Characterizing Social Imaginaries and Self-Disclosures of Dissonance in Online Conspiracy Discussion Communities

SHRUTI PHADKE, University of Washington, USA
MATTIA SAMORY, GESIS, Germany
TANUSHREE MITRA, University of Washington, USA

Online discussion platforms provide a forum to strengthen and propagate belief in misinformed conspiracy theories. Yet, they also offer avenues for conspiracy theorists to express their doubts and experiences of cognitive dissonance. Such expressions of dissonance may shed light on who abandons misguided beliefs and under what circumstances. This paper characterizes self-disclosures of dissonance about QAnon—a conspiracy theory initiated by a mysterious leader "Q" and popularized by their followers "anons"—in conspiratorial subreddits. To understand what dissonance and disbelief mean within conspiracy communities, we first characterize their social imaginaries—a broad understanding of how people collectively imagine their social existence. Focusing on 2K posts from two image boards, 4chan and 8chan, and 1.2 M comments and posts from 12 subreddits dedicated to QAnon, we adopt a mixed-methods approach to uncover the symbolic language representing the *movement*, *expectations*, *practices*, *heroes* and *foes* of the QAnon community. We use these social imaginaries to create a computational framework for distinguishing belief and dissonance from general discussion about QAnon, surfacing in the 1.2M comments. We investigate the dissonant comments to characterize the dissonance expressed along QAnon social imaginaries. Further, analyzing user engagement with QAnon conspiracy subreddits, we find that self-disclosures of dissonance correlate with a significant decrease in user contributions and ultimately with their departure from the community. Our work offers a systematic framework for uncovering the dimensions and coded language related to QAnon social imaginaries and can serve as a toolbox for studying other conspiracy theories across different platforms. We also contribute a computational framework for identifying dissonance self-disclosures and measuring the changes in user engagement surrounding dissonance. Our work provide insights into designing dissonance based interventions that can potentially dissuade conspiracists from engaging in online conspiracy discussion communities.

CCS Concepts: • **Human-centered computing** → **Social media**; *Social content sharing*; **Social network analysis**; **Empirical studies in collaborative and social computing**; HCI theory, concepts and models; • **Applied computing** → *Sociology*.

Additional Key Words and Phrases: cognitive dissonance, conspiracy, social imaginaries, machine learning, semiotics, content analysis, online communities, conspiracy theories, QAnon

ACM Reference Format:
Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2021. Characterizing Social Imaginaries and Self-Disclosures of Dissonance in Online Conspiracy Discussion Communities. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 468 (October 2021), 35 pages. https://doi.org/10.1145/3479855

Authors' addresses: Shruti Phadke, phadke@uw.edu, University of Washington, Seattle, USA; Mattia Samory, mattia.samory@gesis.org, GESIS, Cologne, Germany; Tanushree Mitra, tmitra@uw.edu, University of Washington, Seattle, USA.

2573-0142/2021/10-ART468 $15.00
https://doi.org/10.1145/3479855

1 INTRODUCTION

Conspiracy ideation has widespread consequences ranging from inducing distrust and paranoia in individuals to threats to national security [39, 79]. January 6th 2021 riots at the U.S Capitol by QAnon conspiracy believers show how conspiracy theorizing can lead to harmful collective action [86]. Conspiracy theorists utilize online platforms to enhance and reaffirm their conspiratorial beliefs by discussions with peers [56, 62]. However, conspiracy theories are plagued by inconsistencies and fallacies [87] that could induce a state of cognitive dissonance—belief in contradictory ideas [29]. Consider, for example, the following comment left by a QAnon conspiracy community member on Reddit:

> I really want to believe. But lately I have been wondering whether Q is a psyop launched by the clowns. If the april showers fail to happen, I'm going to take some time off from Q and re-evaluate my interest.

The above statement contains some key aspects relevant to belief and dissonance in the QAnon conspiracy theory. The person indicates their desire to believe in QAnon but at the same time expresses dissonance by doubting the legitimacy and efficacy of the QAnon leader—Q. Cognitive dissonance can motivate people to change their behaviors and attitudes [28, 51]. Thus, experiencing dissonance with conspiracy belief can trigger individuals to depart from conspiratorial views. Hence, studying how conspiracists express dissonance with their beliefs is crucial towards understanding the pathways of recovery from conspiracy theories. But how do we identify expressions of dissonance in conspiratorial discussions? The cryptic, symbolic language, often unintelligible to outsiders of the community, but widely used for communicating within the community, poses significant challenges in interpreting online conspiracy discourse [14, 70]. For example, several parts of the above comment are unintelligible to the outsiders of QAnon discussion communities. "Q", here refers to the leader of the QAnon community who posts prophetic message on image boards for his followers to decipher. "Clowns" are corrupt FBI and CIA agent declared as enemies of the QAnon movement. "April showers" refer to the promise made by Q to their followers about the arrests of corrupt politicians.

Coded language is typical of conspiracy discussions and the primary mechanism for conspiracists to sustain their social imaginaries—collectively imagined realities by a group of similar minded people [47, 81]. Understanding the social imaginaries of conspiracists can reveal the components of their collective existence, such as, their knowledge construction practices, their expectations from reality, their legends and characterization of the outside word [14]. Thus identifying such social imagines will provide the means to understand conspiratorial expressions of belief and disbelief or dissonance. In this paper, we focus on the QAnon conspiracy theory and first ask:

RQ1: What are the social imaginaries of QAnon established by the leader Q?

Given the prominence of Q in establishing the entire belief system of the QAnon community [12], we conduct a qualitative content analysis of over 2000 Q-drops—posts made by Q on 4chan and 8chan—and lay out the social imaginaries that Q puts forth. We find five dimensions of QAnon social imaginaries: *movement*, i.e., the collective identity of the believers who mobilize around the conspiracy; *expectations*, i.e., promises and prophecies made by Q to their followers; *practices*, i.e., collective knowledge construction and conspiracy theorizing of the QAnon community; *heroes*, who are considered the leaders serving the greater good in the social imaginaries; and *foes* are the enemies of the QAnon community.

Social imaginaries help conspiracists maintain separation between conspiracy theory insiders and outsiders [46]. Take, for instance, the use of the term "clown" in the previous example. Clown is connoted negatively so as to show inauthentic and unreliable behavior, thus distancing them as

outsiders. On a meta-linguistic level, insider knowledge is essential for interpreting Q's posts and by extension to decode the expressions of other believers who follow Q. Thus, using the word "clown" that has a specific meaning inside the QAnon community, positions the author of the example as an insider of the QAnon community. In sum, social imaginaries help understand how QAnon believers frame their communication. Hence, we next ask:

RQ2: How do QAnon followers communicate QAnon social imaginaries?

To understand how QAnon followers adapt social imaginaries presented by their leader Q, we analyze Reddit communities where followers often reference and discuss Q-drops. Specifically, using the context of over 1.2M posts and comments from 12 QAnon discussion subreddits, we encode various phrases used to express the concepts of QAnon social imaginaries. We combine quantitative dynamic phrase matching techniques and manual validation to create the QAnon Canon—a lexicon of 403 phrases capturing coded language used by QAnon followers to communicate QAnon social imaginaries. For example, while in Q-drops, Hillary Clinton and Barack Obama, *foes* of QAnon, are mentioned as "HRC" and "HUSSEIN", QAnon followers adapt various expressions, such as "Killary", "HC" and "Obummer", "fraudabama".

While symbolic communication of social imaginaries can facilitate secretive, in-group communication [47], amplify the core conspiratorial ideas and beliefs towards them [43], they can also be used to express dissonance or disbelief in conspiracies [48]. For example, since Q is the leader of the QAnon conspiracy, doubt expressed by a QAnon follower questioning Q's legitimacy can be construed as dissonance with QAnon.

RQ3a: How can we identify belief and dissonance expressions in the QAnon community?
RQ3b: How do users express dissonance within the QAnon social imaginaries?

Based on the QAnon Canon and other theoretically-informed constructs of belief and dissonance, such as the language of doubt or tentativeness [27], credibility cues [54], integrative complexity [21], we create a computational framework to classify Reddit comments as *belief*, *dissonance* or *neutral*. Specifically, we use active learning techniques to surface and label expressions of belief and dissonance, and to distinguish them from general talk in QAnon subreddits. Our classifier achieved precision scores above 0.7 for all three classes. We find that phrases from QAnon Canon representing QAnon social imaginaries are important in predicting expressions of belief and dissonance. For example, expressions of belief contain words related to QAnon *movement* ("patriots", "wwg1wga" short for "where we go 1 we go all"), whereas dissonance self-disclosures frequently mention phrases related to *expectations* ("arrests", "predictions"). To further understand the fracture points in QAnon social imaginaries, we complement our quantitative method with a qualitative analysis of a sample of dissonant comments. Our analysis reveals how dissonance occurs along the dimensions of QAnon social imaginaries. Specifically, we find that dissonance can be triggered because of unfulfilled *expectations* and perceived illegitimacy of QAnon *heroes*.

Dissonance can also change behavior and attitudes [28, 29]. While some people strengthen their beliefs or even seek validation by recruiting more believers, others may choose to leave the community [29]. Hence, we finally pose RQ4, where we analyze how user engagement inside and outside the QAnon communities change after self-disclosure of dissonance.

RQ4: How does engagement in QAnon subreddits change after expressing dissonance?

We conduct an interrupted time series (ITS) analysis of user contributions—number of comments or posts—inside the QAnon communities in the 12 week period surrounding self-disclosures of dissonance. We find that user contributions in QAnon communities decrease significantly, immediately after dissonance disclosure, but not after expressing belief, while their overall engagement on

Reddit stays the same. We corroborate and extend these results through various regression models, showing that not only disclosures of dissonance are followed by a decrease in contributions, but also by the departure of the users from the community. In particular, users who disclose dissonance disproportionately more than belief, are those most likely to leave the community.

Below we outline the contributions and implications of our work:

- We develop a systematic framework for uncovering social imaginaries of conspiracies and for finding various language correlates of social imaginaries in conspiracy discourse (Figure 3).
- We offer the QAnon Canon, a lexicon of over 403 phrases capturing symbolic language and its shared meanings across QAnon social imaginaries that can serve as a toolbox for researchers to extend our study to other platforms (Table 2).
- We offer a computational framework for identifying expressions of dissonance in the QAnon community (Figure 6) which can be used to identify points of fracture in the QAnon belief.
- We detail the points of fracture in QAnon social imaginaries (Table 4) which can be used to design dissonance based interventions for online conspiracy discussion participation.

In the rest of the paper, we first provide a background on the QAnon community and survey literature studying social imaginaries and dissonance in conspiracies. We then discuss the methods and results of each of our research questions. Finally, we conclude with the discussion and ethical considerations of our work.

2 BACKGROUND

2.1 What is QAnon?: Origin and Community Dynamics

In October 2017, a user on the image board, 4chan, signed off as "Q" and posted a comment prophesying the arrests of Hillary Clinton and her staff. In successive posts, "Q" purported the arrests of several other politicians associated with the "Deep state"—a conspiracy theory claiming that a coalition of politicians in the U.S. run a shadow government involved in corruption and cronyism [7]. In the next several months, the discussions around Q's posts took 4chan by the storm. Q presented themselves as an anonymous, high ranking U.S military official, with insider information about the U.S. government. Together, "Q", a mysterious, prophetic leader and "anons", Q's followers comprise the QAnon community. Specifically, Q predicts various political events in their Q-drops—messages posted to image boards such as 4chan, 8chan. Subsequently, followers of QAnon start piecing together the clues left in Q's posts and predictions. In fact, Q-drops are carefully crafted to contain cryptic messages such as "*the wormhole goes deep*" or "*future proves past*" which are meant to be clues for the followers to decipher. By encouraging followers to "open their eyes" and "search for truth" Q has institutionalized knowledge production practices that are unparalleled by other conspiracy movements [61]. Specifically, Q-followers are called "bakers" who assemble the "crumbs" (clues) left by Q into coherent pieces [2, 61]. This social construction of knowledge then produces unambiguous certainty through alternate reality [61], a characteristic commonly associated with new religious movements [6]. In this shared alternate reality, Q and their audience identify themselves as actors in a larger movement by using specific designations such as "anons", "patriots", "bakers". A large part of QAnon's alternate reality and worldview is represented by the use of symbolic language whose shared meaning is understood only within the community.

QAnon community has started attracting research attention due to its clear mobilizing potential. Specifically, previous studies explored the topics and dissemination of Q's messages on various online platforms and found that QAnon borrows theories from other conspiracies such as Pizzagate [60], shares moral values with Christian theology [53] and QAnon followers are likely to use violent rhetoric on Twitter [65]. While the existing studies provide valuable characterization of the QAnon

movement across platforms [1], deeper psychological and sociological exploration is required to deter increased engagement with QAnon [35]. Our work fills this gap by first establishing QAnon social imaginaries symbolizing collective interpretation of reality by QAnon and then using those imaginaries to identify expressions of dissonance in the QAnon community. We start by first providing the background on the role of social imaginaries in the conspiracy communities.

2.2 Conspiracists and Social Imaginaries

Conspiracy theorizing generally consists of a belief that a covert operation is being carried out by a group of conspirators or secret organizations to influence events [42, 63]. Conspiracy theorizing is able to produce a certain aesthetic pleasure that enables people to form social imaginaries—coherent, collective imagination of social existence by a set of people [46, 82]. Social imaginaries refer to the ways people imagine their social existence, their relationship between different social groups, deeper normative notions and the expectations of reality born out of such norms [81]. For example, a social imaginary of a conspiracy theorist can consist of irrational interpretation of reality, such as, "there is a global lobby trying to enslave common people" or "conspiracists being ridiculed by ignorant mass are further proof of the subversion by elites" [46] or QAnon's purported worldview that "America is run by a cabal of pedophiles and Satan-worshippers who run a global child sex-trafficking operation and QAnon are force of good stopping them" [57]. In sum, social imaginaries lie at the heart of the belief systems and are a way for groups to rationalize their sense of reality and even find purpose in the collective action [81]. Thus we argue that it is important to understand the conspiracists' social imaginaries.

In this paper, we study conspiracy social imaginaries by focusing on the QAnon movement. The entire QAnon conspiracy theory is based on the leader Q's early messages posted on image boards such as 4chan and 8chan [86]. In fact, QAnon's social construction of knowledge is largely centered around understanding Q's missives [89]. Hence, in RQ1, to uncover the dimensions of QAnon social imaginaries, we perform a qualitative analysis of over 2000 posts made by Q on 4chan and 8chan. Social imaginaries are typically shared by a group of people and instill the sense of legitimacy to their cause and existence as a group [81]. While in RQ1 we identify social imaginaries based on the leader Q's messages, Q's followers may adopt various linguistic expressions to communicate QAnon social imaginaries on online platforms. Hence, we next consider the role of semiotics—use of symbols to communicate shared meanings—in conspiracy online discourse.

2.3 Conspiracies and Semiotics

Maintaining the separation between the insiders—people who are aware of the conspiracy—and the outsiders—enemies of irrational interpretation of reality—is essential for sustaining and communicating the social imaginaries of conspiracy theories [46]. Semiotics, in the form of secretive communication and symbolic language, is used to maintain separation between the insiders and the outsiders of the conspiracy communities [14, 46]. Specifically the type of semiotics based on paranoid thinking commonly associated with conspiracists [39] is destined to render the communication unintelligible to the outsiders [25]. In fact, Leone Festinger, a renowned social psychologist argues that the main role of semiotics in conspiracies is to disrupt the common sense interpretations of public discourse known to the outsiders [47]. Consider for example the use of the words "clowns" or "swamp" in the QAnon community. While to an outsider clowns mean comic performers and swamp means a wetland, in QAnon community "clowns" mean FBI or CIA agents and "swamp" refers to the collective of people and communities associated with deep state. By disrupting the common sense of the outside world, conspiracists create common sense of their own, shared among the community of insiders [14, 36]. In other words, through the use of semiotics, conspiracists create, develop, and propagate shared meanings that sustain their social imaginaries [14, 46]. In

RQ2, we quantitatively identify semiotic patterns used by QAnon followers and encode them in the form of a phrase lexicon, QAnon Canon.

2.4 Conspiracies and Cognitive Dissonance

On one hand social imaginaries and their semiotic communication play an important role in affirming conspiracy belief [14], on the other, they can be used to understand dissonance with conspiracies as well. While not all conspiracies are false or impossible [8], many suffer inconsistencies, contradictions, and general epistemological challenges [79, 87]. Realizing such inconsistencies may induce a state of dissonance among conspiracy followers. How do conspiracy believers react when their beliefs are contradicted? Researchers found that mistrust in authorities or governing bodies is sufficient to overwhelm the contradictions between individual conspiracy theories [87]. Festinger coined the phenomenon of believing in contradictory, inconsistent ideas as "cognitive dissonance" [28, 30]. In a famous immersive ethnographic study, Festinger and colleagues infiltrated a UFO religion in Chicago where the cult leader had prophesied that the world will end on December 1954. Festinger and colleagues revealed that after the prophesied date and obvious signs of world not ending, members of the group experienced cognitive dissonance. According to the theory of cognitive dissonance proposed by Festinger, dissonance can result from various individual or social factors. For example, dissonance can result from involuntary or voluntary exposure to information that directly contradicts previously held beliefs. Further, the simple act of having to choose between two contradictory ideas can also intensify the experience of dissonance. Moreover, dissonance can result from the conflict between individually held and socially accepted beliefs. When confronted with such dilemmas, individuals may change their perception of contradictory ideas, find overlap between the two ideas or completely reverse their previously held beliefs [29]. Indeed, Festinger's study showed that after experiencing dissonance, different believers reacted in different ways. While some strengthened their convictions and even recruited newer members, others left the cult. In other words, being in the state of dissonance, where beliefs are challenged or contradicted, may lead people to change their behaviors or attitudes [28, 66, 76].

Similar to the cult leader in Festinger's study, Q—the leader of QAnon—has made several predictions that never came true [83]. For example, amongst hundreds of other predictions, Q's very first prediction about Hillary Clinton's arrest in 2017 has provably failed [83]. Can Q's failed predictions, similar to the UFO religion studied by Festinger, induce dissonance in Q's followers? What other fracture points in QAnon belief can induce cognitive dissonance? In this light QAnon makes for an ideal case to study dissonance in conspiracies, where we can analyze both social imaginaries created by Q and the self-disclosures of dissonance by Q followers on the fracture points of the social imaginaries. To our knowledge, there is no study exploring how conspiracy believers express doubt, either in the face of contradictions or other ideals clashing with the social imaginaries. In RQ3, we fill this gap by first identifying the expressions of dissonance among the members of online QAnon community and next, highlighting the points of dissonance in the QAnon belief.

What happens after people express dissonance? Researchers studying addictive behaviors found that higher levels of cognitive dissonance can help people deconstruct their previously held norms and beliefs and thus, pave the pathway for recovery[10, 34, 84]. In RQ4, we analyze changes in user engagement with QAnon discussion communities after they express dissonance with QAnon.

3 DATA

3.1 Q-drops Dataset: Q-drops from 4chan and 8chan

In the beginning of the QAnon conspiracy, Q posted messages (Q-drops) on image boards, and the followers discussed Q-drops across various Reddit communities [12] (See Figure 2). In RQ1,

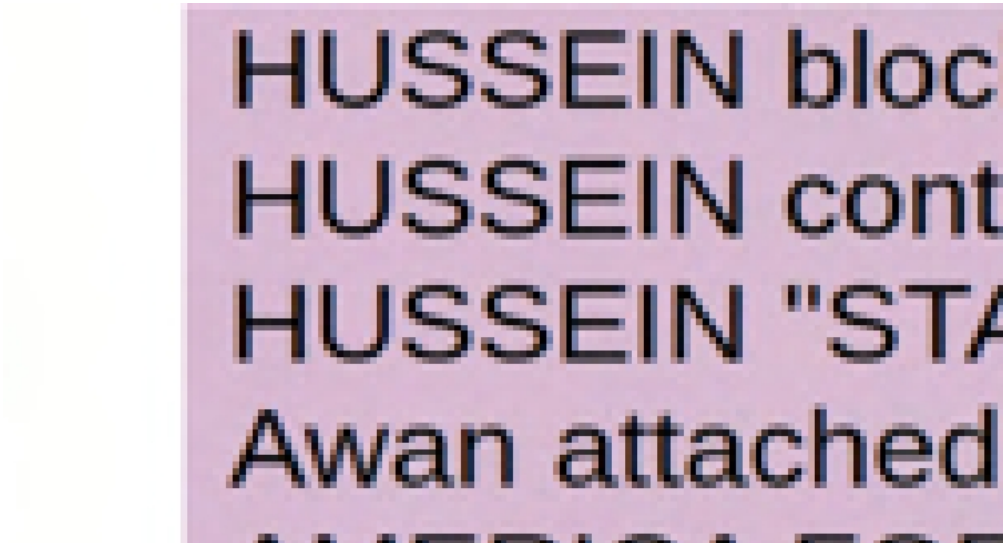


Fig. 1. Example Q-drops (a, b, c) and explanations of cryptic language (d, e) available on a QAnon aggregator site—qalerts.app. In some cases, the meanings behind codes are explained in the Q drops itself. For example the word "clowns" mentioned in (b) is later explained in another Q-drop (c). A common pursuit of QAnon followers is to decode the cryptic language and understand the core message posted by Q. For example, on qalerts.app one of the codewords "HUSSEIN" in a Q-drop (shown in (a)) is decoded by QAnon community members as Barack Hussein Obama (see (d)).

we analyze Q-drops to characterize social imaginaries established by the leader Q for the QAnon community. Specifically, Q drops were made on 4chan and 8chan boards that are anonymous, ephemeral forums revolving around posting images along with text. 4chan and 8chan are infamous for hosting controversial content resulting in multiple temporary bans. We use qalerts.app, a website that aggregates Q-drops from image boards. While the exact agency of qalerts.app is unknown, it is a common resource used by QAnon communities[1] and also by other researchers studying QAnon [1]. There are several other Q-drop aggregation sites, however, most of them contain nearly similar record of Q posts (see Table 1 in Aliapoulios et al. [1]). We downloaded the first 2166 Q-drops that were made between October 2017 and September 2018 to allow for consistent analysis time period between various RQs in the paper[2]. For example, see subreddit timelines in Figure 2. The 12 QAnon discussions subreddits existed in the overall time period of November 2017 to September 2018. Hence, we consider the first 2166 Q-drops spanning over the same time period.

qalerts.app also hosts other resources, such as list of abbreviations specific to the QAnon community and research compiled by QAnon followers. See for example the screenshot of abbreviations provided on qalerts.app in Figure 1 (d) and (e). We utilize these resources to understand and contextualize the content of the Q-drops in RQ1.

3.2 QAnon Subreddits Dataset: Reddit Dataset of 12 Banned QAnon Communities

After the emergence of Q on 4chan in September 2017, the QAnon movement popularized on Reddit [12]. To reach a more mainstream audience, prominent QAnon followers created a subreddit

[1]See QAnon and the Great Awakening group on Gab.com

[2]The last posted Q-drop was in December 2020, making it a total of 4953 Q-drops posted since 2017.

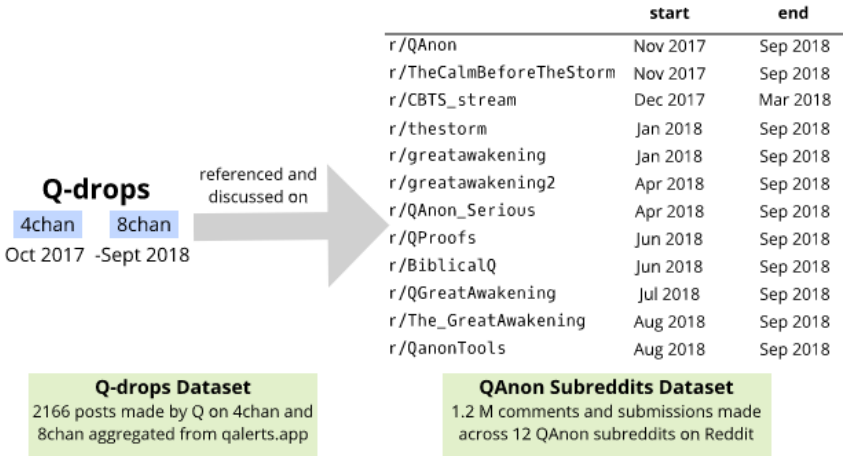


Fig. 2. Figure shows two datasets used in this paper. The Q-drops dataset is used to derive social imaginaries in RQ1 while QAnon subreddit dataset is used in RQ2, RQ3 and RQ4. Usually, Q posts drops—Q-drops—on image boards such as 4chan and 8chan. Q-drops are then referenced and discussed in QAnon discussion subreddits. The 12 subreddits in our dataset comprise of posts made between Nov'2017 and Sept'2018. Hence, to understand relevant social imaginaries, we consider Q-drops from the start of Q in 2017 to Sept'2018.

called r/CBTS_Stream, short for *Calm before the Storm*–a popular saying in the QAnon community indicating impending arrests of the deep state politicians. However, r/CBTS_Stream was banned in March 2018 for violating Reddit's terms of content policies. This ban resulted in the creation of new subreddits such as r/greatawakening2, r/BiblicalQ and others that combined accrued more subscribers than the original r/CBTS_Stream [13]. Finally, 17 of these new communities were also banned by Reddit in September 2018 for repeated violation of content policies [13]. We identified the 17 banned QAnon related subreddits from various press mentions [13, 69]. To obtain the data for banned subreddits, we used Reddit Pushshift Dataset[3] [9]. Specifically, we queried the Pushshift data from Google Bigquery and obtained the submissions and comments from 12 of the 17 banned subreddits. The data for the rest of the 5 subreddits is not present on Pushshift nor through the official Reddit APIs. In total, we have 96,068 submissions and 1,104,096 comments made by 33,561 users across 12 subreddits. Figure 2 provides an overview.

4 RQ1: CHARACTERIZING THE QANON SOCIAL IMAGINARIES ESTABLISHED BY Q

What are the social imaginaries of the QAnon community? By answering, we will uncover how QAnon perceive their community and the outside world, and how they construct their practices and form their expectations from the reality [81]. We first describe our approach to understanding social imaginaries or worldview of the QAnon community. We employ qualitative content analysis to explore various dimensions of the QAnon social imaginaries. Given the prominent use of semiotics in conspiracy discussions [14, 46], we focus on discovering words and phrases with shared meaning that would represent various aspects of the social imaginary. Finally, we present the five dimension of QAnon social imaginaries surfaced from the qualitative analysis along with examples.

[3] https://files.pushshift.io/reddit/comments/

Fig. 3. Figure showing our process for understanding the dimensions of QAnon social imaginaries and recording the ways in which social imaginaries are communicated by the followers inside the QAnon communities. (a) In RQ1 we perform content analysis of over 2K Q-drops and find five dimensions of the social imaginaries. (b) We also recorded 75 phrases—seed lexicon—that use coded language across various dimensions of social imaginaries. In RQ2, we expand the seed lexicon using public discourse in QAnon subreddits. We first (c) resolve coreferences and then (d) find numerical representations for meanings of phrases using PMI. (e) We then iteratively find similar phrases and (e) create QAnon Canon—a lexicon of 403 phrases capturing symbolic communication of QAnon social imaginaries

4.1 RQ1 Method: Qualitative Content and Semiotic Analysis

Q-drops—posts made on image boards by the leader Q—are at the heart of the QAnon movement. QAnon followers consider Q-drops as the main source of insider knowledge [83] and devote themselves to decoding Q-drops in order to expose the deep state [1]. Hence, to understand the social imaginaries of the QAnon community, we analyze the Q-drops dataset (Figure 2). Specifically, we use inductive qualitative content analysis [26, 85]—an inductive reasoning method aimed at revealing themes, patterns and meanings embedded in the data [50]. The core idea behind content analysis is that many words or phrases from the text (Q-drops in this case) are categorized into much fewer coherent categories [74]. Words or phrases in the same category represent similar themes. This similarity can be based on semantic characteristics of the words and phrases [41]. Further, to understand the process of semiosis within the QAnon community, we consider symbolic words and phrases that mediate a common meaning [55]. Consider for example, the Q-drops in Figure 1(a). The phrase "april showers" may not mean anything to an outsider. However, inside the QAnon community "april showers" has specific significance; it symbolizes Q's prophecy about mass arrests of the deep state politicians in April 2018.

4.1.1 ***Extracting information from Q-drops.*** For each of the 2,166 Q-drops, we first manually extracted words and phrases that contributed to shaping the narrative of the QAnon movement.

Category	Example Texts from Q-drops
movement	*Dear **Patriots**, We hear you. We hear all all Americans such as yourself*
	*Think new arrivals. Proofs are important. Thank you, **autists** and **anons***
	*The **wizards and warlocks** will not allow another Satanic Evil POS control our country*
expectations	***HRC extradition** already in motion effective yesterday*
	*Next week. **Boom. Boom. Boom.***
	*Be vigilant today and expect a major **false flag***
practices	***Crumb** dump incoming fast. **Archive** immediately. Upload to graphic.*
	*Shall we play a game? Find @Snowden. Happy **Treasure Hunt!***
	*Keep **digging** and keep **organizing** the info into graphics (critical).*
heroes	*America First. This Is What Happens When **POTUS** Has No Strings Attached.*
	*Panic in DC. Trust **Sessions**. Enjoy the show.*
	*Trust **Sessions**. Trust **Wray**. Trust **Kansas**. Trust **Horowitz**. Trust **Huber**.*
foes	*Realize **Soros**, **Clintons**, **Obama**, **Putin**, etc. are all controlled by 3 families*
	*Why does the **MSM** portray the country as being divided?*
	*Why wasn't **HRC** prosecuted for the emails? These people are EVIL*

Table 1. Five dimensions of QAnon social imaginaries along with examples. These dimensions capture the collective identity of the community (*movement*), QAnon's leaders and legends (*heroes*), QAnon's enemies (*foes*), QAnon knowledge construction practices (*practices*) and the expectations from the reality (*expectations*)

For example, we recorded terms such as "april showers"[4], "WWG1WGA"[5] (Where we go one we go all) and names of politicians associated with the deep state[6]. Next, we proceed to the iterative inductive coding phase.

4.1.2 ***Iterative Inductive Coding.*** After collecting words and phrases from Q-drops, we began to organize them into themes based on the semantic relationships. Two authors of the paper were involved in this inductive analysis. Our focus was on grouping words that represent similar meanings in the QAnon ideology. For example, the words "boom", "moab" (mother of all booms), "april showers", "red october" all generally allude to a future event, significant to the take down of the deep state. Similarly, the words "anons", "bakers", "autists", "patriots" are all used to refer to insiders of the QAnon community. Note that to understand the insider language of QAnon, we refer to the words cross-listed across various Q-drops and also other online resources such as list of abbreviations on `qalerts.app` and various news articles describing the Q-drops and QAnon language [20, 83]. We iteratively developed the categories through a discursive process. We found saturation at five categories that capture the social imaginaries of the QAnon community. Below, we explain the five dimensions of QAnon social imaginaries—*movement*, *expectations*, *practices*, *heroes*, and *foes*—resulting from the qualitative analysis along with examples.

4.2 RQ1 Results: QAnon Social Imaginaries Established by the Leader Q

Table 1 lists all dimensions of QAnon imaginaries along with example Q-drops for each.

(1) **Movement:** The *movement* category signifies the collective identity of the QAnon community. Movement includes Q-team (a group of anonymous people believed to be working with Q), Q

[4] https://qalerts.app/?n=1007
[5] https://qalerts.app/?n=1025
[6] https://qalerts.app/?n=1708

research team (a group of Q followers that organize and research the Q-posts) and several other designations (anons, bakers, patriots) that collectively represent the Q-followers. Slogans such as WWG1WGA (Where we go one we go all), and WRWY (We Are With You) are used to reinforce faith and motivate collective action in the QAnon movement.

(2) **Expectations:** Through the promises of arrests of deep state agents, Q sets expectations for their followers. Expectations are about both good and bad events in the context of the QAnon community. For example, several Q-drops predict arrests of specific deep state politicians while others warn the readers about false flag events, forecasting covert operations of various governments and cabals.

(3) **Practices:** An important part of QAnon community is hunting for clues provided by Q. Q instructs their followers to follow certain knowledge construction practices. For example, Q asks their followers to organize and archive Q-posts, connect the dots and dig for the truth.

(4) **Foes:** Q routinely releases the names of celebrities, politicians and law enforcement agents who are purportedly associated with the deep state. Deep state agents are believed to be involved in a satanic cult with an international child sex trafficking ring. Simply put, foes of QAnon are portrayed as the enemies of the QAnon movement.

(5) **Heroes:** According to Q, while the law agencies, media and a large part of the government is considered to be controlled by the deep state, there are some "good guys" who fight for the American people. Q, Donald Trump and former U.S Attorney General Jeff Sessions are at the top of this list. Heroes often know more than they choose to reveal to the QAnon community for reasons of national security and are believed to be experts at undercover work.

As a byproduct of the content analysis process, we recorded relevant words and phrases in each of the five dimensions of the QAnon social imaginaries. We refer to this as a **seed lexicon** for QAnon social imaginaries. In the next section, we understand how QAnon followers communicate the social imaginaries established by Q, by quantitatively expanding the seed lexicon.

5 RQ2: COMMUNICATION OF SOCIAL IMAGINARIES BY QANON FOLLOWERS

In RQ1, we uncovered five dimensions of the QAnon social imaginaries—*movement*, *expectations*, *practices*, *heroes*, and *foes* put forward by the leader Q. How do QAnon followers communicate these social imaginaries within the QAnon discussion communities? As a result of the qualitative analysis in RQ1, we obtained a seed lexicon of 75 phrases across five dimensions of social imaginaries. For example, see the words in bold in Table 1 and the words mentioned in Figure 3 (b). Note that the phrases in the seed lexicon are directly extracted from the Q-drops. While discussing the Q-drops, the QAnon followers might adapt various expressions of the phrases. For example, the Q-drops frequently mention "deep state" to refer to the collective of allegedly corrupt politicians and celebrities involved in child trafficking rings. However, a manual inspection of the comments in QAnon discussion subreddits revealed that the QAnon followers use various phrases—"swamp", "antiq" (for anti-Q), "evil pedos"—to refer to the deep state. In this section, we explain our methods for identifying various expressions of phrases in the seed lexicon. As a result, we create the QAnon Canon, a lexicon of 403 phrases capturing the communication practices used by QAnon followers.

5.1 Method: Creating QAnon Canon

We quantitatively expand the seed lexicon into QAnon Canon—a dictionary capturing words and phrases in each of the dimension. We use the QAnon subreddits dataset of 1.2M comments and posts to understand various ways in which phrases from the seed lexicon are expressed. We utilize rigorous quantitative methods that use sentence parsing and semantic similarity of words to expand a seed set of 75 phrases to over 403 phrases. Finding various expressions for phrases from public

Fig. 4. (A) An example discussion thread with coreferences where the entity "Obama" from the parent comment is referred by "he" and "him" in the child comment. To resolve coreferences, we consider the pairs of parent-child comments at a time. This way, with the context of the parent comment, "he" in the reply comment will be replaced with the word "Obama". (B) A screenshot of the user interface used to iteratively expand the seed lexicon by finding similar phrases. The interface displays pre-loaded phrases in the seed lexicon (pink tabs) which can be searched in the search bar. The search returns top similar phrases (displayed right of the search bar) and the examples of comments containing the searched term. The search bar also has auto-complete feature that suggests lexically similar words to the search term using fuzzy search. Users can add new phrases to the lexicon by either clicking and adding phrases from the search results or manually typing and adding phrases from the examples. The example shows the results for "HRC". We select relevant search results (example, hillary, killary, hc) and add them back to the seed lexicon.

discourse is a challenging task. Specifically, pronouns are often used to refer to the noun phrases (ex. using 'she' instead of 'Hillary'). This is called *coreference*. In order to find similar phrases, we first need to resolve the coreferences. After resolving the coreferences, we characterize meanings of various phrases as vectors and use interactive mixed-methods approach to find various expressions of phrases in the seed lexicon. We start by introducing our method for coreference resolution.

5.1.1 ***Coreference Resolution.*** Consider the following Reddit comment: "HRC is corrupt. She is Evil." Here, the first sentence mentions HRC (Hillary Rodham Clinton) and the second sentence refers to the same antecedent entity "HRC" by the pronoun "she". To derive correct interpretation of this text, we first need to resolve the coreference (HRC—she) where pronouns and other referring expressions must be connected to the right entities. A successful coreference resolution will result in the replacement of the pronoun with correct entity: "HRC is corrupt. HRC is Evil." Coreference resolution is a well explored problem in computational linguistics for studying discourse [77]. We use `neuralcoref`[7] with `Spacy`[8] pipeline that resolves coreference clusters using neural networks. To further improve the interpretation of a comment, we consider its parent comment while resolving the coreference. For example, see the comment thread in Figure 4 (a). The first comment mentions an entity "Obama" which is referenced with "he" and "him" in comments in the reply. Hence, while

[7] https://github.com/huggingface/neuralcoref
[8] https://spacy.io/

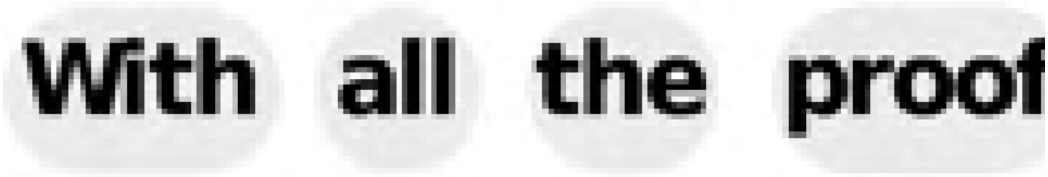


Fig. 5. Figure showing the process of characterizing phrase embeddings and updating the embeddings based on similar phrases. (a) For every target phrase, we first record the context phrases based on the co-occurrence in a sentence. (b) Based on the recorded contexts, we build a count matrix of target phrases vs. context phrases. We calculate the pointwise mutual information (PMI) to first encode the phrase similarity and then derive 200 dimensional embeddings for every phrase based on singular value decomposition (SVD). We use the phrase embedding similarity to search for similar phrases in the interface (Figure 4 (B)) and repeat the process by merging and re-ranking the similarity results of the similar phrases

resolving the coreference in a comment, we also need the context of its parent comment. We use breadth first search to first, find pairs of parent–child comments from the top to the bottom of the comment tree, and then use `neuralcoref` to resolve coreferences in the parent–child comment pair.

5.1.2 ***Characterizing Phrases.*** After resolving coreferences, we characterize similarity between different phrases. To find similarity computationally, we need to represent the meaning of different phrases in the numeric form. This is the process of modeling the meanings of phrases by *embedding* them in the vector space. We first identify phrases by tokening sentences into top unigrams, and bigrams and trigrams using collocations—sequence of commonly occurring words. Note that using collocations allows us to avoid overlap between unigrams and multi-grams. For example, if "Donald Trump" is mentioned in a comment, it will get tokenized only as a bigram and not as individual unigrams, "Donald" and "Trump". We create sparse vector representations, or embeddings, for phrases using the information about their co-occurrence with other phrases in the comment. We first calculate phrase co-occurrence matrix which records how often a phrase co-occurs with another phrase. We consider two phrases to "co-occur" when they are present together in a comment within a span of five phrases (Figure 5 (a)). In other words, to calculate the co-occurring phrases for any particular phrase, we consider the *context* of five phrases. However, raw co-occurrence frequency as a measure of association between two phrases has several limitations. For example, common words like stop words (e.g., "the") naturally co-occur with many other phrases according to raw frequency, but provide little information about the words next to them. Hence, we calculate pointwise mutual information (PMI) for every pair of phrases. High PMI between two phrases indicates that the two phrases share a surprisingly higher number of common context phrases [17]. PMI is specifically helpful in surfacing associations to low frequency phrases. However, because the PMI matrix represents associations between every pair of phrases, it is high dimensional, a square matrix with number of rows and columns equal to the total number of phrases in the entire corpus. Hence, to obtain computationally affordable and dense embeddings for every phrase, we perform singular value decomposition (SVD) on the PMI matrix (Figure 5 (b)). Finally, we capture

movement	practices	expectations	heroes	foes
patriots	hive mind	mockingbird	q	kabal
q analyst	dig the truth	big drop	potus	deep state
q research	treasure hunt	boom	white hats	hillary
anons	crumbs	moab	sessions	obama
qteam	clues	arrests	wray	satanists
wizards	watch water	pyramid collapse	mueller	rothschilds
white hats	spider web	layoffs	huber	big pharma
wwg1wga	future proves past	big news week	kansas	red cross
qanon	trust the plan	maga promise	wizards and warlocks	nwo

Table 2. Table showing example phrases from the QAnon Canon. In total, the lexicon contains over 403 phrases recorded across 5 dimensions of QAnon social imaginaries.

the semantic meaning of every phrase in a 200 dimensional SVD vector. Similarity between the phrases can then be obtained by calculating the cosine similarity between their vectors. In the next subsection, we explain our methods for dynamically updating the phrase embeddings to iteratively find similar phrases.

5.1.3 ***Finding similar phrases.*** Consider that we have embeddings for "Hillary" and "HRC". Given the high similarity between the embeddings, and manual verification, we conclude that "Hillary" and "HRC" refer to the same entity. How can we use this information to further find phrases that are similar to both "Hillary" and "HRC"? We use a mixed manual and quantitative approach to iteratively find similar phrases. To ease the process of finding similar phrases, we built a web interface displayed in Figure 4 (B). The web interface shows the seed phrases that belong to the lexicon. It also enables querying for new phrases via a text box that suggests phrase completions as well as similar spellings, and example comments containing the query. Moreover, it shows the 10 phrases that are most semantically similar to the query. In the absence of a query, the interface shows the 10 phrases most similar to the current lexicon. We expanded the lexicon by inspecting and querying for similar phrases that represent related entities and concepts. For example, the seed lexicon for *foes* included "HRC," and "Hillary" was automatically suggested as a similar phrase. We added it to the lexicon, which automatically updated the list of similar phrases, and iterated the procedure until no similar phrase belonged to the lexicon. In the process, we noted related terms that appeared in the example comments, and queried for them, eventually adding them and their similar phrases to the lexicon when appropriate.

Technically, we compute semantic similarity in two ways. In the case of a query phrase, we compute semantic similarity of a new phrase simply as cosine similarity of the corresponding embeddings. Note that instead of using pre-trained embeddings off the shelf, we use our own trained phrase embeddings to better fit our dataset and similarity context. Using locally generated embeddings also enables us to iteratively modify the embeddings in the interactive lexicon generation phase. When no query is selected, we compute the 10 most similar phrases to each phrase in the lexicon, merge them into a single list, rank the list according to similarity to the closest lexicon phrase, and take the 10 highest-ranking phrases. Following this procedure, we expand our lexicon from 75 phrases to 403. We named the final lexicon the QAnon Canon, which we introduce next.

5.2 Results: QAnon Canon

By iteratively finding phrases similar to the seed lexicon, we obtained QAnon Canon—a lexicon of 403 phrases encoding how QAnon followers communicate QAnon social imaginaries. Table

2 displays ten example phrases in each dimension. We are able to recover various expressions for named entities. For example, QAnon Canon contains multiple expressions for Hillary Clinton (hillary, HRC, HC, killary, billary, alice in wonderland, clintons) and Barack Obama (Obama, Hussein, Obamas, ObamaHillaryCIA, Barack, Obummer, 0bama). Similarly we were also able to find lexically and semantically similar expressions of various phrases symbolizing the movement (q-team, q-analyst, q-research, q-clearance), practices (q-drops, qdrops, q drops, qposts, q-posts) and expectations (layoffs, mass exodus). We have made QAnon Canon publicly available for other researchers to use [9]. In the next research question, we use words from QAnon Canon as linguistic features for identifying expressions of belief and dissonance in the QAnon community.

6 RQ3: IDENTIFYING EXPRESSIONS OF BELIEF AND DISSONANCE IN QANON

How do conspiracists express views that are dissonant with the social imaginaries in their communities? To answer, we classify expressions of belief and dissonance in QAnon subreddits. We use various sampling strategies to create a labeled dataset of comments annotated into one of the three categories: *belief* in QAnon, *dissonance* with QAnon and *neutral*. We use this labeled dataset to build an ensemble of machine learning classifiers to identify belief and dissonance in QAnon. We start by introducing our feature set.

6.1 RQ3a Method: Compiling Factors in Self-Disclosure of Belief and Dissonance

Referring to the QAnon Canon and prior literature on expressions of belief and dissonance, we compile lexical, stylistic, and document level features for classifying belief and dissonance in QAnon. We prefer hand-picked, theoretically motivated features over pre-trained sentence embedding models to preserve the interpretability of various features in identifying belief and dissonance. We discuss the importance of different features in Section 6.6.

(1) **QAnon Canon (403 features):** The lexicon, created in RQ2, captures the social imaginaries of the QAnon community, which can be elicited in affirming belief [3]. Similarly, disagreement with the social imaginaries can signal dissonance with the QAnon worldview. Hence, we calculate the frequency of each of the 403 phrases in the QAnon Canon in user comments.

(2) **Linguistic Inquiry and Word Count (LIWC) (19 features):** LIWC encodes words that capture affective, emotional and cognitive processing expressions and is often used for analysis of online texts [80]. For example, LIWC categories of tentativeness (includes words such as *maybe*, *perhaps*) and certainty (*always*, *never*) were specifically found to be relevant in the expressions of doubts in online reviews [27]. Following a similar rationale, we include 17 other relevant LIWC categories[10]. For example, we include conjunctions (*and*, *but*, *whereas*) that may be present in arguments that combine contradictory claims ("I am trying to trust Q **but** my patience is running out") [23].

(3) **Integrative Complexity (IC) Score (1 feature):** IC is a psychometric that captures people's ability to recognize multiple perspectives and connect them together [78]. IC is closely related to expressing belief and attitudes [21]. IC scores range from 1 to 7, where 1 indicates no evidence of IC and 7 indicates the presence of overarching perspectives with detailed connections. We calculate the IC score using the model published by Robertson et al. [68].

(4) **Credibility Cues (8 features):** Perceived credibility and the evaluation of the common knowledge are strongly associated with belief and disbelief [54, 64]. We include credibility cues such as booster words (*actually*, *evidently*), hedge words (*in my view*, *in general*), modal words

[9] https://social-comp.github.io/ConspiracyTraces/

[10] feel, discrepancy, second person pronouns, differentiation, religion, third person singular pronouns, causation, first person pronouns, anger, hear, third person plural pronouns, insight, sadness, see, negations, conjunctions, anxiety

Fig. 6. Figure showing RQ3a method flow. (a) First we design and extract features for each of the 1.2M Reddit comments and posts. (b) Due to inherently imbalanced dataset (QAnon subreddits are likely to have very few expressions of dissonance compared to belief or neutral) we use various sampling techniques to potentially find comments expressing dissonance. (c) We use labeled data generated from different sampling techniques to train two machine learning classifiers and consider their consensus as final class prediction.

(*hypothetical*, *improbable*) and evidentials (*know*, *guess*) that are associated with perceptions of credibility [54]. We also calculate sentiment scores [11] and number of quotations [22] in a comment that might indicate uncertainty. Additionally, we include number of questions that might signal information needs [55].

(5) **Community Feedback (2 features):** On Reddit, comments that are well-received by the conspiracy community are awarded upvotes while ill-received comments get downvotes [59]. In particular, we expect comments expressing dissonant views to receive negative community feedback. Hence, we calculate the comment score (an aggregation of upvotes and downvotes). To contextualize the feedback received, we also compute the synchronicity of the comment score with the score of its parent comment. To calculate synchronicity, we subtract the parent comment score from the child comment score.

(6) **Generic Document Level features (50 features):** Finally, we calculate *smooth inverse frequency* (SIF) document embeddings that capture the overall semantics of a comment by combining the embeddings of its words [4]. We calculate 50 dimensional SIF embeddings and use them as the baseline for evaluating the features discussed above.

In total, we calculate 483 features for every Reddit comment or post in the 12 QAnon subreddits. Next, we describe various sampling methods used to create a labeled dataset for classification.

6.2 RQ3a Method: Creating a Labeled Dataset

To understand belief and dissonance at scale, we need a large labeled dataset—ideally, the whole 1.2M comments in the study. We rely on a smaller, high quality labeled dataset that is manually vetted and use a high-precision classifier to extend the labeling to the rest of the data. Yet, even annotating the smaller dataset is challenging: labeling expressions of belief and dissonance in a community like QAnon is a complex and nuanced task which requires sizable theoretical background and expertise with the QAnon social imaginaries. Thus, we cannot rely on crowdworkers for the task. Moreover, self-disclosures of belief and dissonance are rare occurrences, with the great majority

of the QAnon subreddits revolving on discussing details of the theories and phatic talk, posing technical challenges to sampling informative instances of belief and dissonance to label. We address these challenges using a expert-in-the-loop mixed-methods approach [24].

A common strategy to make the most out of constrained labeling resources is active learning [72]. It lets an *interim* classifier choose which next comment would be the most informative if labeled, given the ones already labeled. The intuition behind it is that many comments are similar to each other (e.g., phatic comments making up the majority of the discussions), and labeling multiple instances would not offer a downstream classifier any new information. Instead, labeling a diverse set of comments would better serve the classifier to explore the variety in the whole subreddit. Especially, the comments about which classifier is the most uncertain at any point, are the ones that, if labeled, would most likely help it discern between classes in the future. Hence, given a pool of comments, active learning selects the most helpful one for an expert to annotate, and re-trains itself adding the newly annotated comment. This procedure repeats until the classifier performance converges, or until the annotation resources (the experts) are exhausted.

Through a pilot annotation, we found that the classes of interest—belief, dissonance, and neutral comments—are extremely imbalanced. Comments expressing disbelief, especially, amounted to only 6% of the comments in a random sample, while neutral comments amounted to 80%. Hence, we differentiate the pools of candidate comments to feed into the active learning loop, to trade off between exploring the large variety of comments in the whole dataset, and labeling a meaningful number of belief and dissonance comments. We select two distinct pools of comments: the first sampled at *random*, to represent data variety; the second sampled with a *biased* strategy to surface a higher number of instances of belief and dissonance. We perform active learning on each pool separately, and then use the comments labeled in both active learning runs to train the final classifier and extend the labeling to the whole dataset.

For the final classifier, we experimented with different aggregation techniques, from combining the labeled data and training a single classifier, to training classifiers separately on each labeled datasets and combining their predictions into a single score. The latter approach performs best on a held-out validation set. Figure 6 outlines the complete labeling and classification pipeline. Next, we discuss the details of sampling the random and biased data pools, performing active learning, and training the final classifier.

6.2.1 ***Creating Pools of Unlabeled Data.*** Because of the inherent disproportion of neutral comments in comparison to those expressing belief and dissonance in the QAnon subreddits, classifiers trained on such data may be less accurate on the latter classes [5] even after training on large labeled data [15]. In order to accurately model belief and dissonance in QAnon, we need a higher proportion of labeled instances of such classes. At the same time, to build a classifier that generalizes well to real data distributions, we also need labeled samples that represent the overall dataset. Hence, we first create two unlabeled sample pools—*random samples* and *biased samples*.

(1) **Random Unlabeled Sample:** Random sample contains 100K comments and posts selected uniformly at random. Random sampling can be representative of the dataset [45] where various types of expressions of belief and dissonance can occur at their natural frequencies inside the QAnon subreddits.
(2) **Biased Unlabeled Sample:** We use a cluster-based sampling technique to include belief, dissonance and neutral expressions in similar quantities [88]. Specifically, we perform K-Means clustering on the entire dataset and select samples that are closest and farthest from each cluster centroid. We use the elbow method—plotting explained variance in clustering as a function of number of clusters—to determine optimal number of clusters as 3. From every

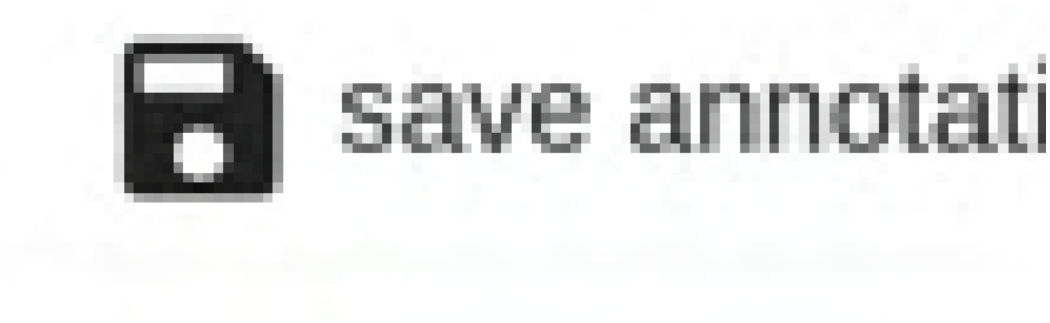


Fig. 7. Figure showing RQ2 annotation interface. Every comment and its parent comment is annotated with categories of QAnon social imaginaries. Annotators label every comment as belief, dissonance or neutral. The interface also displays naive classification scores updated every 10 samples.

cluster, we collect the 20K closest and farthest samples to each centroid, for a total of 120K posts and comments.

We use these two pools of unlabeled data to select the samples to label.

6.2.2 ***Labeling Dataset with Active Learning:*** Combined, the random and biased data contains 220K comments, which are still too large for complete manual labeling. To trade-off between the manual labor of labeling many comments and the downstream classification performance, which depends on a large enough labeled dataset, we devise an interactive labeling process based on active learning. An active learning classifier selects the next unlabeled comment to label according to its sampling strategy; after annotators label the comment, the classifier adds the comment to the labeled dataset, retrains itself, and selects the next comment to label [58]. This process repeats until the satisfactory classification performance is achieved. The crux of the learning strategy lies in sampling the new data to be labeled. We use a popular sampling strategy: uncertainty sampling.

Uncertainty Sampling for Active Learning: A common strategy for finding the best instances to label is to choose those unlabeled instances the classifier is currently most uncertain about. The intuition is that such instances would add the most information to the labeled dataset, and would therefore tighten the classification margin in the fewest iterations. One measure of uncertainty is entropy—a general measure of disorder in a system. In the context of classification, high entropy of predicted class probabilities indicates higher uncertainty of the classifier [40]. Based on the probability distribution $P(y_i|\mathcal{X})$, where y_i is the predicted probability of class i and $\mathcal{X}$ is the sample, entropy for each sample is calculated as:

$$entropy(\mathcal{X}) = -\sum_{i=1}^{C} P(y_i|\mathcal{X}) log_2(P(y_i|\mathcal{X})) \tag{1}$$

Interactive Labeling with Active Learning: Figure 7 showcases the web interface we built to facilitate interactive labeling with active learning. The interface displays the comment for annotation along with its parent comment for the context. Within the comments, phrases belonging to the QAnon Canon are highlighted and tagged with the dimension which they belong to. This helps annotators consider QAnon social imaginaries while identifying expressions of *belief* and *dissonance*, and to determine whether a comment is *irrelevant*. The first two authors of the paper annotated the whole labeled dataset. We use labeling guidelines detailed in Appendix A.1 to label the samples. The active learning classifier estimates its performance in cross-validation and on a held-out test set every 50 labeled samples. Accordingly, the interface also displays the current

change-adjusted balanced accuracy (random performance scores 0, perfect performance 1) and precision and recall for all three classes. We continue the annotation process until precision and recall surpass 0.60 for all classes. Through this process, we labeled 1,204 comments from the random sample and 1,167 comments from the biased sample.

6.3 RQ3a Method: Building Classifiers from Labeled Data

We use this labeled data to develop a classifier that reliably identifies expressions of belief and dissonance in the whole QAnon subreddits. In this section, we explain our selection of the classifier model, and details about its training and prediction procedures.

6.3.1 ***Selecting the Classifier Model:*** Although we labeled a total of 2,371 comments and posts from the random and biased samples, the labeled dissonance instances were still fewer compared to the belief and irrelevant. Since class imbalance may bias classifier performance, we balance the data. We undersample the data while maintaining above 0.6 precision for all classes. We experimented with various undersampling strategies and found the best performance with one-sided undersampling [44]. One-sided sampling removes the noisy, under-performing examples from the majority classes while preserving all the examples of the minority class. After undersampling, we end up with 336 and 900 samples from the random and biased labeled datasets respectively.

To choose best classifier we use the `auto-sklearn`[11] toolkit that searches over a wide variety of models optimizing for performance [31]. We find the optimal model for each of the two labeled datasets—namely, a Random Forest model [38] for the random labeled data and an Extreme Gradient Boosting (XGBoost) model [16] for the biased labeled data.

6.3.2 ***Training and Predicting with Classifier Models:*** We perform hyperparameter optimization to tune model parameters. Different models require tuning of internal parameters such as learning rate, estimators, etc., to generalize to unseen data [18]. We determine the best model parameters through an extensive grid search in a cross-validation scheme. We test over 15,000 combinations of hyperparameters, optimizing for chance-adjusted, balanced classification accuracy.

After fine-tuning both classifiers, we combine their predictions on the whole dataset. Specifically, we use a strict consensus pooling method to determine class assignments. Consensus pooling assigns a particular class (belief, dissonance, or neutral) to a sample if *both* classifiers agree on the predicted class. Disagreements are removed from the predictions. We also experimented with different pooled prediction strategies that do not require removing the ambiguous samples (see Appendix section A.2.1) with slightly lower scores. A single classifier trained on the combination of the two labeled datasets performs significantly worse.

6.4 RQ3b Method: Characterizing the types of Dissonance Self-Disclosures

Using the classifiers designed in the previous steps, we label the dataset of 1.2M comments and posts. We remove around 500K samples with prediction disagreements. In the remaining automatically labeled dataset, we find that over 43K comments and posts (6%) are labeled as dissonant. What are the different ways in which users express dissonance? We qualitatively analyze a random sample of 500 comments and posts expressing dissonance, focusing on how dissonance relates to the dimensions of QAnon social imaginaries, such as how they refer to collective practices and expectations. We report the results of the qualitative analysis in section 6.7.

[11] https://github.com/automl/auto-sklearn

	Random Forest (RF) Classifier				XGBoost Classifier (XGB)				RF+XGB Consensus	
	training		validation		training		validation		validation	
	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall
belief	0.61	0.46	0.55	0.41	0.60	0.54	0.66	0.58	0.71	0.54
dissonance	0.62	0.76	0.52	0.69	0.60	0.59	0.60	0.64	0.70	0.76
neutral	0.71	0.71	0.67	0.69	0.64	0.61	0.70	0.73	0.79	0.79
balanced accuracy	0.66		0.60		0.63		0.69		0.79	

Table 3. Training and validation performances for individual classifiers and final consensus classifier. The final consensus classifier clearly outperforms the individual classifiers across all precision scores.

6.5 RQ3a Results: Classification Performance

Table 3 shows the training and validation performance of the individual classifiers as well as the ensamble classifier consensus-pooling their predictions. Individual classifiers have comparable performances. RF shows 0.66 accuracy in the training set (cross-validated) and 0.60 accuracy on a held-out validation test, while XGB 0.63 and 0.69 respectively. The consensus classifier clearly outperforms both RF and XGB individually. In particular, precision exceeds 0.7 for all three classes. We further include distribution plots for validation precision and recall over 100 training iterations in Figure 10 in the Appendix. The distributions show the mean precision/recall values along with the standard deviations. Overall our results show relatively stable precision and recall values, lessening potential concerns of model overfitting.

6.6 RQ3a Results: Important Indicators of Belief and Dissonance

Which features are most impactful in classifying belief or self-disclosures of dissonance? Usually, important features in the model can be interpreted using classification coefficients. Such model explanations are easiest in binary classification where one coefficient can be interpreted in terms of either of the two classes. However, our classification consists of three classes: belief, dissonance and neutral. We need to understand how different features are impactful in classifying *each* of the classes. To understand feature importance, we use Multitask Elastic Net model, which solves three regression problems with three classes while sharing the same feature space. Hence, we can get feature importance separately for every class. For example, positive feature coefficients for belief class does not imply negative coefficient for the dissonance class. Figure 8 displays the most important features per class. Negative coefficients indicate that the low value of feature is associated with the presence of the class. For example, mention of "Q" is positively associated with both, belief (0.04) and dissonance (0.04) whereas, neutral comments have low mentions of "Q" (0.08). Interestingly, while both belief and dissonance have higher mentions of "trump" (0.01 and 0.01), only comments with belief mention "president trump" (0.02) indicating respect for QAnon *heroes*. Comments expressing belief also have higher proportion of *movement* related words ("wwg1wga" (0.03), "patriots" (0.02),) and higher positivity (compound VADER (0.02)). Dissonance disclosures however, mention more *expectations* related words ("arrests" (0.02), "predictions" (0.01), "moab" (0.01)). Dissonance self-disclosures also have higher integrative complexity (IC) score indicating the presence of multiple argumentative perspectives. In general, we find that semiotic language captured by QAnon Canon is strongly indicative of belief ("wwg1wga", "president trump", "patriots", "msm", "potus", "crumbs", "clowns", "obummer" etc.) and dissonance ("larp", "arrests", "cult", "sessions", "moab" etc.).

Fig. 8. Plot showing most predictive features for each class. Negative coefficients indicate that the low value of feature is associated with the presence of the class. For example "wwg1wga", a QAnon movement related word, is present more in the comments expressing belief (0.03 coefficient for belief) whereas dissonance and neutral comments have less occurrences of the same word (-0.015 for both). Note that the importance of features was separately calculated for each class using a Multitask Elastic Net model. Hence, the coefficients for features for each class are independent of the others. We annotate each feature with the feature category whenever applicable. Abbreviations for QAnon Canon features are; HERO:heroes, FOE: foes, MMT: movement, EXPT: expectations, PRCT: practices.

6.7 RQ3b Results: Points of Dissonance in QAnon

In the qualitative analysis of the dissonant comments, we examined how dissonance is expressed along the social imaginaries of QAnon. Table 4 lists various points of fracture in the QAnon social imaginaries along with example comments. Several users express dissatisfaction with various components of the QAnon movement. For example, some comments expressed how YouTubers profiting off of QAnon movement could harm the unity. Moreover, some users expressed concerns with overzealous nature of and the credibility of the others in the QAnon movement. Next, disappointment over Q's failed prophecies and unfulfilled promises is one of the primary point of dissonance for some users. Several users refer to specific phrases used by Q to express dissatisfaction over unmet expectations. While some users doubted the effectiveness of knowledge construction practices instructed by Q, distrusting the legitimacy and power of heroes is perhaps the most common point of dissonance in the QAnon followers. Some users express concerns over the mysterious identity of Q. We did not find many comments explicitly defending the enemies of the QAnon movement. However, some users expressed concerns over lack of evidence for vilifying deep state politicians. Some also expressed disappointment when Q deligitimized popular right wing celebrities such as Alex Jones.

Legitimacy of the movement	*...training your FOLLOWERS to put your voice before Q's - as we see from some - that can definitely be a problem*
	....this makes me question if this whole thing is even worth it..
	i knew this movement was going to shit when the "Cult of Q" occupied this sub.
	YOU NEED TO CHILL THE F' DOWN
Unfulfilled expectations	*Fool me once shame on Q, fool me twice.. Been following Q from beginning.*
	How many failed predictions does it take before you realize its bs?
	Ok, I'm getting off the Q-Anon train right here. He has consistently said, "Big news week", "Past Unlocks the Future",
	"Have Faith, Patriots are in Control", but nothing ever happens ! We have all been strung along like lemmings.
Ineffective practices	*this is just ridiculous. Q sends us down the rabbit hole, asks us to "dig deeper" but nothing makes sense!*
	In all fairness, Q has only dropped vague crumbs and we are supposed to find the truth off of that?
	Is it too much to ask for more proof?
Distrust in heroes	*I'm seriously starting to doubt Q is a White House insider.*
	Most of what he posts is easily found in the news and on conspiracy sites.
	Trump staff found holding devils signs in their hand.
	I think trump is controlled too by the jews/nwo..
Trust in foes	*Without more evidence than a few vague posts from Q,*
	I'm not going to believe that Barack Obama was sexually abusing that girl.
	I disagree with Q on the AJ [Alex Jones] matter. I have listened to infowars for years.
	What's the point in alienating people like Alex who are clearly on our side

Table 4. Types of dissonance related to social imaginaries in QAnon. In RQ3b, we conducted qualitative analysis of predicted dissonant comments. This table presents examples related to main observations.

7 RQ4: USER ENGAGEMENT AFTER DISSONANCE SELF-DISCLOSURE

Dissonance can induce various behavioral changes, such as strengthening commitment [29], recruiting others [28] or even reversing one's belief [29, 51]. Thus, we study how users change their engagement patterns within the QAnon subreddits after expressing dissonance.

7.1 RQ4 Method: Observing Changes in User Engagement After Dissonance

We use Interrupted Time Series (ITS) analysis to characterize user contributions before and after dissonance self-disclosure. ITS is a quasi-experimental statistical method that is used to analyze change in longitudinal data after an intervention or policy change. For example, researchers have used ITS to observe how dramatic events change user engagement in Reddit conspiracy communities [70]. Here, we consider dissonance self-disclosure as the *intervention* that determines changes in user contributions. With ITS, we can characterize whether, and by what degree, the trends in contributions after the intervention differ significantly from before. ITS analysis involves solving the following linear regression:

$$contributions \sim b_0 + b_1 T + b_2 D + b_3 P \tag{2}$$

where T represents the time step of the observation, D is a binary variable representing whether the time step is before or after the intervention and P encodes the time steps after the intervention. For example, in a model with time step of one week and an observation window starting 2 weeks prior to the intervention, the 2nd week after the intervention will be encoded as $T = 5$ (5 weeks since start of the observation window), $D = 1$ (after intervention), $P = 3$ (3 weeks after intervention, including the week of intervention). b_0 is the model intercept. The coefficient of T, b_1 indicates the slope of trend in the outcome variable (contributions in this case) *before* the intervention. b_2 indicates the change in level *starting at* the intervention whereas b_3 indicates the change in slope *after* the intervention. Therefore, the actual slope of trend after the intervention is derivable adding b_1 to b_3. The ITS regression directly indicates whether the pre-intervention trend b_1 and change in level at the intervention b_2 are statistically significant. Since ITS does not model directly the actual slope of trend after the intervention, but only the change with respect to the trend before, we corroborate its statistical significance via a separate piece-wise linear regression.

7.1.1 ***The ITS Setup:*** Given that the QAnon communities were banned within 11 months of their creation, a week is an appropriately short time-step to measure the immediate effects of the intervention. We define an observation window of total 13 weeks, centered at the intervention[12]. We compute the number of contributions (comments and posts) that users made each week in the QAnon communities, normalized by the number of contributions throughout their lifespan. In other words, each observation shows what fraction of contributions users made in QAnon communities within that specific week. Normalizing this way allows us to compare all users' contributions from the scale of 0 to 1. We employ several other robustness measures to ensure that users with inherently short contribution spans do not influence the analysis. For example, we consider only users who have at least one contribution before and after the 13-week observation window and also at least one contributions before and after the intervention within the observation window. We compare how user engagement changes after expressing dissonance within and outside the QAnon community by repeating the ITS analysis but including contributions made outside of the QAnon subreddits. As a further point of comparison, we also consider belief, instead of dissonance, as an intervention which may determine changes in engagement within the QAnon subreddits.

7.2 RQ4 Method: Correlating Dissonance and Belief with Tenure in the Community

Next, we turn to the question: how do self-disclosures of belief and dissonance affect users' permanence in the QAnon community? We set out to analyze the self-disclosures that users perform in their first 100 comments, and use them to predict their long-term tenure. The subreddits that host the community were active at different times (Figure 2); this may confound the analyses because users may stop posting either as a consequence of their experiences of dissonance, or because the subreddit they primarily posted in were banned. Therefore, we restrict the analyses to users who post on the subreddit `r/greatawakening`, which is the largest and longest running in our dataset. Next, we limit to users who contributed more than 100 comments to the subreddit, and characterize the disclosures of belief and dissonance in their first 100 comments. We compute the total number of disclosures of belief and dissonance as regressors. Moreover, to capture the relationship between belief and dissonance, we compute the dissonance index $\mathcal{D}$ as proposed by Festinger [28]. The dissonance index corresponds to the fraction of disclosures that is dissonant:

$$\mathcal{D} = \frac{dissonance}{dissonance + belief} \tag{3}$$

Thus, we add to the list of regressors, the average and maximum dissonance index $\mathcal{D}$. Moreover, as control variables, we compute the minimum, average, and maximum score of their comments, as well as the time of their first and 100th post. We select first and 100th post mainly because the users may leave the community due to the negative feedback by their peers, or because they post only infrequently to the subreddit. We standardize regressors and control variables and use them as factors determining future user permanence. We consider the last post in the subreddit as the time when the user leaves the community. To avoid confounds, we limit our observation to users who left the community at least one month before the subreddit was banned (a commonly used practice known as censoring in survival analysis [67]). We predict the number of remaining days that the user will spend on the subreddit. To this end, we use negative binomial regression which is suitable for ordinal outcomes (see Table 6 in Appendix). We also predict the number of remaining comments using ordinary least squares regression (log-scaled to account for skew, see Table 7 in Appendix), and the binary outcome of whether the user will remain in the community for more than 10 days via logistic regression (see Table 8 in Appendix). We report results for the subset of regressors that produce the best model fit.

[12] we also experimented with observation windows of 5, 7, 9, 11, 15, 17 and 21 weeks observing similar results.

Fig. 9. ITS plots for user contributions with different interventions. Two asterisks (**) indicate that the coefficient is statistically significant ($p < 0.05$). As shown in (a), immediately after the dissonance (week=0) there is significant decrease in the user contributions inside QAnon subreddits (b_2 = -0.02). However, as indicated in (c) there is no significant change in contributions after expressing belief. In the long term, contributions inside QAnon decrease with higher rate after expressing dissonance ($b_1 + b_3$ = -0.01) compared to belief ($b_1 + b_3$ = -0.004). Moreover, (b) and (c) indicate that there are no significant changes in user contributions outside of QAnon after expressing dissonance or belief.

7.3 RQ4 Results: Changes in User Contributions after Dissonance

To perform ITS analysis, We first identify users who express dissonance in the manually labeled dataset. We rely on the manually labeled dataset, rather than the larger but automatically labeled one, to be fully confident in our identification of dissonant comments and by association, dissonant users. Figure 9 displays the ITS and regression results. We find that user contributions inside QAnon decrease significantly (b_2 = -0.02) immediately after expressing dissonance (Figure 9 (a)). The same effect is *not* significant after the users express belief (Figure 9 (c)). Further, in the long term as well, contributions inside QAnon decrease at a higher rate after expressing dissonance (b_3 = -0.02) (Figure 9 (a)) compared to expressing belief (b_3 = -0.004) (Figure 9 (c)). Is this effect a byproduct of users reducing their overall Reddit activity? This does not appears be the case. We find no significant changes in the user activity outside the QAnon subreddits (Figure 9 (b) and (d)). While the analysis in Figure 9 is based on the 2,371 manually labeled comments spanning over 1,498 users, we also repeat the entire ITS analysis on the complete dataset of 700K comments with labels predicted by the classifier. We find similar results indicating that user contributions decrease significantly soon after dissonance (Figure 11 in Appendix).

7.4 RQ4 Results: Dissonance and Belief Predict Departure from the Community

Since users reduce their participation after expressing dissonance, we test whether self-disclosures of dissonance and belief ultimately lead to users leaving the community. We present here the three models predicting the number of remaining days that the user will spend on the subreddit (Table 6 in Appendix), the number of comments that the users will contribute in the future (Table 7 in Appendix), and whether the user will remain in the community for more than 10 days (Table 8 in Appendix). All three models indicate that a higher number of self-disclosures of dissonance correlates with a shorter tenure. There appears to be an asymmetric effect: whereas dissonance correlates with users leaving, belief does not correlate with them staying longer. While disclosures of belief are not significant correlates per se, it is important to consider them in combination with disclosures of dissonance: the maximum dissonance index $\mathcal{D}$ (Eq. (3)) experienced in the first 100 comments significantly correlates with users leaving the community in the following 10 days. These results support and extend those in section 7.3, showing that not only disclosures of disbelief are followed by a decrease in contributions, but also by the departure of the users from the community.

8 DISCUSSION

In this paper we uncover the dimensions and expressions pertaining to social imaginaries in a conspiracy community. Utilizing the expressions of social imaginaries, we identify how conspiracy community members express their belief and dissonance towards conspiratorial views discussed in the community. Further investigation of their dissonant communication, allows us to outline the points of fracture in their conspiracy belief system.

In RQ1 and RQ2, our analysis yielded novel dimensions and language correlates along which the QAnon conspiracy community aligns their social imaginaries. This led to characterizing QAnon social imaginaries along the conceptual dimensions of *movement*, *practices*, *expectations*, *heroes*, *foes*, and to formalizing the symbolic language used in conveying shared meanings across each dimension. We find that various such symbolic words are indeed important indicators of belief and dissonance in the QAnon community. For example, the phrases symbolizing QAnon *movement* such as "WWG1WGA" (where we go 1 we go all) and "patriots" are the top predictive features of belief expressions. Similarly, words related to *expectations* such as "arrests", "predictions", and "moab" are indicative of dissonant expressions. Moreover, we find that self-disclosures of belief and dissonance are consequential to understanding engagement in the QAnon conspiracy community. Engagement decreases immediately after self-disclosures of dissonance and dissonance experienced early on precedes departure from the community. Taken together, these results suggest that dissonance is followed by behavior change in the conspiracy community. In this light, we next discuss our empirical observation of how users manage their experiences of dissonance. We connect our findings to theoretical accounts of behavior change induced by dissonance.

8.1 Cognitive Dissonance Reduction Strategies

Our RQ3b results indicate that QAnon followers expressed dissonance about the legitimacy of the QAnon movement, unfulfilled expectations, ineffective practices and distrust in the heroes of the movement. What are the consequences of experiencing dissonance? Researchers posit that experiencing cognitive dissonance induces the state of psychological discomfort [28]. To deal with this discomfort, people employ several dissonance reduction strategies [28, 51]. For example, they trivialize the cause of dissonance and self-affirm their belief system [73]. Once an individual trivializes the point of contradiction, other discrepancies can no longer arouse dissonance. Consider this Reddit comment from our dataset for example:

> *I dont really care if Q is real. He has just reinforced my commitment to dig deeper. That's what a real Q dude would do no matter what. DIG DEEPER*

People also find strategies to rationalize the cause of dissonance [51]. Similarly, QAnon followers often rationalized Q's failed predictions by giving alternate explanations of the failures or creating more consonant interpretations of reality.

> *...[you] are not thinking critically in context of everything that comprises Q's message and content. He hasn't "failed". he's made statements that people have misinterpreted and then blamed on Q for the misinterpretations not being correct.*

More importantly, however, researchers state that cognitive dissonance can lead to attitude and behavior changes rejecting previously held beliefs [29]. Meaning, experiencing dissonance with conspiracies may lead people to abandon conspiracy beliefs and pave the way for recovery from conspiratorial worldview. Indeed, our analysis indicates that user contributions lowered after expressing dissonance, and that dissonance increased just before user's departure from the QAnon community. This is mirrored in the findings of our qualitative analysis. We found instances of users indicating that they were leaving the QAnon subreddit as a result of dissonance.

> *Q Anon is A psyop!!!! I am out, this board has been infiltrated. Something good, to something terrible, real quick. Don't fall for this Q stuff. Think for yourselves.*

> *This is the last one for me I think. I got hyped for "the memo" I got hyped for "raw footage". Arrests of the cabal within the week or I am out.*

While acknowledging that dissonance may not always lead to positive behavior change, our results suggest that exploring dissonance as a possible intervention for online conspiracy engagement is a promising future direction. Several other studies have explored dissonance based interventions [32] for reducing implicit racial prejudice [37] and promoting positive social behavior [52]. Below, we discuss how dissonance can be used as an intervention to motivate positive behavior change in conspiracy communities.

8.2 Intervention for Recovery from Online Conspiracy Discussion Engagement

In RQ3, we showed how users spontaneously disclosed experiences of dissonance and how this correlates with changes in behavior, especially focusing on the effects of disengaging from the conspiracy community, as demonstrated in RQ4. Dissonance can also be introduced externally as an intervention to *induce* such behavior. In fact, similar interventions based on "hypocrisy paradigm" [33], that encourage participants to explore the differences between their internally held beliefs and their public expression, have been tested in settings ranging from mental health through addiction recovery to prejudice reduction [32, 37]. For example, participants with high implicit racial prejudice were asked to write an essay on racial justice and fairness. Publicly expressing views contradictory to the implicit beliefs led the participants to reduce the prejudicial behavior [37]. Our results in RQ3b show that such "hypocrisy" exists in the QAnon conspiracy community. For example, the scenario where users want to be part of the QAnon community while at the same time dislike some aspects of the QAnon movement or doubt the QAnon heroes.

The methods in RQ1 and RQ2 offer ways to systematically compile social imaginaries from the point of view of the conspiracist themselves. This may help design community centered hypocrisy interventions based strategies that could nudge the conspiracists to explore the differences between the social imaginaries of the community and their internally held beliefs. For example, building on the qualitative analysis of fracture points in RQ3, one could build interventions that question the infallibility of heroes and the promises made by the movement leaders. Moreover, our results may also help contextualize past successful interventions within the social imaginaries of a specific

community and to recast them as hypocrisy interventions, such as questioning the efficacy of the movement to rigorously derive truth [19] or to be effective against foes [75].

Furthermore, the computational framework in RQ3 can be used to identify the central causes of dissonance in the community. These fracture points can be insisted or expanded via strategic interventions. Our results can also inform which interventions might not be successful. For example, in QAnon, "MSM" (mainstream media) is heavily distrusted and is considered as a foe. Interventions citing news articles from mainstream sources maybe met with instant criticism, despite their credibility. Finally, RQ4 offers ways to measure the outcomes of interventions, and therefore to select the ones that are most effective. In sum, conspiracy social imaginaries and computational dissonance detection offer powerful tools to design contextually-informed interventions.

8.2.1 ***Ethical Considerations.*** However, researchers need to consider social, psychological and ethical consequences of designing such intervention systems. For example, Sunstein and Vermeule argue that sowing the seeds of doubt in conspiracy theory communities is most (or perhaps only) effective when done from within the community [79]. Skepticism coming from outsiders may be deemed illegitimate, or even be construed as part of a larger conspiracy attempting to undermine the conspiracists' truth [49]. Dissonance causes psychological discomfort, and therefore interventions inducing dissonance should weigh harms against benefits. It is also important to consider *whether* certain conspiracies need intervention by accounting for the researchers' socio-political biases. These are but few of the ethical issues that intervention designers should consider before interacting directly with social media participants.

In sum, designing interventions for conspiracy communities is a complex socio-technical and even, a political problem that needs careful considerations of research ethics. Specifically, the design of hypocrisy based interventions described in Section 8.2 will require multi-disciplinary effort of computer scientists, social psychologists, and ethics and privacy experts, who could anticipate various technological, social and privacy implications of the interventions. A careful design of interventions will involve deriving least invasive strategies and systematically evaluating the intervention impact while monitoring for unintended consequences. We do not propose dissonance as an ultimate, or the only solution for recovery from conspiracy theories, but rather hope to initiate a dialogue about carefully designed interventions supported by our findings.

8.3 Conspiracy Belief as Collective Intelligence

Our findings in RQ3 and RQ4 enrich previous quantitative work on online conspiracy communities. Users who join conspiracy communities do so through contacts with existing members in conspiracy-neutral spaces [62]. On the one hand, there is empirical evidence of both self-selection of users into the conspiracy community and of shunning users from non-conspiracy communities. On the other hand, users who remain for a sufficiently long time, or *veterans* [70], may become overly committed. In RQ4, we show that there exists a critical time between users joining the community and committing to it in the long term, where arbitration in terms of self-disclosures and feedback received from the members affects user participation. Especially, users who exhibit heightened experiences of dissonance early on in their tenure in the community tend to leave, whereas users who remain may have successfully resolved existential conflicts between their beliefs and those of the community.

Hence, we suggest moving towards a view of conspiracy theorizing as a collaborative pursuit between individuals and communities. On the one hand, interventions should take into account both individual and collective beliefs, as well as the possible stages of the relationships between users and communities. On the other hand, the study of conspiracy theories may draw insights from research on arbitration between individuals and communities, such as community migration,

inter-community conflict, and moderation, to unpack the relationship between adopting conspiracy beliefs and joining conspiracy theory communities.

8.4 Implications for Studying Conspiracy Semiotics

Previous computational research studying language of conspiracy discussions mostly focuses on uncovering overarching narrative and argumentative elements in discourse [71]. Our methods can complement and enrich these techniques by revealing semiotic patterns in conspiracy discussions.

In our work, we propose a novel approach to uncover symbolic phrases that share common meanings inside the conspiracy communities. By combining manual verification with computationally creating dynamic representation of phrases, we automate the snowballing technique for uncovering symbolic language. Further by characterizing the insider language correlates across the dimensions of QAnon social imaginaries, we also provide a way for interpreting underlying meanings. Our methods and the lexicon QAnon Canon—the lexicon encoding semiotics in QAnon—together provide a toolbox for social-psychology and CSCW researchers to investigate language and expressions in QAnon like conspiracy communities within and across social media platforms. Although QAnon Canon is built from Reddit discussions, it can be useful in studying QAnon communities on other platforms as well, given that the general QAnon discourse is inspired by the same social imaginaries proposed in Q-drops. Specifically given the January 6th 2021 riots on U.S Capitol carried out by QAnon followers, our lexicon can be used to identify the dimensions of social imaginaries that were most prominent in mobilizing the collective action. Using the techniques proposed in this work, social computing researchers can further expand the lexicon to incorporate semiotics in other conspiracy theories.

9 LIMITATIONS AND FUTURE DIRECTIONS

Our work has some limitations that may offer directions for future research. First, the time frame of our study is constrained between the start of QAnon in 2017 to the ban of the QAnon subreddits in 2018. Since then, the QAnon movement has grown across various platforms such as Gab, Voat, and Facebook. Though our observations are temporally constrained, our analytical framework is general and could be applied to study dissonance self-disclosures across those different spaces and time frames. A further limitation concerns the procedure of extracting coded language from Q-drops. We relied on a variety of previous research and online sources to understand the social imaginaries surrounding such language. While this allows us to be confident about the phrases captured in the seed lexicon, it is possible that we overlooked some other coded words due to lack of context. In other words, we prioritized obtaining high precision over recall while building the QAnon Canon. Though time consuming, a potentially more comprehensive process could iteratively surface coded words from Q-drops and interpret them through the online discussion comments that reference them. Moreover, while our classifier gives precision above 0.7 in a complex three class classification problem, it relies on relatively simple features. Its performance would likely improve further after incorporating stylistic and meta-linguistic features and a larger labeled dataset. Finally, while our RQ4 analysis provides initial evidence about changes in user engagement following dissonance disclosures, we do not make causal claims. Studies aiming at deriving causal connections between disclosures of dissonance and behavior change should adopt controlled experimental designs and should account for potential confounders.

10 CONCLUSION

In this work, we studied the social imaginaries within the QAnon conspiracy theory community. We uncovered five dimensions of QAnon social imaginaries and created a lexicon capturing coded language across the five dimensions. We used this lexicon to identify self-disclosures of cognitive

dissonance inside the QAnon community and typified dissonant expressions along the social imaginaries. We further provided evidence that user contributions inside QAnon communities decrease immediately after self-disclosures of dissonance, and that high levels of experienced dissonance correlate with users ultimately leaving the communities. Our results show that users *do* express dissonance inside their communities and dissonance can be explored further as possible intervention for online conspiracy engagement.

11 ACKNOWLEDGMENTS

We want to acknowledge the valuable feedback from the members of Social Computing Lab at Virginia Tech and University of Washington, Seattle towards strengthening this work. This project was partially funded by the Minerva Research Initiative.

REFERENCES

[1] Max Aliapoulios, Antonis Papasavva, Cameron Ballard, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Jeremy Blackburn. 2021. The Gospel According to Q: Understanding the QAnon Conspiracy from the Perspective of Canonical Information. *arXiv preprint arXiv:2101.08750* (2021).

[2] Amarnath Amarasingam and Marc-André Argentino. 2020. The QAnon conspiracy theory: A security threat in the making. *CTC Sentinel* 13, 7 (2020), 37–44.

[3] Brian Arechiga. 2019. Mythic Pizza: Semiotic and Archetypal Significance in the Conspiracy Narrative Known as' Pizzagate'. (2019).

[4] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. (2016).

[5] Josh Attenberg and Foster Provost. 2010. Why label when you can search? Alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 423–432.

[6] Eileen Barker. 1999. *New religious movements: their incidence and significance.* Routledge.

[7] Julian Barnes. 2018. Blaming the Deep State: Officials Accused of Wrongdoing Adopt Trump's Response - The New York Times. https://www.nytimes.com/2018/12/18/us/politics/deep-state-trump-classified-information.html. (Accessed on 02/03/2021).

[8] Lee Basham. 2006. Malevolent global conspiracy. *Conspiracy Theories: The Philosophical Debate* (2006), 93–106.

[9] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 830–839.

[10] Ana-Maria Bliuc, David Best, Muhammad Iqbal, and Katie Upton. 2017. Building addiction recovery capital through online participation in a recovery community. *Social Science & Medicine* 193 (2017), 110–117.

[11] Prashant Bordia and Nicholas DiFonzo. 2004. Problem solving in social interactions on the Internet: Rumor as social cognition. *Social Psychology Quarterly* 67, 1 (2004), 33–49.

[12] Ben Collins Brandy Zadrozny. 2018. How three conspiracy theorists took 'Q' and sparked Qanon. https://www.nbcnews.com/tech/tech-news/how-three-conspiracy-theorists-took-q-sparked-qanon-n900531. (Accessed on 02/17/2021).

[13] Ben Collins Brandy Zadrozny. 2018. Reddit bans Qanon subreddits after months of violent threats. https://www.nbcnews.com/tech/tech-news/reddit-bans-qanon-subreddits-after-months-violent-threats-n909061. (Accessed on 02/17/2021).

[14] Michael Butter and Peter Knight. 2020. *Routledge handbook of conspiracy theories.* Routledge.

[15] Mausam C Lin. 2018. Active Learning with Unbalanced Classes & Example-Generated Queries. In *AAAI Conference on Human Computation.*

[16] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[17] Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16, 1 (1990), 22–29.

[18] Marc Claesen and Bart De Moor. 2015. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127* (2015).

[19] John Cook, Peter Ellerton, and David Kinkead. 2018. Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters* 13, 2 (feb 2018), 024018. https://doi.org/10.1088/1748-9326/aaa49f

[20] Steven Crimando. 2021. Q-Speak: The Language of QAnon. https://www.asisonline.org/security-management-magazine/latest-news/online-exclusives/2021/q-speak-the-language-of-qanon/. (Accessed on 04/12/2021).

[21] Michael R Czaja, Alan D Bright, and Stuart P Cottrell. 2016. Integrative complexity, beliefs, and attitudes: Application to prescribed fire. *Forest Policy and Economics* 62 (2016), 54–61.

[22] Marie-Catherine De Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational linguistics* 38, 2 (2012), 301–333.

[23] Ari Decter-Frain and Jeremy A Frimer. 2016. Impressive words: linguistic predictors of public approval of the US congress. *Frontiers in psychology* 7 (2016), 240.

[24] Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4529–4538.

[25] Umberto Eco. 1979. *A theory of semiotics*. Vol. 217. Indiana University Press.

[26] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of advanced nursing* 62, 1 (2008), 107–115.

[27] Anthony M Evans, Olga Stavrova, and Hannes Rosenbusch. 2021. Expressions of doubt and trust in online user reviews. *Computers in Human Behavior* 114 (2021), 106556.

[28] Leon Festinger. 1962. *A theory of cognitive dissonance*. Vol. 2. Stanford university press.

[29] Leon Festinger and James M Carlsmith. 1959. Cognitive consequences of forced compliance. *The journal of abnormal and social psychology* 58, 2 (1959), 203.

[30] Leon Festinger, Henry Riecken, and Stanley Schachter. 2017. *When prophecy fails: A social and psychological study of a modern group that predicted the destruction of the world*. Lulu Press, Inc.

[31] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2020. Auto-sklearn 2.0: The next generation. *arXiv preprint arXiv:2007.04074* (2020).

[32] Tanya Freijy and Emily J Kothe. 2013. Dissonance-based interventions for health behaviour change: A systematic review. *British journal of health psychology* 18, 2 (2013), 310–337.

[33] Carrie B Fried and Elliot Aronson. 1995. Hypocrisy, misattribution, and dissonance reduction. *Personality and Social Psychology Bulletin* 21, 9 (1995), 925–933.

[34] Marc Galanter. 2014. Alcoholics anonymous and twelve-step recovery: A model based on social and cognitive neuroscience. *The American journal on addictions* 23, 3 (2014), 300–307.

[35] Amanda Garry, Samantha Walther, Rukaya Rukaya, and Ayan Mohammed. 2021. QAnon Conspiracy Theory: Examining its Evolution and Mechanisms of Radicalization. *Journal for Deradicalization* 26 (2021), 152–216.

[36] John Heathershaw. 2012. Of National Fathers and Russian Elder Brothers: Conspiracy Theories and Political Ideas in Post-Soviet Central Asia. *The Russian Review* 71, 4 (2012), 610–629.

[37] Leanne S Son Hing, Winnie Li, and Mark P Zanna. 2002. Inducing hypocrisy to reduce prejudicial responses among aversive racists. *Journal of Experimental Social Psychology* 38, 1 (2002), 71–78.

[38] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. IEEE, 278–282.

[39] Richard Hofstadter. 2012. *The paranoid style in American politics*. Vintage.

[40] Alex Holub, Pietro Perona, and Michael C Burl. 2008. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1–8.

[41] Abraham Kaplan. 1943. Content analysis and the theory of signs. *Philosophy of Science* 10, 4 (1943), 230–247.

[42] Brian L Keeley. 1999. Of conspiracy theories. *The Journal of Philosophy* 96, 3 (1999), 109–126.

[43] Eva Kimminich. 2016. About grounding, courting and truthifying: conspiratorial fragments and patterns of social construction of reality in rhetoric, media and images.

[44] Miroslav Kubat, Stan Matwin, et al. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, Vol. 97. Citeseer, 179–186.

[45] PM Lance and A Hattori. 2016. Sampling and evaluation, a guide to sampling for program impact evaluation. *Chapel Hill, North Carolina: MEASURE Evaluation, University of North Carolina* (2016).

[46] Massimo Leone. 2017. Fundamentalism, Anomie, Conspiracy: Umberto Eco's Semiotics Against Interpretive Irrationality. (2017).

[47] Massimo Leone. 2019. The semiotics of common sense: Patterns of meaning-sharing in the semiosphere. *Semiotica* 2019, 226 (2019), 225–241.

[48] Massimo Leone, Madison Mari-Liis, Ventsel Andreas, et al. 2020. Semiotic Approaches to Conspiracy Theories. (2020).

[49] Stephan Lewandowsky, John Cook, Klaus Oberauer, and Michael Marriott. 2013. Recursive Fury: Conspiracist Ideation in the Blogosphere in Response to Research on Conspiracist Ideation. *Frontiers in Psychology* 4 (2013). https://doi.org/10.3389/fpsyg.2013.00073

[50] Howard Lune and Bruce L Berg. 2017. *Qualitative research methods for the social sciences*. Pearson.

[51] April McGrath. 2017. Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass* 11, 12 (2017), e12362.
[52] Blake M McKimmie, Deborah J Terry, Michael A Hogg, Antony SR Manstead, Russell Spears, and Bertjan Doosje. 2003. I'm a hypocrite, but so is everyone else: Group support and the reduction of cognitive dissonance. *Group Dynamics: Theory, research, and practice* 7, 3 (2003), 214.
[53] Daniel Taninecz Miller. 2021. Characterizing QAnon: Analysis of YouTube comments presents new conclusions about a popular conservative conspiracy. *First Monday* (2021).
[54] Tanushree Mitra, Graham P Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 126–145.
[55] Charles William Morris. 1938. Foundations of the Theory of Signs. In *International encyclopedia of unified science*. Chicago University Press, 1–59.
[56] Kim Mortimer. 2017. Understanding conspiracy online: Social media and the spread of suspicious thinking. *Dalhousie Journal of Interdisciplinary Management* 13, 1 (2017).
[57] CBS News. 2020. What is QAnon? What does WWG1WGA mean? The conspiracy theory that explains everything and nothing - CBS News. https://www.cbsnews.com/news/what-is-the-qanon-conspiracy-theory/. (Accessed on 02/04/2021).
[58] Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. (2009).
[59] Lisa Oswald and Jonathan Bright. 2021. How do climate change skeptics engage with opposing views? Understanding mechanisms of social identity and cognitive dissonance in an online forum. *arXiv preprint arXiv:2102.06516* (2021).
[60] Antonis Papasavva, Jeremy Blackburn, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. 2020. " Is it a Qoincidence?": A First Step Towards Understanding and Characterizing the QAnon Movement on Voat. co. *arXiv preprint arXiv:2009.04885* (2020).
[61] William Clyde Partin and Alice Emily Marwick. 2020. THE CONSTRUCTION OF ALTERNATIVE FACTS: DARK PARTICIPATION AND KNOWLEDGE PRODUCTION IN THE QANON CONSPIRACY. *AoIR Selected Papers of Internet Research* (2020).
[62] Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2021. What Makes People Join Conspiracy Communities? Role of Social Factors in Conspiracy Engagement. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–30.
[63] Charles Pigden. 1995. Popper revisited, or what is wrong with conspiracy theories? *Philosophy of the Social Sciences* 25, 1 (1995), 3–34.
[64] Toby D Pilditch, Jens K Madsen, and Ruud Custers. 2020. False prophets and Cassandra's curse: The role of credibility in belief updating. *Acta psychologica* 202 (2020), 102956.
[65] Samuel Planck. 2020. Where We Go One, We Go All: QAnon and Violent Rhetoric on Twitter. *Locus: The Seton Hall Journal of Undergraduate Research* 3, 1 (2020), 11.
[66] Daniel Priolo, Audrey Pelt, Roxane St Bauzel, Lolita Rubens, Dimitri Voisin, and Valérie Fointiat. 2019. Three decades of research on induced hypocrisy: A meta-analysis. *Personality and Social Psychology Bulletin* 45, 12 (2019), 1681–1701.
[67] Charles P Quesenberry Jr, Bruce Fireman, Robert A Hiatt, and Joseph V Selby. 1989. A survival analysis of hospitalization among patients with acquired immunodeficiency syndrome. *American journal of public health* 79, 12 (1989), 1643–1647.
[68] Alexander Robertson, Luca Maria Aiello, and Daniele Quercia. 2019. The language of dialogue is complex. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 428–439.
[69] Mike Rothschild. 2018. QAnon Followers Have Limited Options After Reddit Ban. https://www.dailydot.com/debug/qanon-movement-reddit-ban-voat-facebook/. (Accessed on 07/10/2021).
[70] Mattia Samory and Tanushree Mitra. 2018. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[71] Mattia Samory and Tanushree Mitra. 2018. 'The Government Spies Using Our Webcams' The Language of Conspiracy Theories in Online Discussions. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.
[72] Burr Settles. 2009. Active learning literature survey. (2009).
[73] Linda Simon, Jeff Greenberg, and Jack Brehm. 1995. Trivialization: the forgotten mode of dissonance reduction. *Journal of personality and social psychology* 68, 2 (1995), 247.
[74] Joshua Smithson. 2013. Gamechangers: A content and semiotic analysis of Super Bowl commercials during recession and non-recession years. (2013).
[75] Ana Stojanov, Jesse M. Bering, and Jamin Halberstadt. 2020. Does Perceived Lack of Control Lead to Conspiracy Theory Beliefs? Findings from an online MTurk sample. *PLOS ONE* 15, 8 (08 2020), 1–18. https://doi.org/10.1371/journal.pone.0237771

[76] Jeff Stone and Nicholas C Fernandez. 2008. To practice what we preach: The use of hypocrisy and cognitive dissonance to motivate behavior change. *Social and Personality Psychology Compass* 2, 2 (2008), 1024–1051.
[77] Nikolaos Stylianou and Ioannis Vlahavas. 2021. A neural entity coreference resolution review. *Expert Systems with Applications* 168 (2021), 114466.
[78] Peter Suedfeld and Philip E Tetlock. 1992. 27 Conceptual/integrative complexity. (1992).
[79] Cass R Sunstein and Adrian Vermeule. 2009. Conspiracy theories: Causes and cures. *Journal of Political Philosophy* 17, 2 (2009), 202–227.
[80] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
[81] Charles Taylor. 2004. *Modern social imaginaries*. Duke University Press.
[82] John B Thompson. 1984. *Studies in the Theory of Ideology*. Univ of California Press.
[83] Edward Tian. 2021. The QAnon Timeline: Four Years, 5,000 Drops and Countless Failed Prophecies - bellingcat. https://www.bellingcat.com/news/americas/2021/01/29/the-qanon-timeline/. (Accessed on 03/16/2021).
[84] Isaac Vaghefi and Hamed Qahri-Saremi. 2017. From IT addiction to discontinued use: A cognitive dissonance perspective. (2017).
[85] Robert Philip Weber. 1990. *Basic content analysis*. Number 49. Sage.
[86] QAnon Has Become The Cult That Cries Wolf. 2021. QAnon Has Become The Cult That Cries Wolf | FiveThirtyEight. https://fivethirtyeight.com/features/qanon-has-become-the-cult-that-cries-wolf/. (Accessed on 04/11/2021).
[87] Michael J Wood, Karen M Douglas, and Robbie M Sutton. 2012. Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science* 3, 6 (2012), 767–773.
[88] Weiwei Yuan, Yongkoo Han, Donghai Guan, Sungyoung Lee, and Young-Koo Lee. 2011. Initial training data selection for active learning. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*. 1–7.
[89] Ethan Zuckerman. 2019. QAnon and the emergence of the unreal. *Issue 6: Unreal* 6 (2019).

A APPENDIX

A.1 RQ3a Codebook used for Manually Labeling Samples

Our codebook was developed over multiple pilot labeling experiments where the authors went back and forth referring to Q-drops and Reddit comments to understand discourse in the QAnon community.

(1) **Belief:** It is belief when atleast part of the message shows support for Q/the movement/the subreddit/the social imaginaries or shows interest in the QAnon activities or shows strong opposition/dislike of the enemies of QAnon. It is not belief when the message presents belief in something outside of QAnon (aliens, religion).
(2) **Dissonance:** It is dissonance if at least part of the message is such that the speaker contradicts, doubts or expresses uncertainty about the social imaginaries or argues against the movement/redditors/a specific redditor when the latter defends the social imaginaries or proposes alternatives as being preferable to Q (god, other celebrities, politicians). It is not dissonance when the speaker confronts individual users without reference to the social imaginaries or argues for the relative/speculative nature of the social imaginaries, without denying its usefulness or truth.
(3) **Neutral:** The comment is neutral if the message does not contain any reference to beliefs/dissonance with respect to QAnon.

While labeling, we also skipped a few samples that we could not understand or interpret with confidence.

A.2 RQ3a Classification Additional Reports

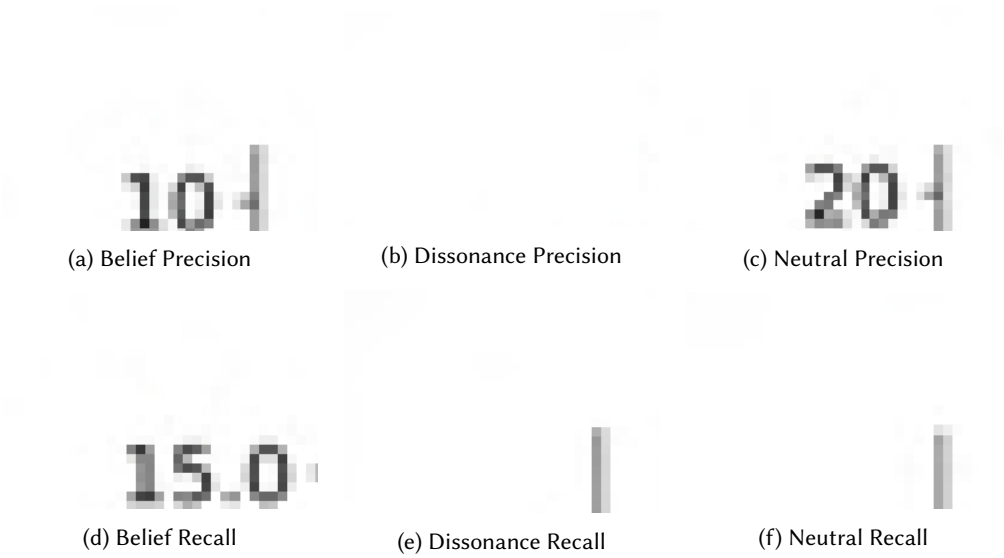
A.2.1 Experiments with various Prediction Pooling Strategies. While we select consensus pooling for the identifying dissonance, we also experimented with other pooling strategies described below. We also report the precision and recall results on validation set for each of the strategies in Table 5.

	RF+XGB Max Pooling		RF+XGB Avg Pooling		RF+XGB Stacking	
	validation		validation		validation	
	precision	recall	precision	recall	precision	recall
belief	0.63	0.59	0.62	0.61	0.61	0.63
dissonance	0.61	0.62	0.59	0.61	0.65	0.64
neutral	0.67	0.70	0.69	0.68	0.75	0.72

Table 5. Table showing validation performances with different prediction pooling strategies.

(1) **Max Pooling:** Consider the prediction with highest probability value of both classifiers
(2) **Average Pooling:** Calculate the average of class probabilities generated by both classifiers and consider the class with maximum averaged probability
(3) **Stacking:** Train a third classifier based on both, the predictions of two classifiers and all the features. Consider the predictions of the third classifier as final

A.2.2 Stability of Validation Performance. We tested the validation scores for final consensus classifier over 100 training iterations. Training tree based classifiers involves certain degree of randomness. Hence, we calculate the average precision and recall scores for all classes over 100 training iterations. Figure 10 displays the precision recall values for all classes over 100 training iterations. Overall, the plots suggests our validation scores are balanced and thus our model has less chances of overfitting.



(a) Belief Precision (b) Dissonance Precision (c) Neutral Precision

(d) Belief Recall (e) Dissonance Recall (f) Neutral Recall

Fig. 10. Validation precision and recall distributions of three classes over 100 training iterations. The dotted red line represents the mean and the shaded area around it represents the standard deviation.

A.3 RQ4 User Engagement Additional Reports

A.3.1 ITS Analysis of Dissonance in Predicted Data. While in RQ4 we present the ITS analysis of the labeled data, we also repeat the same experiment for the entire set of predicted comment. Figure 11 displays trends plotted using all predicted instances of dissonance and belief. The trends in the predicted data follow the same pattern as the results presented in Figure 9 in the main paper.

Fig. 11. ITS based on automatically labeled comments. We observe similar trends as manually labeled comments reported in Figure 9.

Dep. Variable:	days until user leaves	**No. Observations:**	1773
Model:	NegativeBinomial	**Df Residuals:**	1767
Method:	MLE	**Df Model:**	5
Date:	Wed, 14 Apr 2021	**Pseudo R-squ.:**	0.09317
Time:	10:52:21	**Log-Likelihood:**	-7910.1
converged:	True	**LL-Null:**	-8722.7

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	3.6545	0.014	261.392	0.000	3.627	3.682
born	-0.0648	0.018	-3.686	0.000	-0.099	-0.030
max(score)	0.0391	0.015	2.645	0.008	0.010	0.068
dissonance	-0.0301	0.014	-2.141	0.032	-0.058	-0.003
belief	-0.0090	0.014	-0.642	0.521	-0.037	0.019
created	-0.7951	0.019	-41.482	0.000	-0.833	-0.758
alpha	0.3141	0.013	24.610	0.000	0.289	0.339

Table 6. Negative binomial regression results, predicting in how many days the user will leave the community. The maximum comment score is positively correlated with longer permanence on the subreddit. Disclosures of dissonance, instead, are negatively correlated. Hence, disclosures of dissonance signal that users will leave the subreddit soon.

A.3.2 Regression Analysis of User Engagement . Here we present the results of regression analyses that further corroborate our observations of decreased engagement in the ITS analysis. We predict the number of remaining days that the user will spend on the subreddit (Table 6), number of comments user will likely make before exiting the community (Table 7) and the binary outcome of whether the user will remain in the community for more than 10 days (Table 8).

Received April 2021; accepted July 2021

Dep. Variable:	log(comments left)	**R-squared:**	0.507
Model:	OLS	**Adj. R-squared:**	0.504
Method:	Least Squares	**F-statistic:**	226.5
Date:	Wed, 14 Apr 2021	**Prob (F-statistic):**	3.15e-264
Time:	11:56:51	**Log-Likelihood:**	-2531.5
No. Observations:	1773	**AIC:**	5081.
Df Residuals:	1764	**BIC:**	5130.
Df Model:	8		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	4.6015	0.024	191.560	0.000	4.554	4.649
born	0.4513	0.029	15.552	0.000	0.394	0.508
min(score)	0.0608	0.024	2.519	0.012	0.013	0.108
max(score)	0.0608	0.024	2.484	0.013	0.013	0.109
avg($\mathcal{D}$)	-0.0185	0.059	-0.311	0.756	-0.135	0.098
max($\mathcal{D}$)	0.0339	0.052	0.655	0.513	-0.068	0.135
dissonance	-0.0691	0.033	-2.095	0.036	-0.134	-0.004
belief	0.0056	0.028	0.198	0.843	-0.050	0.061
created	-1.2059	0.029	-41.294	0.000	-1.263	-1.149

Omnibus:	60.154	**Durbin-Watson:**	1.950
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	81.072
Skew:	-0.354	**Prob(JB):**	2.49e-18
Kurtosis:	3.771	**Cond. No.**	4.99

Table 7. Ordinary least squares regression results, predicting the log-scaled number of comments that the user will post in the future. Similarly to the negative binomial regression, comment scores are positively correlated with more future comments, while disclosures of dissonance correlate negatively.

Dep. Variable:	remains after 10 days	**No. Observations:**	1773
Model:	Logit	**Df Residuals:**	1766
Method:	MLE	**Df Model:**	6
Date:	Wed, 14 Apr 2021	**Pseudo R-squ.:**	0.3708
Time:	11:03:12	**Log-Likelihood:**	-508.19
converged:	True	**LL-Null:**	-807.64

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	3.1499	0.170	18.528	0.000	2.817	3.483
born	-0.2703	0.080	-3.388	0.001	-0.427	-0.114
avg(score)	0.1447	0.075	1.938	0.053	-0.002	0.291
min(score)	0.1694	0.070	2.424	0.015	0.032	0.306
avg($\mathcal{D}$)	0.3087	0.173	1.786	0.074	-0.030	0.648
max($\mathcal{D}$)	-0.3363	0.169	-1.993	0.046	-0.667	-0.006
created	-2.4774	0.171	-14.523	0.000	-2.812	-2.143

Table 8. Logistic regression results, predicting whether the user remains in the community for more than 10 days after posting the 100th comment. The results corroborate the previous analyses. In particular, the negative coefficient for the maximum dissonance index $\mathcal{D}$ indicates that users who experience the highest dissonance are the ones who leave the community the soonest.