around harm and privacy that need to be addressed as research is planned and executed – as was intimated in Chapter 2 on mapping out the CIO research stakeholder landscape. Such research potentially places researchers and institutions at considerable risk where there are perceptions that specific values are being imposed on subjects in an effort to influence the ways they think and behave. This is not to say that such research should be prohibited, and there may be cases where the societal benefits far outweigh the potential risks. But the risks may be extreme – especially if interventions touch on politically contentious issues.

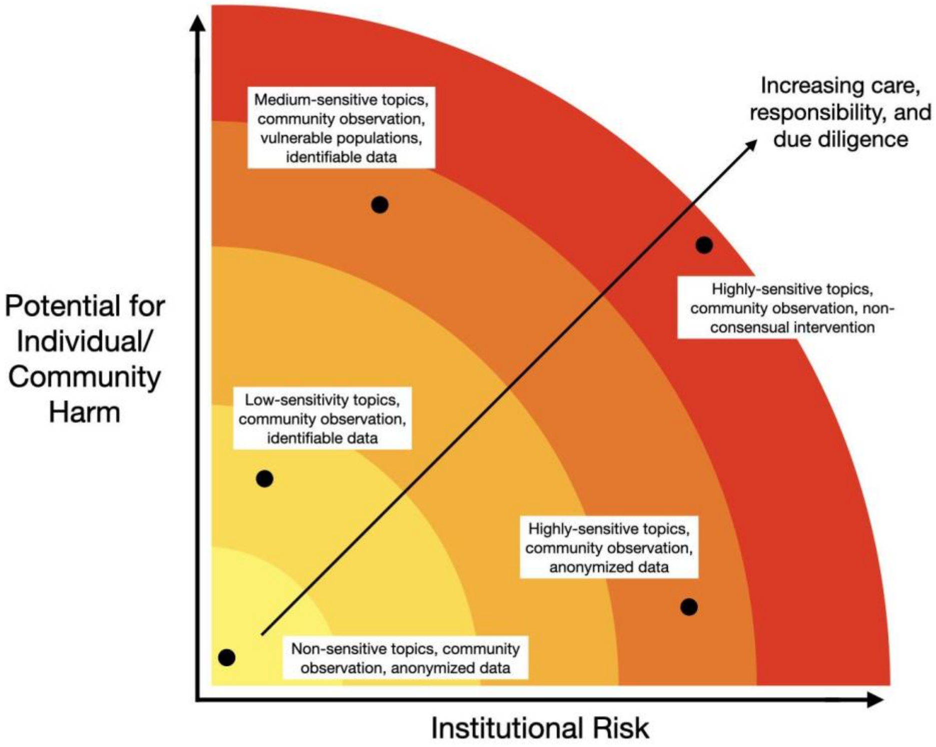

Potential for Individual/ Community Harm
Institutional Risk
Increasing care, responsibility, and due diligence
Medium-sensitive topics, community observation, vulnerable populations, identifiable data
Highly-sensitive topics, community observation, non-consensual intervention
Low-sensitivity topics, community observation, identifiable data
Highly-sensitive topics, community observation, anonymized data
Non-sensitive topics, community observation, anonymized data

Figure 3. A schematic representation of the relationship between institutional risk, potential for harm, and levels of care, responsibility, and due diligence, in CIO research.

7. Applying a Risk Innovation Framework to Counter IO Research

Risk innovation is based on a reframing of risk as a threat to value, both to a principal agent and to key stakeholders that are impacted by, and in turn have the ability to impact, an enterprise. Identifying stakeholders and understanding what is most important to them is foundational to a risk innovation approach. But this is not a one-time process. The Risk Innovation Framework encourages operationalizing stakeholder value at multiple points throughout a project. When practiced consistently over time, a project team can more successfully navigate risks and safeguard stakeholder value.

At each project phase, the Risk Innovation Framework encourages naming key stakeholders, identifying ethical considerations, assessing orphan risks, and implementing guidelines to help navigate obstacles and maintain integrity.

In Figure 4, we map the Risk Innovation Framework onto a CIO project and introduce tools and resources for practical application. Given how rapidly this field is evolving, this process guide should be treated as a guideline only, and freely adapted to specific circumstances and situations. It does, however, provide a useful starting point for applying a Risk Innovation Framework to CIO research.

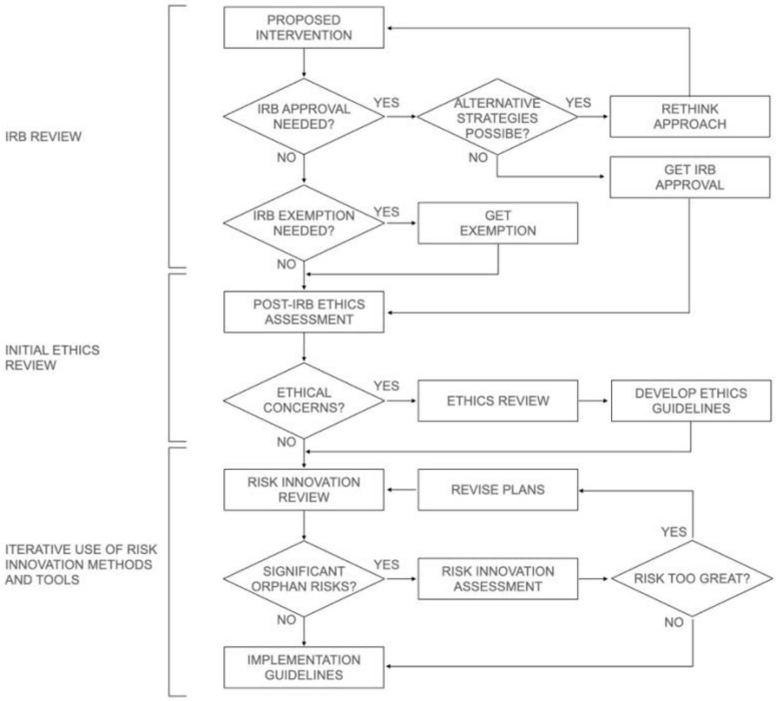


Figure 4. A conceptual decision-based flow chart for developing and implementing CIO research that meets appropriate IRB and ethical standards while navigating a complex risk landscape.

Complimenting the decision points outlined in Figure 4, our work on a practical instance of CIO research resulted in a step-based approach to conducting CIO research projects that is based on four project phases (Figure 5). Within each phase there are five identical steps, reflecting the iterative nature of applying risk innovation to CIO research.

This approach is particularly applicable to project-based research and development, where decisions and actions are driven by specific goals and outcomes. It is not necessarily generally applicable to all CIO research, but nevertheless forms an approach that is informative, adaptable, and useful for avoiding risks that may otherwise be overlooked.

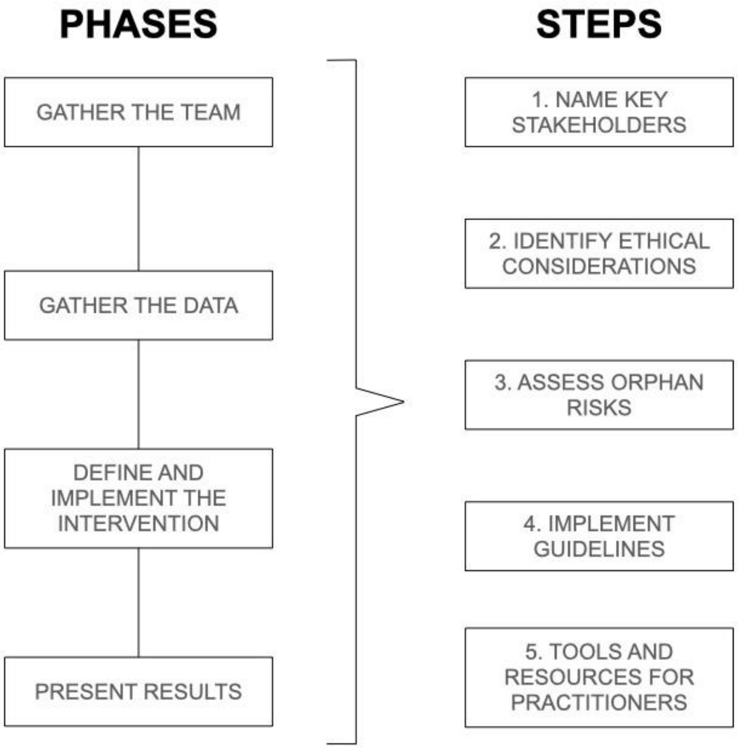


Figure 5. A phases and steps approach to applying a risk innovation framework to outcomes-oriented CIO research

We developed and tested the process above with a team comprised of engineers and data scientists developing and using a social media-modeled (but offline) CIO research platform. The team was specifically engaged in developing a sample intervention which would alert and stem the spread of false narratives on social media. This intervention used a chat-bot designed to engage with and develop a trusted relationship with participants, and through this to positively their behaviors and thinking.

While the research platform was designed to operate within a closed system, it was intended to provide insights on future platforms that potentially utilize public social media platforms. This, together with the use of human subjects in the research, made it especially useful for developing a risk innovation-based approach to CIO research.

An Example of Mapping the Stakeholder Landscape to Develop a Specific Countermeasure

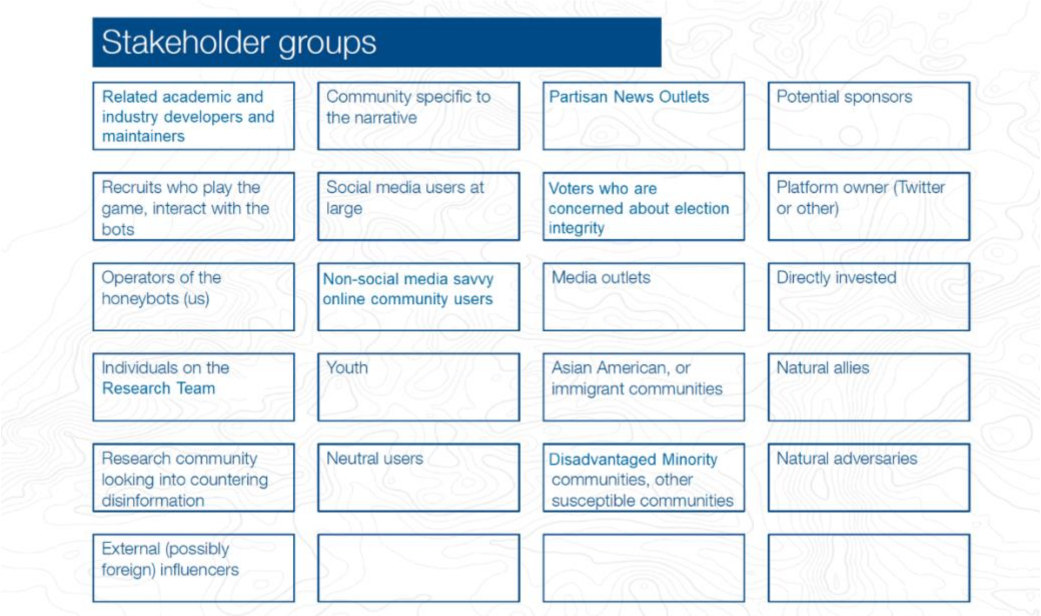
To illustrate the underlying process associated with applying this approach to IO countermeasures, we will show an example of how the risk innovation templates were used. In this case the hypothesis was to apply AI as a member of an online community which is being targeted with disinformation. The AI would be a resilient countermeasure, applied defensively, to alert the community to harmful narratives by engaging in discussions when disinformation topics arose. Upon request, the AI bot would provide rapid fact-checking and would maintain transparency by revealing that it was not a human user, but would do so in a natural, conversational way to maintain the uniqueness of the community, and present itself as a member with similar values and interests.

During the ideation phase, the research team was encouraged to identify and categorize key stakeholders, and to use these personas to imagine and explore potential risks and risk navigation strategies. For this publication, specific details have been altered in these figures as appropriate to protect the integrity and security of the specific research project these were developed for, but the essence of the completed templates remains the same.

As the team engaged in the risk innovation process the templates shown in the figures below were found to be extremely useful to help identify overlooked but still key stakeholders and ethical considerations related to their demographic. This partnership, between technical developers who are focused on applications and implementation, and social and political sciences enabled the engineers and data scientists who were unfamiliar with the thinking and processes behind risk innovation. Although new to exploring orphan risks, the team discovered gaps in their initial thinking and uncovered orphan risks through guided discussions who might potentially be impacted by their work and how, and potential consequences.

Figure 6 shows the results of the first brainstorming session where researchers were asked to think about who might have a stake in their research – in its conceptualization, its execution, its dissemination, and advice or actions arising from it. The template illustrates both the breadth of potential stakeholders and the rapidity with which the team were able to begin imagining a sophisticated stakeholder landscape beyond what they were typically used to.

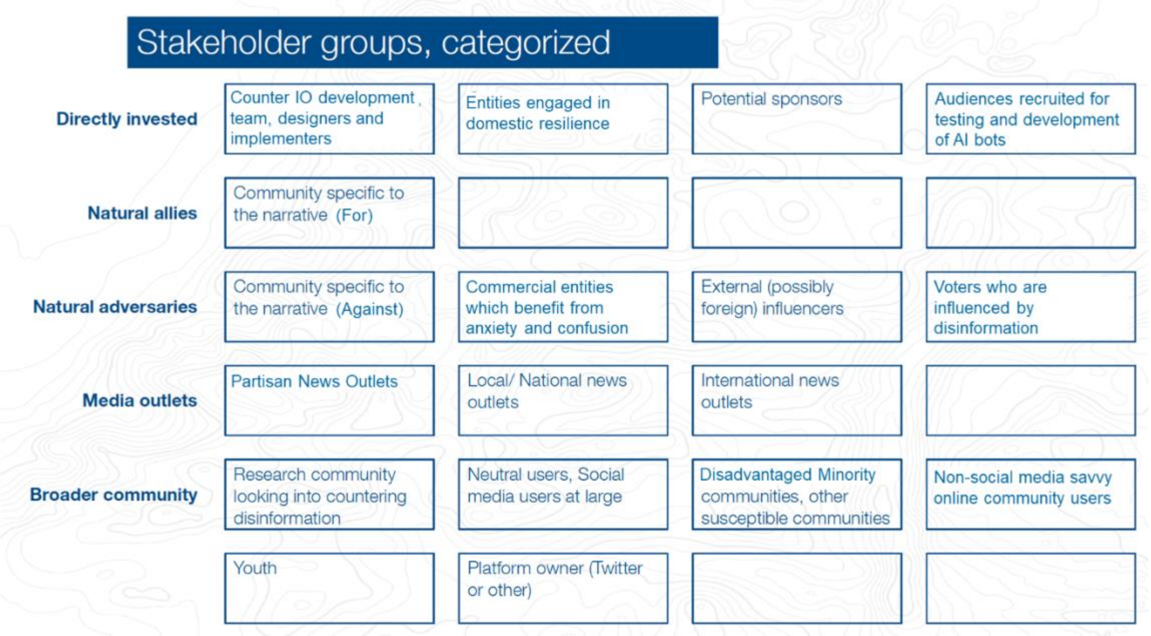
To help further define the stakeholder landscape, these stakeholders were loosely grouped into five categories, thus providing a broad perspective on communities that researchers needed to actively consider when developing and executing their research and disseminating the results (figure 7).



Stakeholder groups

Related academic and industry developers and maintainers	Community specific to the narrative	Partisan News Outlets	Potential sponsors
Recruits who play the game, interact with the bots	Social media users at large	Voters who are concerned about election integrity	Platform owner (Twitter or other)
Operators of the honeybots (us)	Non-social media savvy online community users	Media outlets	Directly invested
Individuals on the Research Team	Youth	Asian American, or immigrant communities	Natural allies
Research community looking into countering disinformation	Neutral users	Disadvantaged Minority communities, other susceptible communities	Natural adversaries
External (possibly foreign) influencers			

Figure 6. An example of a completed risk innovation template used by a CIO research group to identify key stakeholder groups potentially impacted by their work, and who in turn have the potential to impact the planned research and its impact. Researchers were asked to brainstorm stakeholders. Content has been altered where appropriate to protect confidentiality.



Stakeholder groups, categorized

Directly invested	Counter IO development, team, designers and implementers	Entities engaged in domestic resilience	Potential sponsors	Audiences recruited for testing and development of AI bots
Natural allies	Community specific to the narrative (For)			
Natural adversaries	Community specific to the narrative (Against)	Commercial entities which benefit from anxiety and confusion	External (possibly foreign) influencers	Voters who are influenced by disinformation
Media outlets	Partisan News Outlets	Local/ National news outlets	International news outlets	
Broader community	Research community looking into countering disinformation	Neutral users, Social media users at large	Disadvantaged Minority communities, other susceptible communities	Non-social media savvy online community users
	Youth	Platform owner (Twitter or other)		

Figure 7. Categorization of stakeholder groups identified in figure 6 as a critical step in exploring the stakeholder landscape around a proposed CIO research project. Content has been altered where appropriate to protect confidentiality.

While these templates formed part of the scene-setting, they were invaluable in ensuring that the process was streamlined and productive. They were also instrumental in developing the steps outlined in figure 5 for integrating a risk innovation approach into CIO research. This deeper dive, envisioning how specific individuals would react to the presence of an AI bot in online communities, helped to craft a number of stakeholder personas in more detail. These helped put a realistic face to stakeholders who might otherwise have remained an abstraction, and in turn guided discussions around possible ethical concerns and potential orphan risks. Finally, a User Story, which is an Agile software development method of illustrating in a few simple sentences the desired outcome for each stakeholder category, was generated for each category. These User Stories then formed a basis for the development of software requirements for countermeasure development, testing, and verification.

The complete approach is now presented as steps and further described below:

CIO Research Risk Innovation Phases and Steps

Gather the team

At its initiation, a project brings together its core team. This is a crucial first opportunity to apply the Risk Innovation Framework.

1. **Name key stakeholders:** At this phase, it's important to consider the internal team of researchers, engineers, data analysts, sponsors and decision makers.

2. **Identify ethical considerations:** Are multiple disciplines, demographics, or experiences represented? If not, can additional perspectives be gathered? What defines success and failure for each internal team member? Is there anything that would prompt an internal stakeholder walk away from this project?

3. **Assess orphan risks:** In particular, pay attention to**:** *Organizational values and culture* – Risks from tensions between business practices, both internal and external, and the set of values that reflect what is important to a business' founders and members. *Worldview* – Risks from people's deeply held beliefs about how they view the world and how it should function (Table 1).

4. **Implement guidelines:** Does your organization have institutional guidance, de facto norms, or legal requirements for human subject research? How does the ethics and values mindset of your organization or sponsor differ from your own?

5. **Tools and resources for practitioners:** Write down best- and worst-case outcomes for your internal stakeholders. As a team, discuss what steps you will take to navigate toward the best-case outcomes and avoid the worst-case outcomes. Optionally, take

advantage of the Risk Innovation Stakeholder Value Identification exercise to help name areas of value for each stakeholder (Risk Innovation Nexus 2019).

Gather the data

This phase ushers in a number of outside stakeholders, including entities and individuals not directly associated with the project.

1. **Name key stakeholders:** Internet, Data Storage and Management Entities; Social and Tribal Entities; and, again, Individuals internal to the project team who are directly responsible for the data.

2. **Identify ethical considerations:** How transparent is your data gathering process? Will you share the process with others? Where will your data be stored? Will others have access to your data? How might data gathering approaches potentially threaten the value of social and tribal entities?

3. **Assess orphan risks:** Consider in particular: *Privacy* – Risks from the social pitfalls associated with the use and misuse of an individual's data. *Loss of Agency* – Risks from products or business practices that reduce the ability of organizations and individuals to have agency. *Standards* – Risks from a business' lack of engagement with an evolving operational framework for businesses that spans legal requirements, informal guidelines, and norms and codes (Table 1).

4. **Implement guidelines:** How do you and your organization view data ownership; what privacy rights does an individual have once they've submitted their data and who is responsible for enforcing/guaranteeing those rights? How comfortable are you in using PAI datasets?

5. **Tools and resources for practitioners:** Continue to consider best- and worst-case outcomes and give name to stakeholder value. Take advantage of the Risk Innovation Planner as a place to note – and reference – your growing list of stakeholders and what is most important to them (Risk Innovation Nexus 2019).

Define and implement the intervention

This phase of the project is when the work really comes to life. Clarifying the larger context within which your project will operate allows the team to identify and navigate additional ethical issues. Understanding who is being impacted and how is key to moving forward ethically.

1. **Name key stakeholders:** Consider in particular: Media Entities; Science and Education Entities; Internet, Data Storage and Management Entities; Social and Tribal Entities;

and, again, Individuals internal to the project team who are directly responsible for the outputs.

2. **Identify ethical considerations:** Revisit the best- and worst-case outcomes that your internal stakeholders identified earlier in the process. Now that your intervention is more fully developed, are you still on track to navigate toward the best-case outcomes and avoid the worst-case outcomes? Have you considered your participants, their ability to consent to their role in your experiment, and the impacts - both real and perceived - your experiment might have on them?

3. **Assess orphan risks:** Pay especial attention to: *Geopolitics* - Risks from a lack of awareness of or strategies for navigating a shifting geopolitical landscape. *Loss of Agency* - Risks from products or business practices that reduce the ability of organizations and individuals to make decisions. *Perception* - Risks created from how people perceive a technology to impact/threaten what they think is important. *Social Justice & Equity* - Risks from business practices and technologies that marginalize or disadvantage specific segments within society. *Worldview* - Risks from people's deeply-held beliefs about how they view the world and how it should function (Table 1).

4. **Implement guidelines:** What external factors could come into conflict with the ethics of your experiment? Have you considered how your experiment may be perceived externally? What are the implicit and explicit risks to your organization, "tribe", and yourself personally? Can you articulate what constitutes "informed consent" for your experiment?

5. **Tools and resources for practitioners**: Continue to populate the Risk Innovation Planner (Risk Innovation Nexus 2019), paying specific attention to your growing list of orphan risks and the action steps you're taking to navigate the opportunities and challenges they represent. Review the hypothetical risk scenarios on the Risk Innovation Nexus website (Risk Innovation Nexus 2019) that illustrate the challenges presented by orphan risks within different contexts and try to draw connections between these hypothetical scenarios and your own project.

Present the results

As your project comes to an end, it's important to consider which outputs you choose to share and with whom. How can you best share your results without misrepresenting your work or your stakeholders, and how can you work to ensure that your results are misused by outside entities?

1. **Name key stakeholders:** Consider in particular: Federal, State and Local Governments; Media Entities; Science and Education Entities; Social and Tribal Entities; and Individuals internal to the project team who are directly responsible for presenting the

outputs as well as the Individual(s) funding the project.

2. **Identify ethical considerations:** Do you plan to release information that might put your organization or individuals on your team at risk? Do you plan to release information that might put members of your community at risk? Is this information you plan to release detailed enough that it could be used maliciously against you, your organization, or in a global or political context?

3. **Assess orphan risks:** Pay special attention to: *Bad Actors* - Risks from enterprises that behave in ways that are ethically questionable or that lead to unacceptable harm. *Geopolitics* - Risks from a lack of awareness of or strategies for navigating a shifting geopolitical landscape. *Loss of Agency* - Risks from products or business practices that reduce the ability of organizations and individuals to make decisions. *Reputation & Trust* - Risks from a business having only a rudimentary understanding of how their behavior and actions strengthen or weaken reputation and trust (Table 1).

4. **Implement guidelines:** Have you considered how your experiment may be perceived externally? What are the implicit and explicit risks to your organization, “tribe”, and yourself personally? How can your results be misused against (or, how can your results enable) disadvantaged groups or individuals?

5. **Tools and resources for practitioners**: Now is the time to review all of your stakeholders and what is important to them as well as the best- and worst-case scenarios you have captured throughout your process. Refer back to your Risk Innovation Planner and ensure your action steps are in alignment with stakeholder value and that you’re continuing to navigate orphan risks until your project comes to an end (Risk Innovation Nexus 2019).

8. Powers of Ten

As the background, framing and assessment above indicates, conducting socially responsible and ethical counter influence operations research is neither straight forward or formulaic. Rather, the frameworks, guardrails, considerations, processes, and checks and balances that need to be put in place, will differ widely across different contexts.

Here, making informed decisions is a process of asking relevant questions, being informed by relevant guidelines, drawing on critical resources, and developing a mindset that is agile and receptive enough to recognize and navigate orphan risks that could otherwise undermine research or halt it completely.

To aid in this process, this paper concludes with three chapters that provide guiding lists of questions, principles, and resources. They are not comprehensive and should be approached for what they are -- a guide and a prompt to support context-dependent approaches to responsible and ethical research. Nevertheless, they provide an effective starting point for anyone engaging in CIO research who wants to ensure that their work is responsible and ethical, and that it is not hampered by avoidable but often novel and easy to overlook risks.

9. Ten Guiding Questions

These ten questions are intended to guide researchers and research institutions in considering factors that may be easily overlooked in developing countermeasures yet may have a profound impact on how research is perceived, its social legitimacy and how effective and impactful it potentially is. They are also intended to help keep research within acceptable ethical boundaries, and avoid crossing ethical lines, whether intentionally or inadvertently.

The questions are not comprehensive but are intended to stimulate thinking and perspectives that will guide ethical and socially responsible research.

1. Does your organization have institutional guidance, de facto norms, or legal requirements for human subject research?
2. Can you distinguish between rule-based and principles-based approaches to scientific experimentation?
3. How does the ethics and values mindset of your research team, organization or sponsor differ from your own?
4. What external factors could come into conflict with the ethics of your experiment?
5. How do you and your organization view data ownership; what privacy rights does an individual have once they've submitted their data and who is responsible for enforcing/guaranteeing those rights?
6. How comfortable are you in using PAI datasets even if the acquisition of the set was not acquired in ethically normative and acceptable ways? (Note: this includes secondary use)
7. Have you considered how your experiment may be perceived externally? What are the implicit and explicit risks to your organization, "tribe", and yourself personally?
8. How can your results be misused against (or, how can your results enable) disadvantaged groups or individuals?
9. Can you articulate what constitutes "informed consent" for your experiment?
10. What actions, above and beyond IRB approval, have you implemented to ensure the safety and wellbeing of your participants and audience?

Discussion

Starting a new effort, especially in an area which blends science, technology, engineering, and mathematics (e.g., computer science, statistics) with the more qualitative social, behavioral, and political sciences can be bewildering. Adding to the complexity are the pressures and unwritten influences of sponsors and the organizations to which we belong. These Ten Guiding Questions provide the fundamental ethical questions that should be considered at program kickoff, regardless of whether ethics has been formally identified as a concern or not.

Unlike medicine and psychology, where the ethics of human subject testing is well documented and the ethical evaluation process institutionalized, CIO has yet to be formally recognized even though its immediacy, impact, and cost are well documented and validated by popular belief. With this in mind, responses to questions one through three may require some adaptation when applying rules meant to protect test subjects from harm or bias. For example, could repeated exposure to a “flat earth” narrative in testing and evaluation of a CIO approach make someone vulnerable to believing the earth is flat?

Rules versus principles-based approaches to ethics are an important distinction, often because the science and technology focused community cannot easily grok how anonymized data might cause harm to the original sources. In a rules-based approach we tend to decide if a rule does or does not apply to our project and, if the latter, end our consideration of consequences. However, in principle-based approaches (from which the rules initially stemmed) the consideration of consequences needs to be continually reassessed in a data-driven environment, especially due to the speed at which online information habits and patterns evolve. Thus the reader is encouraged to consider two points from which these questions arise: first, understanding the underlying principles essential to your stakeholders, and second, identifying gaps in organizational rules (which stem from those principles) which might emerge as you conduct your research. The guiding questions, listed above, serve as a means of eliciting the information needed to address both these points regardless of the stakeholder maturity in considering and documenting institutional ethics.

10. Ten Overarching Principles, Plus One

I/we:

1. Will not use unethical means to combat disinformation
2. Will not intentionally promote harmful stereotypes
3. Will adhere to institutional core values and IRB guidelines
4. Will exercise care when using public information which could characterize or identify individuals
5. Will maintain diligence in identifying and mitigating bias in algorithms and implementation
6. Will perform risk analysis to understand unintended consequences; including how the capabilities we develop could be weaponized or misused
7. Will not violate EULAs or User Agreements
8. Will conduct R&D with neutrality and without bias, and strive to remain apolitical
9. Will provide means for transparency and accountability
10. Will ensure that stakeholders, including researchers, are aware of their personal risks

Plus one: Will take reasonable precautions to prevent harm (as listed above) and, in the event that inadvertent harm occurs, will own and learn from it

Discussion

It is all too easy, in the course of normal work, to develop “technical tunnel vision” where the only objective is algorithmic, experimental, or demonstration results. The thought is “there is time after the milestone” to consider the ethical implications, or “it’s not within my purview of concerns - I’m just doing my job”. However, as history has demonstrated, ethics concerns everyone at every level, and the Ten Overarching Principles (plus one) can be easily adapted to a variety of project roles and functions.

The principles, as listed above, range from decisions and actions at the individual level (we are all capable of deciding for ourselves if something is ethical or not), to the team (performing a risk analysis and taking the time to discuss the implications), to leadership (ensuring that results are presented without bias), to the institution (checking that IRB rules and core values are adhered to during reviews).

However, regardless of good intentions and due diligence, it’s not possible to predict all the ways in which a capability could be misused in the future. The “plus one” in this case, takes into consideration the ten principles but prepares for potential harm and appropriate action by putting in place rules and practices that can evolve.

11. Ten Critical Resources

The resources below represent a “top tier” set of papers, reports, and books, that are useful for researchers, administrators, funders, practitioners, and others involved in CIO research. The list is not comprehensive, but it is an important starting point for helping to frame and guide ethical and responsible CIO research.

1. Recommendation on the ethics of artificial intelligence (UNESCO 2020)

Link: https://en.unesco.org/artificial-intelligence/ethics

Why it’s useful: the authors recognize the potential power that AI has in terms of disrupting (or enhancing) international relations. The need to remind nations of the importance of applying emerging technologies based on both international law as well as respect for social, economic, and minority equality was recognized in this document to raise awareness of the dangers of weaponizing AI.

2. Bit by Bit: Social Research in the Digital Age, Princeton University Press. (Salganik 2017)

Link: https://www.bitbybitbook.com/

Why it’s useful: In addition to a chapter on ethics, Matthew Salganik touches on the various aspects in conducting human-in-the-loop experimentation, laying out the conflict between data scientists and social and behavioral studies, and outlining the challenges faced when trying to draw inferences from big data.

3. Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0) (Markham and Buchanan 2012); Internet Research: Ethical Guidelines 3.0 (franzke, Bechmann et al. 2020)

Link: https://aoir.org/reports/ethics2.pdf

Why they are useful: These documents, from the Association of Internet Researchers, are often used to inform IRBs who are faced with understanding and adjudicating ethical practices for research which uses online participants and communities. The document defines “internet research” and provides a basis from codified policies for ethical standards.

4. The OKCupid dataset: A very large public dataset of dating site users (Kirkegaard and Bjerrekær 2016)

Link: https://doi.org/10.26775/ODP.2016.11.03

Why it’s useful: A call-to-arms for transparency in sharing datasets as well as a contrast of how scientific results, scraped from (relatively) publicly available data, can be weaponized. Taken from the dating website OkCupid, the authors were able to identify potential relationships in cognitive ability to social and behavioral patterns such as religious beliefs, political interest and participation. This paper demonstrates the high risk run by data and social scientists in both feeding bias, as well as the ethical questions of misleading the dating service users and violating terms of service. A cautionary tale.

5. It's time for tech startups and their funders to take "orphan risks" seriously. (Maynard 2018)

Link: https://medium.com/edge-of-innovation/its-time-for-tech-startups-and-their-funders-to-take-orphan-risks-seriously-3f7813976a07

Why it's useful: An applied view of why understanding risk is essential to technology development. A clear statement of how unplanned off-shoots can threaten key values, and why it can be a threat to national and cultural values. This resource also compliments the tools and resources on applied risk innovation that are available at http://riskinnovation.org.

6. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI (Madaio, Stark et al. 2020)

Link: https://doi.org/10.1145/3313831.3376445

Why it's useful: Explores the practical and competing challenges faced by practitioners and organizations in developing ethical guidelines for the development of AI enhanced systems. The authors explain their methodology to "co-design" guidelines that include and integrate both perspectives in an attempt to satisfy both practical needs as well as organizational principles for fairness.

7. Developing a framework for responsible innovation. (Stilgoe, Owen et al. 2013)

Link: https://doi.org/10.1016/j.respol.2013.05.008

Why it's useful: Another example risk/ethics framework based on "four integrated dimensions of responsible innovation: anticipation, reflexivity, inclusion and responsiveness" to complement Risk Innovation methods. This is a broader approach, and well suited to structure organizational standards for ethics and risk into a structure which can align with practical research guidelines.

8. Web-based Game "The Evolution of Trust" (Case 2017)

Link: https://ncase.me/trust/

Why it's useful: A representational model showing how to apply game theory to simulate trust, and a practical and clever visualization of the risks involved when trust breaks down. The game provides a template for simulating and better understanding how to model ethical decision making and the risks when "our environment acts against the evolution of trust".

9. Assessing and Evaluating Department of Defense Efforts to Inform, Influence, and Persuade. Desk Reference. (Paul, Yeats et al. 2015)

Link: https://www.rand.org/pubs/research_reports/RR809z1.html

Why it's useful: While there are only two mentions of the word "ethics" within this copious tome (one being in the index), there is a brief but useful section on preserving integrity, accountability and transparency in assessment. Note that in this context assessment pertains to the assessment of applied influence operations. This document provides an excellent reference for readers who need to understand the mindset of military uses of influence operations, and where ethics could be introduced, and how to present the ideas in a synergistic manner.

10. National Strategy for Countering Domestic Terrorism (Executive Office of the President National Security Council 2021)

Link: https://www.whitehouse.gov/wp-content/uploads/2021/06/National-Strategy-for-Countering-Domestic-Terrorism.pdf

Why it's useful: A clear and succinct statement of motivation - why there is a need to apply AI to counter Influence Operations and the visionary boundaries that must be respected in order to ensure democracy is preserved. The approach as laid out in this document "honors and protects" both America's security as well our civil rights and liberties, points which are at the heart of ethical considerations for countermeasure development.

In addition to the resources listed above it is strongly recommended that scientists and engineers become familiar with the IRB policies of their own organizations, as well as the legal and ethical guidelines of the platforms they will be interacting on (i.e., Twitter guidelines for experimentation) can be found here: https://developer.twitter.com/en/use-cases/do-research/academic-research)

12. Summary

Influence operations are well documented and a clear and present danger to the public. There is no recipe for conducting ethical counter influence operations, only continued and careful consideration of the research, the effect it could have, and how it may be perceived at each phase of development. This paper argues that academia and government research institutions are best suited to conduct this research responsibly and as an unbiased entity with technical expertise. From a government perspective, the authors outlined the need for counter influence operations research, the stakeholder landscape, the application of artificial intelligence, and the ethics and risks involved.

Although this paper does not propose a process to conduct ethical counter influence operations, it does provide a framework to evaluate and address the social, ethical and institutional risks of novel research and innovation. Core to the "risk innovation" approach taken is a discussion of guiding questions, overarching principles, and critical resources for institutions to evaluate risk from multiple perspectives and ultimately develop plans of mitigation. The discussion herein is meant to be thought provoking and raise awareness of the potential impacts of human subject research at a high level with broad impact, rather than focusing solely on the novelty and results of a research and development approach.

References

Algorithm Watch. (2020). "AI Ethics Guidelines Global Inventory." Retrieved December 25, 2021, from https://inventory.algorithmwatch.org/.

Barratt, M. J. and A. Maddox (2016). "Active Engagement withStigmatised Communities through Digital Ethnography,." Qualitative Research, **16**(6): 701-719.

Bauder, D., M. Liedtke and The Associated Press. (2021). "Whistleblower says Facebook routinely chose 'profit over safety' when it came to misinformation." Retrieved January 29, 2022, from https://fortune.com/2021/10/04/facebook-whistleblower-social-media-misinformation-hate-algorithm/.

BBC. (2021). "Twitter's algorithm favours right-leaning politics, research finds." Retrieved December 30, 2021, from https://www.bbc.com/news/technology-59011271

Bellamy, R. K. E., K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney and Y. Zhang (2019). "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." IBM Journal of Research and Development **63**(4/5): 4:1-4:15.

Benigni, M. C., K. Joseph and K. M. Carley (2017). "Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter." PLoS ONE **12**(12: e0181405).

Benjamin, V., J. S. Valacich and H. Chen (2019). "DICE-E: A Framework for Conducting Darknet Identification, Collection, Evaluation with Ethics." MIS Quarterly **43**(1).

Berghel, H. (2018). "Malice domestic: The Cambridge analytica dystopia." Computers in Human Behavior **51**(5): 84-89.

Bloustein, E. J. (1964). "PRIVACY AS AN ASPECT OF HUMAN DIGNITY: AN ANSWER TO DEAN PROSSER." New York University law review **39**(6): 962-1007.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle and J. H. Fowler (2012). "A 61-million-person experiment in social influence and political mobilization." Nature Biotechnology **489**(7415): 295-298.

Case, N. (2017). "“The Evolution of Trust” web-based game."   Retrieved January 29, 2022, from https://ncase.me/trust/.

Defense Innovation Board (2019). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense, Defense Innovation Board.

Deputy Secretary of Defence (2021). Memorandum: Implementing Responsible Artificial Intelligence in the Department of Defense U. D. o. Defence.

Edelson, L. and D. McCoy. (2021). "How Facebook Hinders Misinformation Research." Retrieved December 25, 2021, from https://www.scientificamerican.com/article/how-facebook-hinders-misinformation-research/.

Ess, C. (2002). Ethical decision-making and Internet research. Recommendations from the AoIR ethics working committee, AoIR ethics working committee.

Executive Office of the President National Security Council (2021). National Strategy for Countering Domestic Terrorism, US National Security Council.

Fiesler, C. and N. Proferes (2018). "“Participant” Perceptions of Twitter Research Ethics." Social Media + Society **4**(1): 2056305118763366.

franzke, a. s., A. Bechmann, M. Zimmer and C. M. Ess (2020). Internet Research: Ethical Guidelines 3.0, Association of Internet Researchers.

Heath and Human Services. (2018). "45 CFR 46 “Common Rule”."   Retrieved 4/11/21.

Hesse, A., L. Glenna, C. Hinrichs, R. Chiles and C. Sachs (2018). "Qualitative Research Ethics in the Big Data Era." American Behavioral Scientist **63**(5): 560-583.

Hicks, K. H., A. H. Friend, J. Federici, H. Shah, M. Donahoe, M. Conklin, A. Akca, M. Matlaga and L. Sheppard (2019). By Other Means. Part I: Campaigning in the Gray Zone, Center for Strategic & International Studies International Security Program.

Jasanoff, S. (2007). "Technologies of Humility." Nature **450**: 33.

Kirkegaard, E. O. W. and J. D. Bjerrekær (2016). "The OKCupid dataset: A very large public dataset of dating site users." Open Differential Psychology.

Kramer, A. D., J. E. Guillory and J. T. Hancock (2014). "Experimental evidence of massive-scale emotional contagion through social networks. ." Proceedings of the National Academy of Sciences **111**(24): 8788-8790.

Kretser, A., D. Murphy, S. Bertuzzi, T. Abraham, D. B. Allison, K. J. Boor, J. Dwyer, A. Grantham, L. J. Harris, R. Hollander, C. Jacobs-Young, S. Rovito, D. Vafiadis, C. Woteki, J. Wyndham and R. Yada (2019). "Scientific Integrity Principles and Best Practices: Recommendations from a Scientific Integrity Consortium." Science and Engineering Ethics **25**(2): 327-355.

Lewis, K., J. Kaufman, M. Gonzalez, A. Wimmer and N. Christakis (2008). "Tastes, ties, and time: A new social network dataset using Facebook. com." Social Networks **30**(4): 330-342.

Lima, C. (2021). "A Whistleblower's Power: Key Takeaways from the Facebook Papers." Retrieved January 29, 2022, 2022, from https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/.

Madaio, M. A., L. Stark, J. W. Vaughan and H. Wallach (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.

Marcellino, W., M. Magnuson, A. Stickells, B. Boudreaux, T. D. Helmus, E. Geist and Z. Winkelman (2020). Counter-Radicalization Bot Research, Using Social Bots to Fight Violent Extremism, Rand Corporation.

Markham, A. and E. Buchanan (2012). Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0), Association of Internet Researchers.

Maynard, A. (2018, September 16 2019). "It's time for tech startups and their funders to take "orphan risks" seriously." from https://medium.com/edge-of-innovation/its-time-for-tech-startups-and-their-funders-to-take-orphan-risks-seriously-3f7813976a07.

Maynard, A. D. (2015). "Why we need risk innovation." Nature Nanotechnology **10**: 730-731.

Maynard, A. D. and E. Garbee (2019). Responsibe innovation in a culture of entrepreneurship: A US perspective. International Handbook on Responsible Innovation: A Global Resource. R. Von Schomberg and J. Hankins. Cheltenham, UK, Edward Elgar**:** 488-502.

Maynard, A. D. and M. Scragg (2019). "The Ethical and Responsible Development and Application of Advanced Brain Machine Interfaces." J Med Internet Res **21**(10): e16321.

Mazarr, M. J., A. Casey, A. Demus, S. W. Harold, L. J. Matthews, N. Beauchamp-Mustafaga and J. Sladden (2019). Hostile Social Manipulation: Present Realities and Emerging Trends. Santa Monica, CA, RAND Corporation.

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

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1970). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research., U.S. Department of Health and Human Services.

New York Times. (2021). "Facebook sent flawed data to misinformation researchers." Retrieved December 25, 2021, from https://www.nytimes.com/live/2020/2020-election-misinformation-distortions#facebook-sent-flawed-data-to-misinformation-researchers.

Nissenbaum, H. (2004). "Privacy as contextual integrity." Washington Law Review **70**(1): 119-157.

Nissenbaum, H. (2009). Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford, Stanford University Press.

Paul, C., J. Yeats, C. P. Clarke and M. Matthews (2015). Assessing and Evaluating Department of Defense Efforts to Inform, Influence, and Persuade. Desk Reference, RAND Corporation.

Powers, S. and M. Kounalakis (2017). Can Public Diplomacy Survive the Internet? Bots, Echo Chambers, and Disinformation, US Department of Energy. **US Department of Energy Publications. 377.**

Risk Innovation Nexus. (2019). "Hypothetical Risk Scenarios." Retrieved January 29, 2022, from https://riskinnovation.org/resources/scenarios/.

Risk Innovation Nexus. (2019). "Risk Innovation Nexus Stakeholder Value Identification Exercise." Retrieved January 29, 2022, from https://riskinnovation.org/wp-content/uploads/2020/09/RiskInnovation_StakeholderValue_Activity.pptx.

Risk Innovation Nexus. (2019). "Risk Innovation Planner." Retrieved January 29, 2022, from https://riskinnovation.org/wp-content/uploads/2020/09/RiskInnovation_Planner_Template.pptx.

Risk Innovation Nexus. (2021). "Orphan Risks." Retrieved 4/11/21, from https://riskinnovation.org/think-differently/orphan-risks/.

Risk Innovation Nexus. (2021). "RISK INNOVATION NEXUS: Connecting ethical and responsible innovation with value growth." Retrieved 4/11/21, from https://riskinnovation.org/.

Salganik, M. J. (2017). Bit by Bit: Social Research in the Digital Age, Princeton University Press.

Sanders, B. (2019). "Democracy Under The Influence: Paradigms of State Responsibility for Cyber Influence Operations on Elections." Chinese Journal of International Law **18**(1): 1-56.

Sarawitz, D. (2016). "Saving Science." The New Atlantis Retrieved 4/11/21, from https://www.thenewatlantis.com/publications/saving-science.

Shrestha, R., K. Kafle and C. Kanan (2021). "An Investigation of Critical Issues in Bias Mitigation Techniques." arXiv **arXiv:2104.00170 [cs.LG]**.

Stilgoe, J., R. Owen and P. Macnaghten (2013). "Developing a framework for responsible innovation." Research Policy **42**(9): 1568-1580.

The United Nations (1948). Universal Declaration of Human Rights.

The White House (2017). National Security Strategy of the United States of America, December 2027, The White House, US Fedewral Government.

UNESCO (2020). Recommendation on the ethics of artificial intelligence., UNESCO.

White House (1995). Executive Order 12968: Access to Classified Information.

Williams, M. L., P. Burnap and L. Sloan (2017). "Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation." Sociology **51**(6): 1149-1168.

Zimmer, M. (2018). "Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity." Social Media + Society **4**(2): 2056305118768300.