Perceived Social Media Bias, Social Identity Threat, and Conspiracy Theory Ideation During the COVID-19 Pandemic

Kevin Craig
Auburn University
kac0117@auburn.edu

Valeria Sadovykh
University of Auckland
valeriasadovykh@gmail.com

Abstract

Social media organizations have an obligation to filter and sometimes exclude content, often based on machine learning algorithms. This has resulted in perceptions of bias in social media. When individuals perceive that a social media system is designed to exclude their point of view, they may experience a loss of self-worth, based on their excluded point of view. As a result, they may resist and avoid the technology that seems biased against them to prevent further loss of self-worth. They might also believe in conspiracies about why social media is marginalizing their point of view and find new self-worth as a conspiracy theorist. Data from 225 individuals who are interested in the risks associated with vaccines indicate that Perceived Bias presents a Social Identity Threat, which, in turn, is associated with Resistance to IT and Conspiracy Theorist Ideation.

1. Introduction

This article addresses perceptions of bias in social media and the conspiracy theories that arise when individuals perceive bias against their point of view. Conspiracy theories are not new. They have been observed whenever 1) social events have shifted power and resources in society, and 2) individuals had opportunities to imagine and propagate false motives that could serve as alternative explanations of those events [1]. What is new is modern technology's ability to rapidly generate power imbalances in obscure ways, creating the opportunity for more and more conspiracy theories.

The context for this work is social media during the COVID-19 pandemic, which presents an important opportunity to observe individual reactions to perceived social media bias in a time when information propagation is critical. For better or worse, many people rely on social media for news and information. As a result, social media sites such as Facebook, Twitter, and YouTube have political and financial power through their ability to control content. These sites have a social obligation to block harmful content and are motivated by profit to curate beneficial information. This is done through machine learning (ML), in which algorithms are developed and refined to categorize social media posts as harmful or beneficial. Herein lie some challenges. Individuals read popular press articles about bias in the ML process [e.g., 2], and the conclusions drawn from those articles and their own experiences with ML in social media may have social consequences. Thus, the two conditions for conspiracy theory emergence are met: social media is shifting power in the form of access to information, and the means by which it does so are opaque, leaving room for conspiracy theories to form.

ML often relies on human supervision, as panels of individuals sample social media posts and categorize them. The judgment of these panels is then used to evaluate and refine algorithms used to determine what is shown on social media sites, and to whom it is shown. Because these panels (and the development teams that employ them) are not perfectly inclusive, it is inevitable that they are the source of some bias [3]. In spite of efforts to reduce bias and transparency in this process, conspiracy theories have emerged that bias is actively employed by social media companies for political and financial gain.

This work is based on Social Identity Theory [4] and the Identity Threat Framework [5]. In this context, "identity" refers to beliefs held by an individual about who they are [6]. The study presented here explores how the power associated with ML can threaten one particular type of identity: social identity, which is composed of beliefs about the self as an individual member of a social category.

Social Identity Theory proposes that individuals form "ingroups" based on shared characteristics such as nationality, religion, and political affiliation. Members of ingroups categorize others as members of "outgroups," and often perceive phenomena that redistribute power and wealth as tools used to benefit outgroups at the expense of the ingroup [4]. In the context of social media filtering, it may be that individuals who affiliate with an unpopular opinion (opposition to vaccines, for example) see themselves as members of a social group that is de-valued and

URI: https://hdl.handle.net/10125/80066
978-0-9981331-5-7

marginalized by a process that does not include or even consider their point of view. If these individuals are not aware of the motives and methods behind social media filtering, they might seek out and support conspiracy theories that offer comfort in the form of validating and celebrating their own point of view.

The Identity Threat Framework [5] predicts how individuals react to a system that seems to promote an outgroup at the expense of their ingroup. This work offers a model based on that framework, testing the relationships among Perceived Bias, Social Identity Threat, Resistance to IT and Conspiracy Theorist Ideation. The model's hypotheses are drawn from disparate works in IS and our reference disciplines. However, this work combines them in ways that create new knowledge.

The overarching theme of this work is that: 1) Individuals may perceive bias in social media against an opinion they share with others; 2) this threatens value to society as one among those who hold an unpopular point of view, lowering their self-worth; 3) this causes them to resist and denigrate the social media in order to protect self-worth; and 4) this may cause them to seek a new source of self-worth in the form of idealizing the self as a conspiracy theorist. Each of these logical steps is sourced from the literature.

The identity threat literature and IS have proposed that people lose self-worth, a type of self-esteem, when they are marginalized by virtue of membership in a stigmatized social category [7, 8]. The identity threat literature has proposed that a reconstruction of identity (i.e., the addition of new beliefs about the self) may restore self-worth [5]. Finally, conspiracy theory research has proposed that self-worth emerges from idealizing the self as a conspiracy theorist [1, 9, 10]. This work integrates and extends these theoretical propositions, along with the logic of counter-normativity, to connect social media bias with conspiracy theorist development.

Further, the development of ML for uses such as social media filtering is associated with bias, and bias has been associated with social identity, which has been studied as the source of self-worth [11]. However, while industry seeks advice on how to mitigate the harm associated with bias, the effects of ML bias on social identity has never been studied as the cause of coping behaviors such as resistance to information technology (IT), or on conspiracy theories. The bias in ML results in a power imbalance between the ingroup whose opinions are excluded and the outgroup that implements ML.

Data collected from 225 individuals in the context of COVID-19 vaccine information propagation indicates that individuals who perceive bias against their point of view on social media report higher levels of Social Identity Threat, Resistance to IT, and Conspiracy Theorist Ideation.

2. Literature Review

2.1. Social media and ML-based filtering

Before the emergence of social media, individuals relied on print and broadcast media for information, and the editors of such outlets often fell under the suspicion of ideological bias, reflecting a "desire to affect reader opinions in a particular direction" [12].

As a result, social media is appealing to those individuals who have felt disenfranchised by ideological bias in traditional media. These individuals have turned to social media as a source of information that is not under the control of (and thus not subject to the biases of) editors [13]. Social media, however, may have its own biases, as the result of algorithms used to sort and filter content [14].

With the rapid growth of volume, variety, and velocity of data [15], statistical algorithms built to apply ML are becoming more difficult and complex for organizations to manage. To offset this, ML is often allowed to manage content with minimal human supervision, which can lead to the unintended filtering of content. Organizations face trade-offs with the possible consequences of greater reliance on artificial intelligence (AI), which can be prone to bias, often due to a lack of diversity in ML development [16].

This can have terrible ramifications, for example in the case of healthcare institutions that rely on predictions of AI and ML models, models that have been associated with biases such as those based on race. For example, in 2019 a study was published in the Washington Post that discussed Optum Health-Care's algorithm, which gave preference to white patients over black patients who were more ill [17]. As a result of the high visibility of such mistakes and the difficulty inherent in communicating how ML works, the ML aspect of social media systems generates the opportunity for individuals to imagine motives behind filtering; such an opportunity is a facilitating condition for conspiracy theorizing.

Many people are aware of the fact that online content is filtered, sorted, and distributed according to user characteristics such as demographics and opinions [18]. The process of using ML to filter content can be opaque and flawed. This creates a foundation for users to stop trusting online content due to a lack of transparency and well-publicized examples of bias. As a result, conspiracy theories can be elevated to the level of credibility formerly only associated with professionally edited news outlets; as the World Health

Organization has termed it, society now faces a flood of misinformation [19].

Due to the fact that much of society has lost trust in organizational AI services, many institutions and world leading organizations are developing and applying mechanisms to de-bias AI systems. For example, PwC released a Responsible AI framework in response to a Federal AI Standards Engagement Plan with the National Institute of Standards and Technologies (NIST) that provided information on Artificial Intelligence standards [20]. The framework focuses on AI technical standards to support a reliable, robust, and trustworthy build and usage of AI systems. It is a practical guide that focuses on five dimensions of control for "Responsible AI." It looks at the tool or service governance mechanism, its compliance with ethics and regulation, assessment of robustness, security, and privacy, as well as interpretability and "explainability" of decision models and what drives prediction and uncovers biases in the underlying data and model development [20].

2.2. Social identity threat and resistance to IT

Characteristics shared among individuals often form the basis for inclusion in a social category, and each individual who shares those characteristics realizes self-worth based on the value society places on that social category [21]. Individuals may self-categorize, and they also may find themselves categorized by the perceptions of others [21]. The basis for social categories can be as broad and involuntary as ethnicity [22] or as narrow and voluntary as holding a shared opinion [23].

Membership in a social category, when recognized by an individual, is the source of beliefs about the self, collectively termed social identity [24, 25:31]. Social identity is a source of self-esteem in the form of self-worth, or how valued an individual believes themselves to be by society as a member of a social category.

Social Identity Threat is a phenomenon that arises when individuals are disparaged or disempowered by powerful and influential groups or individuals on the basis of social category [7, 21]. The effect of this disparagement and disempowerment is to reduce the value of the categorized individual in society, thus jeopardizing their level of self-worth [26].

An identity threat is any phenomenon that harms identity by reducing valued beliefs about the self, causing individuals to experience reduced levels of self-worth, a type of self-esteem [5, 26], and the type of self-esteem associated with social identity is self-worth [11]. Thus, when a social category loses value in society, an individual member of that category will experience a loss of the self-worth associated with that social category. To cope with this, they are likely to take action to protect or replace the value associated with their social category [4, 5, 26]. Petriglieri's Identity Threat Framework [5] summarizes the research surrounding this phenomenon and proposes that individuals cope with such a threat by seeking experiences that protect, restore, and replace the self-worth associated with identity (Figure 1).

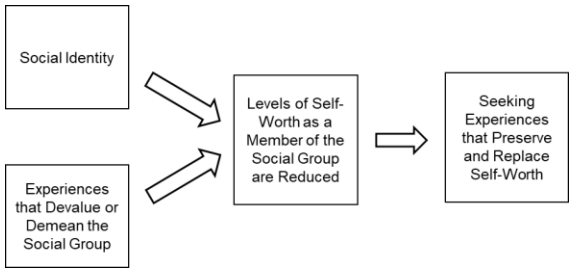

Figure 1. The Social Identity Threat Process

The behaviors proposed in the Identity Threat Framework include avoiding and derogating the threat source to preserve self-worth and constructing a new identity to generate self-worth [5].

2.3. Conspiracy theories and self-worth

Existing research has associated conspiracy with social identity and the appeal of non-conformity when predominant opinions are in conflict with shared individual beliefs [10]. Faced with the loss of self-worth due to prevailing societal norms, some adopt a conspiracy theorist identity as an alternative source of self-worth [1]. When stressed by the societal impacts of phenomena such as disease, those who speak out against the experts often frame conspiracy theorists as "lone crusader[s] for the truth against a… conspiratorial scientific establishment" [27, 28:463]. This is particularly true in the context of scientific knowledge, which often elicits beliefs that elites are controlling information for their own benefit. This was observed during the H1N1 pandemic of 2009-2010, when some believed that the pharmaceutical industry was manipulating information for "villainous" ends [28]. This matches the pattern described by the Identity Threat Framework, in which derogating the source of threatening information as villainous would protect self-worth by eroding the credibility of a threat source [5].

A further similarity between the psychology of conspiracy theory and the Identity Threat Framework is how each predicts the development of identity as a means to generate self-worth. In the case of conspiracy theory, this identity is based on the value an individual may hold as one who sees deeper "truths" that are

invisible to the public. It has been observed that when an individual recognizes the validity (imagined or not) of a single conspiracy theory, and demonstrates the courage to support that conspiracy theory against prevailing norms, they then recognize and support a wide range of conspiracy theories [29]. This investment in a range of conspiracy theories reflects an ideation of the self as a conspiracy theorist, and this phenomenon is known as "Conspiracy Theorist Ideation" [9]. The shared logic between this and the Identity Threat Framework is that both describe an expansion of self-beliefs to compensate for harm to self-worth. A deeper explanation of this identity development process is offered by conceptualizing Conspiracy Theorist Ideation as the product of "counter-normative identity," an identity that is defined by opposition to popular social norms [30, 31].

2.4. Counter-normative identity

When an individual perceives conflict between popular societal norms and their own beliefs, they may engage in "counter-normative" behavior. This is a reaction designed to generate self-worth through the rejection of normative behavior [31]. This rejection could be accompanied by claims that normative behavior is inauthentic and motivated by weaker individuals' need to conform to society. It could also be supported by a claim of authenticity by virtue of being true to a consistent self that does not change under pressure. Because of these two things, counter-normative behavior may generate self-worth by signaling a lack of weakness and the strength needed to maintain authenticity in the face of pressure to conform.

It has been proposed that counter-normative behavior is actually the expression of a counter-normative identity [31]. Such an identity would be marked by a rejection of popular beliefs. The development of this identity is a reaction to attacks on self-worth that are the result of holding an unpopular opinion; the counter-normative identity generates self-worth by embodying a rare and courageous point of view [32] (Cambon et al., 2006). A similar phenomenon has been observed among conspiracy theorists, who gain self-worth from the belief that "I know things they don't know" [9, 33] (Imhoff & Lamberty, 2017; Lantian et al., 2017).

Counter-normativity is not inherently harmful; sometimes it is the force that expands or overthrows standards of normal behavior that are narrow and harmful [34]. At its best, counter-normativity can be an attitude taken in advance of positive social change, promoted by individuals who are aware of the genuine negative consequences of un-checked social norms. For example, Sparkman and Walton [30] describe the emergence of the counter-normative behavior of eating less meat among those who are aware of the ecological implications of raising livestock. However, when individuals expand identity on the basis of a counter-normative attitude, they may find themselves in opposition to social norms that they would otherwise find beneficial. The expansion of identity to include Identity Theorist Ideation may be one such example.

3. Research hypotheses and model

Drawing from the Identity Threat Framework [5], our conceptual model addresses perceptions of people categorized by association with an unpopular opinion (in this context, that vaccines may be dangerous). This research begins with perceptions of an IT as the means by which an outgroup of powerful normative-conforming people and institutions exert control over others by limiting access to information ("outgroup control").

The dynamic between an ingroup and an outgroup is often perceived by members of an ingroup as a "zero-sum game," in which any power and resources gained by an outgroup must come at the expense of the ingroup [35, 36]. This results in a threat to social identity because people may perceive the IT as changing social structure in favor of others at their expense, diminishing the power and value of their ingroup. As a tool that reduces the power and resources of the ingroup, the IT poses an identity threat by reducing the value and worth associated with a social identity (thus reducing self-worth).

The right side of our conceptual model explains how individuals react to an identity threat. They may try to protect an identity that generates self-worth by taking power away from a threatening IT. This might be accomplished by persuading the self and others that the threatening IT is itself a negative phenomenon. A threat to self-worth is reduced when the source of that threat is not respected [37]. In addition, the source of threat may be avoided or sabotaged to reduce that IT's effect on the self and society.

People also may recoup self-worth by expanding and extolling the values of their threatened identity. This expanded identity may be counter-normative, generating self-worth to compensate for that lost by the IT's effects on the self and society. From this model (Figure 2), we develop our hypotheses.

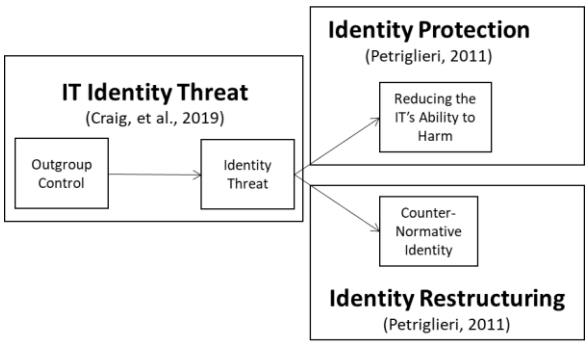

Figure 2. Conceptual model

Our hypotheses are based on the logic of our conceptual model, with specific and measurable constructs and relationships. Figure 3 illustrates our hypothetical model.

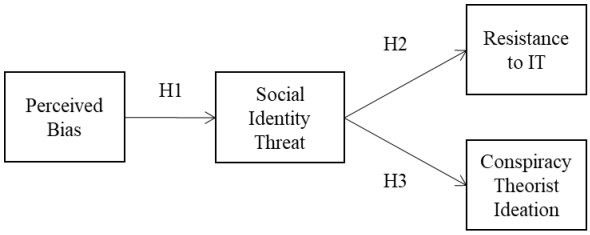

Figure 3. Hypotheses model

H1: The Product of Outgroup Control

Perceived Bias, defined as "the extent to which a system seems to affect reader opinions against the observer's opinions," captures the spirit of outgroup control in our conceptual model. When a social media organization influences public opinion in a way that demeans the opinions of an ingroup, the opinions of that ingroup are valued less by society. As a result, individual members of that ingroup recognize a loss of their value to society, i.e., their self-worth. This is Social Identity Threat, defined as "a reduction of how valued the individual feels they are as a member of a social group." Thus:

H1: Perceived Bias will positively relate to Social Identity Threat

H2: Identity Protection

We define Resistance to IT in the broadest possible way, to capture the full range of identity protection behaviors proposed by the Identity Threat Framework [5]. Thus, it is defined as "derogation of and opposition to an information system." We hypothesize that this is the result of the threat implied by the loss of self-worth reflected by Social Identity Threat. When individuals experience a loss of self-worth as the result of an identity threat, they anticipate further loss of self-worth from the source of that identity threat [26]. Derogation of the threat source is one way to reduce this anticipated loss of self-worth. If people disrespect a system, that system's ability to harm social identity is reduced, because the ability of any phenomenon to reduce self-worth is limited to the extent to which that phenomenon is respected [37:28, 38]. Derogating the social media organization that is harming social identity may reduce the respect associated with that social media organization, which would reduce its ability to further harm social identity. Thus, individuals are motivated to reduce their own and society's respect for the source of Social Identity Threat by thinking and speaking ill of it.

Opposition to an information system represents another way to reduce further harm to self-worth from a social media organization. Identity threat depends on exposure to a source of reduced self-worth [5]. If an individual avoids or successfully undermines or avoids a social media system, then that system can no longer harm identity. Thus, individuals are motivated to resist that system by avoidance and opposition:

H2: Social Identity Threat will positively relate to Resistance to IT.

H3: Expanding Identity to Compensate for Lost Self-Worth

Conspiracy Theorist Ideation, defined as "the adoption of theories that explain important events as secret plots by powerful and malevolent groups" [1] is thought to provide self-worth [9, 33]. This may compensate for the self-worth lost when a social media system harms the prestige of the ingroup. Psychologists have described a process that leads to "conspiracy ideation" in which individuals develop a propensity to reject scientific consensus [23]. It is thought that experiences that lead to the acceptance of one conspiracy theory change individuals' attitudes toward conspiracy theories in general [1, 23]. Taken together with the logic of identity restructuring to compensate for lost self-worth [5], we hypothesize that beliefs about the self as a conspiracy theorist are formed in reaction to Social Identity Threat:

H3: Social Identity Threat will positively relate to Conspiracy Theorist Ideation

4. Research method

Our unit of analysis is the individual, and our population of interest consists of individuals who have varying opinions about COVID-19 vaccines, and have

varying levels of Conspiracy Ideation. Our measures consisted of 5-point Likert scales of agreement, and can be found in Table 3. For Perceived Bias, we relied on scales from Eveland and Shah [39] and Gibbon and Durkin [40]. For Social Identity Threat, we used the social identity dimension from Craig, et al. [26]. For Resistance to IT, we used items adapted from Kim and Kankanhhalli [41] for opposition and avoidance, and items from van Prooijen [42] for derogation. We measured Conspiracy Theorist Ideation with items from Stojanov and Halberstadt [29].

Following pre-tests, we conducted two pilot tests with a subject pool of individuals using Amazon Mturk to recruit subjects. Next, we conducted a third pilot test with responses from 90 students from a large public university located in the Southeastern United States. Subjects for this pilot test were recruited by email and motivated by interest in the topic of our study. Qualitative data was informally collected during each pilot test and used to improve our study.

For our formal study, 225 completed surveys were collected from a panel through the firm, Qualtrics. Each subject answered "Yes" to the question "Do you enjoy reading stories or social media posts about the risks associated with vaccines?" This filtering question was designed to provide a subject pool of individuals who would have opinions, and possibly feelings of social affiliation, regarding vaccines. Before seeing the rest of the survey, subjects read a paragraph about social media filtering and were shown three controversial tweets expressing COVID-19 suspicion.

As shown in Table 1, skew and kurtosis were within the generally accepted threshold of +/- 1 [43]. Regarding reliability, we measured the alpha for each construct and none were below the cutoff value of 0.8 [44]. To establish discriminant validity, we compared the average variance explained by our items with construct correlations (see Table 2 and Table 3) [45].

Table 1. Mean, Std. Deviation, Skew and Kurtosis

	Mean	Std. D.	Skew	Kurt.
P. Bias	3.497	1.055	-0.428	-0.398
SID.Thr	3.081	1.116	-0.022	-0.781
Resist.	3.099	1.054	-0.077	-0.622
Consp.	3.655	0.867	-0.312	-0.234

Table 2. Construct Correlations, with AVE in the Diagonal

	Resist	Cons. Id.	SID. Thr.	P. Bias
Resist	**0.745**	0.414	0.546	0.5
Cons.	0.414	**0.751**	0.45	0.539
SID. Thr.	0.546	0.45	**0.788**	0.519
P. Bias	0.5	0.539	0.519	**0.733**

Table 3. Item Loadings (Principal Component Analysis, Varimax Rotation with Kaiser Normalization)

Perceived Bias, preceded by "Please rate your agreement with these statements about Twitter."	
Blocking these posts is biased against my views about vaccines.	0.785
Blocking these posts prevents people from sharing my views about vaccines.	0.723
This social media site is biased against my views about vaccines.	0.687
Social Identity Threat, preceded by "For these questions, "peer group" refers to people who share your beliefs about vaccines."	
Embracing Twitter makes me feel less respected by others in my peer group.	0.84
If I seem to support Twitter, I feel that others will consider me to be a poor member of my peer group.	0.812
Cooperating with Twitter makes me feel less worthy to belong to my peer group.	0.787
If I use Twitter with enthusiasm, people in my peer group will lose respect for me.	0.777
If I enthusiastically cooperate with Twitter, I feel that others will respect me less as a member of my peer group.	0.719
Resistance to IT, preceded by "How much do you agree with these statements about you and Twitter?"	

People are not learning the truth from Twitter.	0.801
I will avoid using Twitter.	0.772
Twitter is an important cause of harm.	0.76
If I could, I would try to undermine Twitter.	0.719
People are not learning from Twitter.	0.708
If I could, I would obstruct Twitter.	0.703
Conspiracy Theorist Ideation	
Some things that everyone accepts as true are in fact hoaxes created by people in power.	0.788
Many so-called coincidences are in fact clues as to how things really happened.	0.768
The alternative explanations for important societal events are closer to the truth than the official story.	0.752
Events which seem to lack a connection are often the result of secret activities.	0.737
There are secret groups that greatly influence political decisions.	0.705

To satisfy the assumptions for regression, Q-Q plots were drawn for all variables and these indicated normal data. Mahalanobis distance and standardized deletions were used to identify multivariate and univariate outliers. Five multivariate outliers were identified, leaving us with 220 usable cases (112 male, 102 female, 6 other/prefer not to answer; minimum age of 18, maximum age of 84, mean age of 41). All hypotheses were tested using SPSS v26 and ordinary least-squares regression [46].

5. Results

As detailed in Table 4, the data supports our hypotheses at the $p < 0.001$ significance level, so we are highly confident that perceptions of bias in social media are associated with Social Identity Threat, as measured by a loss of self-worth. Likewise, individuals who experience a loss of self-worth in the context of bias against their opinions over a form of social media are more likely to resist that social media as a technology, and more likely to embrace a wide range of conspiracy theories, indicating an idealization of the self as a conspiracy theorist.

Table 4. Hypothesis Results (Controlled for Age and Gender)

H#	IV	DV	B	St.E.	t	Sig (p)
H1	P. Bias	SID Thr.	0.533	0.061	8.746	<0.001
H2	SID Thr.	Res.	0.591	0.059	10.073	<0.001
H3	SID Thr.	Cons. Id.	0.370	0.054	6.849	<0.001

The R^2 values for the dependent variables were 0.378 for H1, 0.353 for H2, and 0.232 for H3. Because Social Identity Threat serves in a mediating position in our model, we performed post-hoc tests, using the bootstrap method [47]. Mediation was supported at the 99% confidence interval for both dependent variables.

6. Discussion

Broadly speaking, among those who perceive a bias in the development and deployment of ML in social media against the social groups to which they belong, there is a likelihood that they experience a loss of self-worth. This loss of self-worth is likely to be accompanied by opposition to a social media that seems to discriminate against their point of view, and positive beliefs about the self as one who appreciates the value of conspiracies.

6.1 Contributions to Theory

This work opens new ground for exploring the effects of social media and perceptions of its biases, and it does so by drawing on existing literature and combining current theory in new ways. It establishes a connection between bias and resistance to IT and sheds light on how ML can contribute to the formation of conspiracy theories. It also introduces the concept of counter-normative identity to IS research.

Specifically, we offer theoretical support and evidence that perceptions of bias against an individual's point of view can make that individual feel less valued. It also reinforces prior work indicating that when individuals feel threat to their self-worth from an information system, they are likely to resist that system. Finally, we provide evidence that the de-valuation of an individual by an information system is likely to be associated with the development and propagation of conspiracy theories as a reaction.

6.2 Implications for Practice

Conspiracy theory development depends on gaps in public knowledge; without such gaps, there is no logical space for imagined motives and alternative explanations for events. Thus, it is advisable for social media firms to engage the public with transparency and inform the wider world about their efforts to address bias in their systems.

Some corporate institutions already have well-defined managerial processes that may apply to ML bias. However, many organizations now realize that their traditional management approaches must change to address current technology. Unlike traditional software development, ML requires iteratively driven development to address challenges such as the need for stewardship and governance practices to maintain effective oversight and transparency as algorithms evolve.

Based on our data indicating that ML algorithms may be tied to conspiracy theories through social processes, organizations should recognize that the social implications of their ML development represent an important source of risk. Organizations should thus focus on the efforts to control the social risks of ML deployment and work to ensure credibility. Placing emphasis on regulation, governance and transparency within data, analytics, and AI-driven decision making-solutions may help address that risk. Specifically, organizations should implement frameworks to embed ethics into the ML development process

One such framework is described by Felzmann, et al. [48]. This framework (see Figure 4) integrates transparency into the development of ML systems [48]. This goes from the first step, an approval of the initiation of having AI in place and check of its ethicality, as well as provide transparency of the data usage, test, processing and analysis.

Enterprises need to focus on how they govern AI systems and the associated data. With the attentive governance over ML development processes and how those processes are communicated to the public, social media organizations can reduce the harm caused by conspiracy theories related to how they filter content.

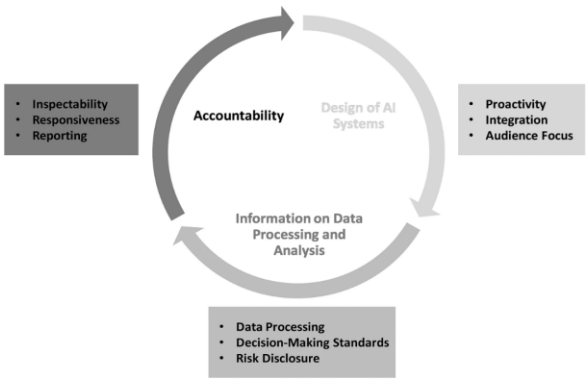

Figure 4: From Felzmann, et al. [48]

6.3 Directions for Future Research

This work may have implications for a rarely-studied management phenomenon: employee conspiracy theories. Employee conspiracy theories [42] emerge when an organization undergoes IS-related change and employees develop beliefs about why management is forcing them to change their work habits. These beliefs could involve deskilling to reduce the value (and cost) of employees [49], or of promoting the work goals of management over that of labor [50].

Also, the counter-normative identity concept presented in this work could lead the way toward defining and studying anti-IT identity. This concept has been the subject of conjecture in works based on IT Identity [51]. It may be that there is an anti-IT identity that an individual may form based on counter-normative behavior, defined in opposition to an IT when a threatening IT seems to be popular among their peers.

Finally, this work may inform the reference disciplines of psychology and sociology by providing evidence that counter-normative behaviors are tied to identities, transparency of motive, and self-worth. Counter-normative behavior in the face of cruel societal norms has long been studied in the contexts of sexuality and religion, especially in countries whose laws do not guarantee civil rights [52]. By contrast, insights from this work on counter-normative behavior in the face of beneficial societal norms may help researchers form a more complete picture of resistance behaviors. The dynamics of identity and self-worth may provide more insight than berating those who hold normative views and merely encouraging those who are forced by identity into counter-normative behaviors (or vice versa). If individuals know and understand more about why others seem to oppose them, they may be less likely to imagine dark motives.

7. Conclusion

Our results indicate that social media systems developed to help people communicate and understand the world may cause harm to self-worth and thus spur resistance and possibly even conspiracy theorization, when those systems seem biased. COVID-19 presents the most disruptive pandemic in the post-internet world; it is almost certainly not to be the last. The next pandemic could be more lethal and spread more quickly. It is critical that social scientists learn from the events of the past year so that the efforts of the world's epidemiologists and medical staff will be as successful as possible, and not hindered by irrational (but predictable) behaviors. Hopefully, the theory and findings in this work will help the field of Information Systems contribute to this most important human effort.

8. References

[1] Douglas, K.M., R.M. Sutton, and A. Cichocka, "The psychology of conspiracy theories", *Current Directions in Psychological Science 26*(6), 2017, pp. 538–542.

[2] Metz, R., "How one employee's exit shook Google and the AI industry", *CNN*, 2021. https://www.cnn.com/2021/03/11/tech/google-ai-ethics-future/index.html

[3] Hill, K., and J. White, "Designed to Deceive: Do These People Look Real to You?", *The New York Times*, 2020. https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html

[4] Haslan, S., *Psychology in Organizations: The Social Identity Approach*, Sage Publications, Inc., Thousand Oaks, CA, 2004.

[5] Petriglieri, J.L., "Under threat: Responses to and the consequences of threats to individuals' identities", *Academy of Management Review 36*(4), 2011, pp. 641–662.

[6] Burke, P.J., and J.E. Stets, *Identity theory*, Oxford University Press, Oxford, UK, 2009.

[7] Mummendey, A., "Positive distinctiveness and social discrimination: An old couple living in divorce", *European Journal of Social Psychology 25*(6), 1995, pp. 657–670.

[8] Mummendey, A., T. Kessler, A. Klink, and R. Mielke, "Strategies to cope with negative social identity: Predictions by social identity theory and relative deprivation theory.", *Journal of Personality and Social Psychology 76*(2), 1999, pp. 229.

[9] Imhoff, R., and P.K. Lamberty, "Too special to be duped: Need for uniqueness motivates conspiracy beliefs", *European Journal of Social Psychology 47*(6), 2017, pp. 724–734.

[10] Sternisko, A., A. Cichocka, and J.J. Van Bavel, "The dark side of social movements: Social identity, non-conformity, and the lure of conspiracy theories", *Current Opinion in Psychology 35*, 2020, pp. 1–6.

[11] Rubin, M., and M. Hewstone, "Social identity theory's self-esteem hypothesis: A review and some suggestions for clarification", *Personality and Social Psychology Review 2*(1), 1998, pp. 40–62.

[12] Mullainathan, S., and A. Shleifer, *Media bias*, National Bureau of Economic Research, 2002.

[13] Ardèvol-Abreu, A., and H. Gil De Zúñiga, "Effects of editorial media bias perception and media trust on the use of traditional, citizen, and social media news", *Journalism & Mass Communication Quarterly 94*(3), 2017, pp. 703–724.

[14] Kulshrestha, J., M. Eslami, J. Messias, et al., "Quantifying search bias: Investigating sources of bias for political searches in social media", *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, (2017), 417–432.

[15] Jones, M., "What we talk about when we talk about (big) data", *The Journal of Strategic Information Systems 28*(1), 2019, pp. 3–16.

[16] Yapo, A., and J. Weiss, "Ethical implications of bias in machine learning", *Proceedings of the 51st Hawaii International Conference on System Sciences*, (2018).

[17] Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations", *Science 366*(6464), 2019, pp. 447–453.

[18] Verma, N., K.R. Fleischmann, and K.S. Koltai, "Understanding online trust and information behavior using demographics and human values", *International Conference on Information*, Springer (2019), 654–665.

[19] Benson, T., "Twitter Bots Are Spreading Massive Amounts of COVID-19 Misinformation - IEEE Spectrum", *IEEE Spectrum: Technology, Engineering, and Science News*, 2020. https://spectrum.ieee.org/tech-talk/telecom/internet/twitter-bots-are-spreading-massive-amounts-of-covid-19-misinformation

[20] PwC, "PwC response to NIST RFI: Developing a Federal AI Standards Engagement Plan", https://www.nist.gov/system/files/documents/2019/06/03/nist-ai-rfi-pwc_001.pdf

[21] Branscombe, N.R., N. Ellemers, R. Spears, and B. Doosje, "The context and content of social identity threat", *Social identity: Context, Commitment, Content*, 1999, pp. 35–58.

[22] Giannakakis, A.E., and I. Fritsche, "Social identities, group norms, and threat: On the malleability of ingroup bias", *Personality and Social Psychology Bulletin 37*(1), 2011, pp. 82–93.

[23] Lewandowsky, S., J. Cook, K. Oberauer, S. Brophy, E.A. Lloyd, and M. Marriott, "Recurrent fury: Conspiratorial discourse in the blogosphere triggered by research on the role of conspiracist ideation in climate denial", *Journal of Social and Political Psychology 3*(1), 2015, pp. 142–178.

[24] Stets, J.E., and P.J. Burke, "Identity theory and social identity theory", *Social Psychology Quarterly*, 2000, pp. 224–237.

[25] Turner, J.C., "Towards a cognitive redefinition of the social group.", *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 1981.

[26] Craig, K., J.B. Thatcher, and V. Grover, "The IT Identity Threat: A Conceptual Definition and Operational Measure", *Journal of Management Information Systems 36*(1), 2019, pp. 259–288.
[27] Cohen, J., "The Duesberg phenomenon", *Science 266*(5191), 1994, pp. 1642.
[28] Wagner-Egger, P., A. Bangerter, I. Gilles, et al., "Lay perceptions of collectives at the outbreak of the H1N1 epidemic: heroes, villains and victims", *Public Understanding of Science 20*(4), 2011, pp. 461–476.
[29] Stojanov, A., and J. Halberstadt, "The Conspiracy Mentality Scale", *Social Psychology*, 2019.
[30] Sparkman, G., and G.M. Walton, "Dynamic norms promote sustainable behavior, even if it is counternormative", *Psychological Science 28*(11), 2017, pp. 1663–1674.
[31] Yarrison, F.W., "Normative Vs. Counter-Normative Identities: The Structural Identity Model", 2013.
[32] Cambon, L., A. Djouari, and J.-L. Beauvois, "Social judgment norms and social utility: When it is more valuable to be useful than desirable", *Swiss Journal of Psychology 65*(3), 2006, pp. 167–180.
[33] Lantian, A., D. Muller, C. Nurra, and K.M. Douglas, "I know things they don't know!", *Social Psychology*, 2017.
[34] Jagose, A., "The trouble with antinormativity", *differences 26*(1), 2015, pp. 26–47.
[35] Halevy, N., O. Weisel, and G. Bornstein, "'In-group love' and 'out-group hate' in repeated interaction between groups", *Journal of Behavioral Decision Making 25*(2), 2012, pp. 188–195.
[36] Hodge, R.W., P.M. Siegel, and P.H. Rossi, "Occupational prestige in the United States, 1925-63", *American Journal of Sociology 70*(3), 1964, pp. 286–302.
[37] Neu, J., *Sticks and stones: The philosophy of insults*, Oxford University Press on Demand, 2008.
[38] Bond, M.H., K.-C. Wan, K. Leung, and R.A. Giacalone, "How are responses to verbal insult related to cultural collectivism and power distance?", *Journal of Cross-Cultural Psychology 16*(1), 1985, pp. 111–127.
[39] Eveland Jr, W.P., and D.V. Shah, "The impact of individual and interpersonal factors on perceived news media bias", *Political Psychology 24*(1), 2003, pp. 101–117.
[40] Gibbon, P., and K. Durkin, "The third person effect: Social distance and perceived media bias", *European Journal of Social Psychology 25*(5), 1995, pp. 597–602.
[41] Kim, H.-W., and A. Kankanhalli, "Investigating user resistance to information systems implementation: a status quo bias perspective", *MIS quarterly*, 2009, pp. 567–582.
[42] van Prooijen, J.-W., and R.E. de Vries, "Organizational conspiracy beliefs: Implications for leadership styles and employee outcomes", *Journal of Business and Psychology 31*(4), 2016, pp. 479–491.
[43] Muthén, B., and D. Kaplan, "A comparison of some methodologies for the factor analysis of non-normal Likert variables", *British Journal of Mathematical and Statistical Psychology 38*(2), 1985, pp. 171–189.
[44] Nunnally, J., *Psychometric Theory*, McGraw-Monte, New York, 1978.
[45] Straub, D., M.-C. Boudreau, and D. Gefen, "Validation guidelines for IS positivist research", *The Communications of the Association for Information Systems 13*(1), 2004, pp. 63.
[46] Tabachnick, B.G., L.S. Fidell, and J.B. Ullman, *Using multivariate statistics*, Pearson, Boston, MA, 2007.
[47] Preacher, K.J., and A.F. Hayes, "SPSS and SAS procedures for estimating indirect effects in simple mediation models", *Behavior Research Methods, Instruments, & Computers 36*(4), 2004, pp. 717–731.
[48] Felzmann, H., E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Towards transparency by design for artificial intelligence", *Science and Engineering Ethics*, 2020, pp. 1–29.
[49] Mirrlees, T., and S. Alvi, "Taylorizing Academia, Deskilling Professors and Automating Higher Education: The Recent Role of MOOCs.", *Journal for Critical Education Policy Studies (JCEPS) 12*(2), 2014.
[50] Lyytinen, K., and M. Newman, "Explaining information systems change: a punctuated socio-technical change model", *European Journal of Information Systems 17*(6), 2008, pp. 589–613.
[51] Carter, M., S. Petter, V. Grover, and J. Thatcher, "Information technology identity: A key determinant of IT feature use and exploratory usage", *Management Information Systems Quarterly 44*(3), 2020.
[52] Francis, D.A., "What does the teaching and learning of sexuality education in South African schools reveal about counter-normative sexualities?", *Sex Education 19*(4), 2019, pp. 406–421.