Conspiracy Narratives on Voat: A Longitudinal Analysis of Cognitive Activation and Evolutionary Psychology Features

Veronika Batzdorfer
veronika.batzdorfer@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

ABSTRACT

Understanding defining features of conspiracy narratives presents a significant challenge. Existing research predominantly focuses on online platforms as alternative information sources, overlooking the self-reinforcing dynamics influencing content creators. This study addresses this gap by investigating the role of prior cognitive activation, specifically users' engagement with conspiracy cues, in facilitating cognitive accessibility to such content. Central to our inquiry is the concept of *activation burden*, which refers to the cognitive effort required for individuals to engage with conspiracy narratives. We explore how features derived from evolutionary psychology, such as pattern recognition, detecting groups, or threat management, may contribute to the lowering of this activation burden, thereby fostering continued engagement with conspiratorial content.

To empirically examine these dynamics, we use data from Voat.co [26], a platform known for hosting de-platformed conspiracy-related discussions, sharing similarities in structure with Reddit. To characterize conspiracy content as multifaceted narratives, we investigate the utility of instruction-tuned Large Language Models (LLMs) for annotating text spans that represent various evolutionary facets of conspiracy features ($N = 3,384$, between 2014-06-20 and 2020-12-23). Our findings highlight the self-reinforcing effects of cognitive activation, indicating that users responding to *pattern*, *secrecy* and *threat* show carry-over effects and persistence of these features. The results could contribute to the understanding of antecedents of conspiracy beliefs and platform moderation practices, enhanced by LLM-annotation.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; • **Hardware → Emerging tools and methodologies**.

KEYWORDS

conspiracy beliefs, annotation, LLM, voat, self-reinforcement, volitional theory, multilevel vector auto-regressive model

ACM Reference Format:
Veronika Batzdorfer. 2024. Conspiracy Narratives on Voat: A Longitudinal Analysis of Cognitive Activation and Evolutionary Psychology Features. In *ACM Web Science Conference (WEBSCI '24), May 21–24, 2024, Stuttgart, Germany.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3614419.3644019

WEBSCI '24, May 21–24, 2024, Stuttgart, Germany

ACM ISBN 979-8-4007-0334-8/24/05
https://doi.org/10.1145/3614419.3644019

1 INTRODUCTION

The social web has changed how people consume and discuss information and form belief systems and opinions. Consuming information on online platforms has been sped up and democratised, circumventing traditional gatekeepers like journalists [4]. This comes at the expense of content accuracy, as social web platforms like Twitter or Reddit prioritise user engagement over information accuracy [1]. Forming beliefs in such contexts is essentially influenced by collective sensemaking [31]. [40] showed that trust in certain information depends on the number of postings or retweets by others active on the platform. Peer influence plays a vital role in amplifying "fake news", as people searching for social approval are more likely to share information [40]. Collective sensemaking in the form of common narratives is essential when official details on a salient event are absent or ambiguous (see [4]). Conspiracy theories are a form of collective sensemaking of salient events that reduce complexity by overinterpreting malevolent, secret intentions of powerful agents, bringing emotional relief to the individual [4, 36]. In the last decade, research sought to understand factors that might predispose people to interpret cues as threatening, such as network properties and cascade effects, individual differences in analytical thinking, cognitive skills or interpersonal distrust [19, 31]. A large body of research stressed the effect on those recipients exposed to conspiracy ideation but, to a lesser extent, how those who express such beliefs are affected by their behaviour (for self-reinforcement of political beliefs, see [11]).

We cast a broader look at self-regulation in expressing online conspiracy content and reacting to it. We focus on social web data as a source of naturally occurring behaviour traces to study individual changes with high ecological validity. More specifically, we study Voat.co, as this platform, which is similarly structured to Reddit, played a vital role for individuals in self-disclosing and spreading ideas after being de-platformed on other sides [26, 27]. On Voat.co, engaging with content with a low effort was done by *up-or-down voting* or *commenting*, responding to an original submission [26]. By voting, commenting on, and submitting conspiracy content, Voat.co represents a technological system to self-reinforce users through feedback loops of observing, reacting, and engaging with such content types. The present study investigates the lagged relationship between engagement with conspiracy features and posting behaviour longitudinally. By doing so, we emphasise the potential of a vicious circle between observing the content and actively generating conspiracy content by engaging in low-level reactive behaviour. Moreover, we investigate the suitability of identifying conspiracy narratives by five evolutionary psychology features.

The main contribution of the present paper is two-fold:

- We unobtrusively study the self-reinforcing cues for engagement with conspiracy content and self-disclosure.
- We present a multi-label conspiracy dataset suitable for assessing the multifaceted structure of conspiracy narratives, as well as evaluating the efficacy of annotation practices augmented by Large Language Models and their potential to assist in moderation. The data (and prompts) can be retrieved from: https://github.com/nika-akin/LLMs-and-conspiracy-beliefs

2 RELATED WORK

Impactful events (e.g., epidemics, natural disasters, economic crises, or terror attacks) are routinely echoed by behaviour on the social web [25, 33]. Besides expressing variants of affective reactions (e.g., empathy, worries, anger), a different type of response is expressing conspiracy theories that mainly thrive in uncertain crises and emotional shock [34]. In the context of uncertainty, caveats to the social web are the propagation of false information (intentionally or unintentionally), the generation of click baits (i.e., attention-seeking, exaggerated headlines), the self-reinforcement of conspiratorial clusters, or, taken further, the expression of pro-violence attitudes (i.e., political radicalisation) [15]. This increases the importance of research not only concerned with manifestations of different conspiracy beliefs but also investigating how individuals evolve to hold such beliefs and the role of motivational factors that lead to different self-expression types [31]. Motives for self-disclosure might range from maintaining a consistent self in the wake of social pressure or observing one's behaviour and self-constraining oneself to maintaining a consistent self [11]. Research so far has focused on established canonical conspiracy narratives based on top-down established motifs (e.g., the claimed governmental staging of the 9/11 terror attack) and argumentative styles (such as dogmatic epistemology) (see [2]). Hence, a considerable proportion of the literature is concerned with fact-checking efforts or establishing degrees of biases and political orientation of websites, which have shown limited effectiveness in curbing misinformation [24].

We address this bias of ex-ante defining known themes by drawing on the evolutionary framework by [39]. Based on the framework, conspiracy beliefs can be defined in the context of the human evolutionary adaptation to manage threats, detect secret, harmful alliances that pursue malicious motives, or recognise causal patterns which are essential for survival and learning [39]. In the context of conspiracy beliefs, these mechanisms have become dysfunctional. People overreact to cues that indicate **secrecy** (attempts to mislead and cover-up), powerful **actors** (a collective, individuals, institutions, personified entities), a causal **pattern** (connecting events or observations to an integrated whole), an evil motive or intent of an **action**, and a **threat** (a criminal event or danger of systemic harm) (see [12, 39]) (see Figure 1). They further fail to serve the epistemic (need to understand), existential (need for control) or social needs of the individuals [12, 39]. Elements like compensating for losing control [12] can be transferred to online platforms when the probability of encountering divergent facts is reduced even more as individuals self-reinforce previous opinions. Regarding the process of conspiracy theorising, different degrees of self-reinforcement can be differentiated.

2.1 Self-Reinforcement of Conspiracy Beliefs

Our research model for understanding self-expressions on Voat.co is based on premises derived from volitional theory and self-regulation theory (e.g., [23]). Initialising actions requires volitional effort—that is, investment of cognitive resources such as attention—and actions vary about this initialisation effort depending on their degree of routine and automatisation, attentional focus (vs. distraction or active avoidance of stimuli) or emotional processes involved (e.g., anger or enthusiasm) [3, 35]. By drawing on a volitional perspective, radicalisation processes can be regarded as a stepwise process, each step reducing volitional barriers to engaging further—either by increasing attention to the issue in question, avoiding information not fitting with a narrative (e.g., counter-information), increasing the emotional intensity, or developing routines with lower-intensity activities.

Applying this theoretical process to Voat.co's users allows to conceptualise comments, upvotes and downvotes as a low(er) effort activity and self-initialized submissions as a high(er) effort activity: in order to post a submission a specific number of upvotes by other users is required [26] and hence require more effort to deliberate and initialise the behaviour and not simply react to the content provided by other persons. A further implication of the presented theoretical perspective is that behaviour with a threatening cue can initialise a feedback loop and, thus, contribute to a vicious cycle. By threatening cues the person becomes further preoccupied with content, leading to an increased awareness of information provided by others of interest [31]. Likewise, based on traditional cognitive dissonance theory [14], it can be assumed that the person will more actively search for behaviour-fitting information and avoid contrasting information to prevent dissonance. This tendency to uphold and reaffirm prior beliefs leads to conspiracy-related content being more accessible, which will lower the necessary initialisation effort to produce different features of conspiracies [23, 35]. Based on these theoretical considerations, we formulate the following research questions:

RQ 1: Do specific features of conspiracy content (*threat*, *secrecy*, *action*, *actor*, or *pattern*) longitudinally reinforce themselves?

RQ 2: Do conspiracy features facilitate engagement, such as submissions, comments or votes of various intialization effort?

3 METHOD

3.1 Sample

We use a dataset from Voat.co covering the whole period from 2014 until 2020 that has been collected and made accessible by Mekacher and Papasavva [26]. The data set has been used to research the deplatforming effects of Reddit and migration to Voat.co and resultant community engagement and behaviour changes [27]. In the realms of these data, work has shown that conspiracy narratives gain importance in less moderated platforms [20], making the analysis of individual trajectories relevant. For annotation of the evolutionary conspiracy features, we randomly sampled submissions and comments from subverses (comparable to subreddits) that are presumably more conspiracy-related ($N = 1,880$): */v/anon*, */v/Conspiracy*, */v/GreatAwakening*, */v/pizzagate*, and */v/theredpill*. Moreover, more non-conspiracy-related subverses ($N = 1,504$) were sampled as well such as: */v/gaming*, */v/news*, */v/Science*, or */v/Showerthoughts*. Inclusion criteria involved (i) text length (min. 100 characters); (ii) distinct (text) body; (iii) excluding embedded links/external references.

3.2 Conspiracy Feature Annotation

In order to address the intricate problem of multi-label classification with the Voat.co dataset $N = 3,384$, we used OpenAI's GPT-3.5-turbo function calls with a temperature of 0.01 to annotate text spans of the five features indicative for conspiracy narratives (*action, actor, threat, pattern, secrecy*) (see Figure 1). This comprised three subtasks for each posting, (i) a binary labeling of the presence of each conspiracy feature, (ii) the text span detection of potentially overlapping labels, (iii) a binary label of the overall conspiracy-relatedness of the posting. Labeling was done based zero-shot instructions that explain each feature and discriminate it from each other. Prompt engineering was oriented at the best practices of [41] in, for instance, providing precise instructions after first introducing a context for each feature and developing prompts. Further,

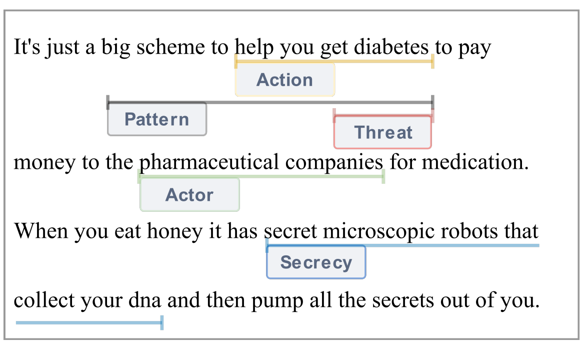

Figure 1: Example of a 5-class multi-label annotation of evolutionary conspiracy text spans (*action, actor, pattern, threat, secrecy*) with GPT-3.5

we conducted a 5-class multi-label classification using BERT (bert-base-cased), implemented with PyTorch Lightning to understand the usefulness of the conspiracy features for identifying conspiracy narratives. The model was fine-tuned using a linear classifier and trained with BCEWithLogitsLoss for 15 epochs, a learning rate of $2e - 05$ and a batchsize of 32.

3.3 Modelling Dynamics with Multilevel VAR Models

We employed multilevel VAR (vector autoregressive) models. VAR models belong to the broader class of autoregressive (AR) models, which have a long tradition in longitudinal research. While prominent forms such as the cross-lagged panel model [32], the random-intercept cross-lagged panel model [17] or combinations with growth curve models [6] are focused on a limited number of waves, VAR models extend the framework to time series with many ($T > 40$) waves. While the historical VAR models focus on a single time series of two variables based on $N = 1$, multilevel VAR models present a statistical framework for a nested sample of units. Briefly, a VAR model allows regressing a variable on former measures of a supposed cause. The model enables cross-lagged effects of both variables and is, thus, suited for the research question on feedback loops [18]. Furthermore, the model allows exploring the number of lags necessary for a variable to affect the other (and vice versa). From a causal perspective, VAR models represent "Granger causality" estimates [16], which are the estimated predictive effects while controlling for prior states of the dependent variable (i.e., the autoregression).

The model delivers *within-person* lagged regression coefficients, which inform about the within-person stability of the respective outcome (e.g., threat) in the form of an autoregressive parameter and the target cross-lagged effect. These are of substantial value as potential trait-like person-level confounding (e.g., due to personality) can be ruled out ("using the person as his or her control", [5] p. 71). Moreover, *within-person contemporaneous* relations reflect within-person partial correlation coefficients after adjusting for all autoregressive and cross-lagged effects. As such, these are correlations among the residuals of the variables after accounting for temporal effects. As within-person fixed parameters, they concern an average person's covariations of the residuals across time that may reflect the consequence of time-varying confounders or lagged effects that occur on a substantially smaller scale as the measured lag (in our case, three days). While these parameters concern fixed effects (i.e., effects for the average person), the mlVAR model also delivers random effects (standard deviations of the estimated within-person effects and relationship), allowing to evaluate interindividual differences in the effects and relationship.

We fitted competing multilevel autoregressive models with different temporal lags (lags 1-3). The optimal lag was evaluated using the Bayesian Information Criterion (BIC) [7], for which a lag 3 model showed the lowest BIC values for all variables in comparison.

4 RESULTS

In order to understand if the five features derived from evolutionary psychology are sufficient, we estimated a logistic regression of the labeled data set for the odds of being a conspiracy as defined by ChatGPT. The results indicate at least four evolutionary conspiracy features associated with the odds (OR) of a conspiracy and, hence, an overlap with the annotated ChatGPT notion of conspiracies. Notably, *threat* ($OR = 3.21$ [95% $CI = 2.45 - 4.21$]), *pattern* ($OR = 11.44$ [95% $CI = 9.13 - 14.40$]), *secrecy* ($OR = 15.15$ [95% $CI = 10.30 - 22.83$]) and *actor* ($OR = 5.35$ [95% $CI = 3.47 - 8.54$]) significantly increased the likelihood of the conspiracy label for a post (all $p < 2e - 16$). Conversely, *action* did not significantly influence the conspiracy notion ($OR = 0.95$, [95% $CI = 0.74 - 1.22$], $p = 0.7$). This could suggest that the instructional design for this feature may not be robust enough to clearly distinguish it from non-conspiracy content.

The results of the multi-label classification model point toward a similar direction. We show the performance for each conspiracy class, alongside aggregated measures across classes in Table

1. Notably, classes such as *secrecy* demonstrate high precision and recall scores resulting in an F1-score of 0.86. This suggests that the associated features are distinct and easily recognizable by the model. Similarly, for the *threat* class the model demonstrates a balanced performance, indicating that the model reliably captures instances labeled as *threat* while minimizing false positives. Regarding the *pattern* class, while the model achieves relatively high recall (0.82), indicating minimizing the number of instances missed (false negatives) the lower precision score (0.63) suggests a higher inclusiveness of false positives for the pattern class. In contrast, the *action* class shows a relatively low precision and recall score, indicating that while the model correctly identifies a substantial portion of instances labeled as *action* with moderate precision (54%), it tends to miss some relevant instances, resulting in a lower recall rate (41%). Therefore, model predictions are conservative, capturing only those instances where the model is highly confident, while potentially overlooking other instances that may also belong to the class due to the ambiguity or variability in features associated with this class. Lastly, regarding the actor class, with modest precision score of 0.51 and a recall of 0.65, the model indicates a higher tendency for false positives. Overall, the micro-average F1-score across all classes is 0.74, indicating a moderate balanced performance of the model across all classes.

Table 1: Multi-label results of the BERT classification model

	Precision	Recall	F1-score
Pattern	0.63	0.82	0.71
Secrecy	0.77	0.97	0.86
Threat	0.68	0.70	0.69
Action	0.54	0.41	0.47
Actor	0.51	0.65	0.57
Micro avg	0.68	0.81	0.74
Macro avg	0.63	0.71	0.66
Weighted avg	0.67	0.81	0.73
Samples avg	0.57	0.60	0.55

These results underscore the need for further refinement and investigation, particularly in improving the model's performance on categories with lower precision and recall values, such as *action* or *actor*.

4.1 Temporal Dynamics of Conspiracy Features

The estimated temporal lagged effects showed the significance of self-reinforcing conspiracy features. Regarding RQ1, we estimated fixed effects and random effects for the within-person autoregressive and lagged effects. This shows a significant positive autoregressive effects (ρ) for pattern, threat, and secrecy (ρ_{pattern} = .263, p < .001, ρ_{threat} = .087, p < .001, ρ_{secrecy} = .148, p < .001), respectively, indicating a carry-over effect and persistence of these features for a three-day period. Furthermore, there were significant cross-lagged effects for both variables pattern and secrecy (β_{pattern} = .236, p < .001, and β_{secrecy} = .178, p < .001), indicating that pattern predicted secrecy after three days and vice versa.

Regarding RQ2, the link between conspiracy features and user engagement cannot be established. Although, weak positive cross-lagged effects for secrecy and upvotes might suggest this low initialization effort (β_{secrecy} = .067, p = .017, the effect for threat and comments is negative (β_{threat} = .068, p = .013 suggesting that threat decreases the number of comments.

The standard deviations of the autoregressive coefficients and cross-lagged effects were low, suggesting little variation of these parameters across subjects. When controlling for the temporal effects, the correlations of the residuals for threat and pattern were positive and significant (r = .653, p < .001). Thus, threat and pattern co-occur simultaneously within the same measurement time (short-term temporal patterns), possibly reflecting additional processes faster than the specified lag-3 or due to co-occurring confounders.

5 CONCLUSION

This study investigated the temporal ordering of intraindividual differences in reacting to and producing conspiracy features. The results of our multilevel VAR model suggest a feedback loop of some conspiracy features such as *pattern*, *threat* and *secrecy*. Sustained and short-term dynamics of users can be differentiated from the results. On the one hand, from a sustained perspective, the within-person autoregressive effects indicate that features such as *threat* or *pattern* persist once they are initiated. In addition, we found a three-day feedback loop by estimating cross-lagged effects while controlling for the prior state of the dependent variable. On the other hand, the contemporaneous relations also showed a significant positive correlation for most conspiracy features, indicating either fast causal processes or omitted confounders [13, 21]. Yet, a direct link to facilitated user posting by individual features could not be established. These findings underscore the importance of reinforcing evolutionary psychology cues, such as pattern recognition and threat management, which are relevant in various contexts, including platform moderation.

In the realm of misinformation spread, it has been shown that users might divert attention to something other than the content's accuracy [30]. As social web platforms filter and suggest content based on previous user search history or preferences, further reinforcement opportunities for individuals are constituted to increase engagement. The role of algorithmic biases toward a particular content selection and repetition of content is one of the keys to lowering volitional barriers to persuasion [38]. Not only a repetition of information sources but also the repetition of the behaviour of social others might be causally relevant for the observed self-reinforcement loops. Experimental evidence from [9] found that a social reinforcement component contributes to behaviour travelling faster in networks. Besides content and social features, individual self-reinforcement, as a product of attentional biases, frames the reinforcement feedback loops of conspiracy content that might harden pre-existing beliefs (see [22]). One instantiation is confirmation bias if people seek information that reaffirms their prior beliefs to sustain congruence [22]. One relevant factor for lowering volitional barriers is the emotional valence of conspiracy content, as using emotional language increases the probability of contagion in social media contexts [8]. These contexts demonstrate the difficulty of moderation practices, correcting or removing content, blocking users, or debunking conspiracy content [10]. On the one hand, users overstate their confidence in the accuracy of their information [29]. Hence, corrections might backfire and even further harden the opinions believed to be accurate (conspiracy believers resist contrary evidence and migrate to other platforms that further foster isolation) [10]. On the other hand, debunking attempts are usually short-lived, and individuals tend to revert to the old behaviour, as our results suggest the short-lived nature of self-reinforcement in the online context. This renders top-down moderation attempts as interventions too coarse-grained. The importance of the credibility of peers in the online sphere seems much more promising. A first initiative would be to mark information as credible by peers before sharing content. This connects to our research results, as the importance of significant others to lower thresholds might play a role in processing threatening cues within this evolutionary psychology processing framework.

5.1 Limitations

In the realm of multilevel VAR models, [28] emphasize the challenge of causal inference owing to time-varying confounding variables. These lagged effect confounders may be for example (i) the degree of media exposure to political or social events, (ii) significant life events experienced by individuals, or (iii) fluctuations in psychological states, including stress and depression. This underscores a limitation in our study: the restricted number of covariates available to capture nuances of psychological dispositions. To enhance our understanding of individual variability, future investigations should consider incorporating additional covariates and quasi-experimental studies.

The potential impact of employing large language models to augment crowd worker labeling tasks related to conspiracy features requires exploration, similar to investigations conducted on large-scale dataset labeling [41]. Previous studies have demonstrated the efficacy of ChatGPT in tasks involving explanations [41]. Importantly, experts in the domain should validate the adequacy of the five features for annotating conspiracy content, as well as assess the validity of LLMs in labeling overarching conspiracy theories as there my be ostensible biases due to the dataset used [37]. This validation is essential for constructing both a general and specific dataset, facilitating the identification of multifaceted conspiracy narratives. Additionally, the data labeling process could be iteratively expanded with human annotators integrated in an active learning cycle ([41]).

REFERENCES

[1] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. 2020. Exposure to social engagement metrics increases vulnerability to misinformation. *arXiv preprint arXiv:2005.04682* (2020).

[2] Christian Baden and Tzlil Sharon. 2021. Blinded by the lies? Toward an integrated definition of conspiracy theories. *Communication Theory* 31, 1 (2021), 82–106. https://doi.org/10.1093/ct/qtaa023

[3] Albert Bandura. 1991. Social cognitive theory of self-regulation. *Organizational behavior and human decision processes* 50, 2 (1991), 248–287.

[4] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Trend of narratives in the age of misinformation. *PloS one* 10, 8 (2015), e0134641. https://doi.org/10.1371/journal.pone.0134641

[5] Niall Bolger and Jean-Philippe Laurenceau. 2013. *Intensive longitudinal methods: An introduction to diary and experience sampling research.* Guilford press.

[6] Kenneth A Bollen and Patrick J Curran. 2004. Autoregressive latent trajectory (ALT) models a synthesis of two traditions. *Sociological Methods & Research* 32, 3 (2004), 336–383. https://doi.org/10.1177/0049124103260222

[7] Laura F Bringmann, Nathalie Vissers, Marieke Wichers, Nicole Geschwind, Peter Kuppens, Frenk Peeters, Denny Borsboom, and Francis Tuerlinckx. 2013. A network approach to psychopathology: new insights into clinical longitudinal data. *PloS one* 8, 4 (2013), e60188. https://doi.org/10.1371/journal.pone.0060188

[8] Jason W Burton, Nicole Cruz, and Ulrike Hahn. 2021. Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour* 5, 12 (2021), 1629–1635. https://doi.org/10.1038/s41562-021-01133-5

[9] Damon Centola. 2010. The spread of behavior in an online social network experiment. *science* 329, 5996 (2010), 1194–1197.

[10] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26. https://doi.org/arXivpreprintarXiv:2009.11483

[11] Jaeho Cho, Saifuddin Ahmed, Heejo Keum, Yun Jung Choi, and Jong Hyuk Lee. 2018. Influencing myself: Self-reinforcement through online political expression. *Communication Research* 45, 1 (2018), 83–111. https://doi.org/10.1177%2F0093650216644020

[12] Karen M Douglas, Robbie M Sutton, and Aleksandra Cichocka. 2017. The psychology of conspiracy theories. *Current directions in psychological science* 26, 6 (2017), 538–542.

[13] Sacha Epskamp, Denny Borsboom, and Eiko I Fried. 2018. Estimating psychological networks and their accuracy: A tutorial paper. *Behavior research methods* 50 (2018), 195–212. https://doi.org/10.3758/s13428-017-0862-1

[14] L Festinger. 1962. A theory of cognitive dissonance, vol. 2. redwood.

[15] 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.

[16] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* (1969), 424–438.

[17] Ellen L Hamaker, Tihomir Asparouhov, Annette Brose, Florian Schmiedek, and Bengt Muthén. 2018. At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate behavioral research* 53, 6 (2018), 820–841. https://doi.org/10.1080/00273171.2018.1446819

[18] Jonas MB Haslbeck, Laura F Bringmann, and Lourens J Waldorp. 2021. A tutorial on estimating time-varying vector autoregressive models. *Multivariate Behavioral Research* 56, 1 (2021), 120–149. https://doi.org/10.1080/00273171.2020.1743630

[19] Sara Hughes and Laura Machan. 2021. It's a conspiracy: Covid-19 conspiracies link to psychopathy, Machiavellianism and collective narcissism. *Personality and individual differences* 171 (2021), 110559. https://doi.org/10.1016%2Fj.paid.2020.110559

[20] Vincenzo Imperati, Massimo La Morgia, Alessandro Mei, Alberto Maria Mongardini, and Francesco Sassi. 2023. The Conspiracy Money Machine: Uncovering Telegram's Conspiracy Channels and their Profit Model. *arXiv preprint arXiv:2310.15977* (2023).

[21] D Gage Jordan, E Samuel Winer, and Taban Salem. 2020. The current status of temporal network analysis for clinical science: Considerations as the paradigm shifts? *Journal of clinical psychology* 76, 9 (2020), 1591–1612. https://doi.org/10.1002/jclp.22957

[22] Daniel Kahneman, Patrick Egan, et al. 2011. Thinking, fast and slow: Farrar. *Straus and Giroux* 1 (2011).

[23] Julius Kuhl. 1992. A theory of self-regulation: Action versus state orientation, self-discrimination, and some applications. *Applied psychology* 41, 2 (1992), 97–129. https://doi.org/10.1111/j.1464-0597.1992.tb00688.x

[24] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131. https://doi.org/10.1177/1529100612451018

[25] Yu-Ru Lin, Drew Margolin, and Xidao Wen. 2017. Tracking and analyzing individual distress following terrorist attacks using social media streams. *Risk analysis* 37, 8 (2017), 1580–1605. https://doi.org/10.1111/risa.12829

[26] Amin Mekacher and Antonis Papasavva. 2022. " I Can't Keep It Up." A Dataset from the Defunct Voat. co News Aggregator. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1302–1311.

[27] Corrado Monti, Matteo Cinelli, Carlo Valensise, Walter Quattrociocchi, and Michele Starnini. 2023. Online conspiracy communities are more resilient to deplatforming. *arXiv preprint arXiv:2303.12115* (2023).

[28] Agnes Norbury, Shelley H Liu, Juan José Campaña-Montes, Lorena Romero-Medrano, María Luisa Barrigón, Emma Smith, Pedro Ortega Ana José Rodríguez-Blanco Lucia Segura-Valverde Marta Villalba Hospital, Madrid Alcón-Durán Ana Di Stasio Ezequiel García-Vega Juan Manuel Martín-Calvo Ana López-Gómez, Madrid Bañón-González Sara María Codesal-Julián Rosana Frade-Ciudad Ainara Hernando-Merino Elena Vallejo-Oñate Silvia Infanta Elena Hospital, et al. 2021.

Social media and smartphone app use predicts maintenance of physical activity during Covid-19 enforced isolation in psychiatric outpatients. *Molecular psychiatry* 26, 8 (2021), 3920–3930. https://doi.org/10.1038/s41380-020-00963-5

[29] Gerry Pallier, Rebecca Wilkinson, Vanessa Danthiir, Sabina Kleitman, Goran Knezevic, Lazar Stankov, and Richard D Roberts. 2002. The role of individual differences in the accuracy of confidence judgments. *The Journal of general psychology* 129, 3 (2002), 257–299.

[30] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780. https://doi.org/10.1177/0956797620939054

[31] 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. https://doi.org/10.1145/3432922

[32] David Rogosa. 1980. A critique of cross-lagged correlation. *Psychological bulletin* 88, 2 (1980), 245.

[33] Koustuv Saha and Munmun De Choudhury. 2017. Modeling stress with social media around incidents of gun violence on college campuses. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–27. https://doi.org/10.1145/3134727

[34] 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. https://doi.org/10.1609/icwsm.v12i1.15039

[35] Fritz Strack and Roland Deutsch. 2004. Reflective and impulsive determinants of social behavior. *Personality and social psychology review* 8, 3 (2004), 220–247.

[36] Cass R Sunstein and Adrian Vermeule. 2009. Conspiracy theories: Causes and cures. *Journal of political philosophy* 17, 2 (2009), 202–227. https://doi.org/10.1111/j.1467-9760.2008.00325.x

[37] Harry Thornburg. 2001. *Introduction to Bayesian Statistics*. Retrieved March 2, 2005 from http://ccrma.stanford.edu/~jos/bayes/bayes.html

[38] Christian Unkelbach, Alex Koch, Rita R Silva, and Teresa Garcia-Marques. 2019. Truth by repetition: Explanations and implications. *Current Directions in Psychological Science* 28, 3 (2019), 247–253. https://doi.org/10.1177%2F0963721419827854

[39] Jan-Willem Van Prooijen and Mark Van Vugt. 2018. Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on psychological science* 13, 6 (2018), 770–788.

[40] Quansheng Wang, Xue Yang, and Wanyu Xi. 2018. Effects of group arguments on rumor belief and transmission in online communities: An information cascade and group polarization perspective. *Information & Management* 55, 4 (2018), 441–449. https://doi.org/10.1016/j.im.201

[41] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? *arXiv preprint arXiv:2305.03514* (2023).

Received ; revised ; accepted