Conspiracy Theorising: Self-Reinforcing Feedback Loops through Online Expressions

Keywords: multilevel autoregressive models; conspiracy beliefs; Twitter; self-reinforcement; word embeddings

Extended Abstract

Social media have been shown as a virtual environment for conspiracy theories. Conspiracy theories are a form of collective sensemaking of salient events that reduce complexity by overinterpreting malevolent, secret intentions to powerful agents, bringing emotional relief and justification to the individual (Sunstein & Vermeule, 2009). As conspiracy beliefs can influence decision making (e.g., acceptance of vaccination), understanding initial conditions for adopting and spreading conspiracy content is important (Bertin, Nera & Delouvée, 2020). 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 (Barron et al., 2018; Hughes & Machan, 2021). Yet, research so far stressed the effect on those recipients exposed to conspiracy ideation but to a lesser extent how those who express such beliefs are affected themselves (see also Cho et al., 2018).

In the present study, we investigate whether prior cognitive activation of users (i.e., processing and passing conspiracy-related content) leads to such content being more readily accessible, hence resulting in lower activation burdens to produce own content. The present study investigates longitudinally the lagged bidirectional relationship between an individual's reactive diffusion of content and the active initiation of posting behaviour. By doing so, we emphasise the potential of a vicious circle between passively observing content and actively generating conspiracy content by engaging in low-level reactive behaviour such as retweeting observed content. As retweeting implies a low initiation effort, it accelerates further cognitive preoccupation with the content and lowers the barriers for more active self-initialization of conspiracy behaviour (e.g., actively generating and distributing conspiracy content).

The main contribution of the present paper is two-fold: (i) we unobtrusively study the short-term and long-term patterns of individuals in their natural environment (Jordan, Winer & Salem, 2020); (ii) by modelling effects from multiple subjects (i.e., fixed and random effects), we can identify self-regulation behaviour across persons, which can inform the development of platform moderation approaches.

We study a sample of Twitter users between January 2020 − November 2020. We used Twitter's REST API to sample the users' timelines (up to 3,200 tweets per user). As a result, we could identify $N = 109$ accounts. After sampling potential conspiracy-related accounts based on a keyword search, we manually annotated a set of tweets based on whether they matched at least three generic criteria out of five to be indicative for conspiracy theorising (measures were derived from agency, pattern, threat, secrecy, coalition, see Van Prooijen & Van Vugt, 2018) (see Fig.1). Thus, rather than taking a specific theory (e.g., antisemitic keywords) as indicative for conspiracy theorising, we concentrated on these minimal sufficient generic criteria from evolutionary psychology. Further, the inclusion of accounts was based on an activity threshold of users of at least three months, a rate of English tweet content of at least 80%, and an exclusion of bots with at least 80 % probability. Conspiracy content of tweets was quantified by means of representing words as word embeddings with GloVe (Pennington et al.,

2014) and quantifying the relative engagement with concept mover's distance (Stoltz & Taylor, 2019) (see Fig.2) and differentiating it from non-conspiracy content (see Fig.3). We used multilevel vector autoregressive time-series models to investigate: [H1] Retweeting conspiracy content longitudinally affects a person's own active conspiracy posting behaviour; and [H2] actively tweeting conspiracy content longitudinally increases a person's retweeting behaviour. Briefly, a VAR model allows regressing a variable on former measures of a supposed cause. By integrating traditional VAR models within the multilevel framework, researchers get the full array of between-person and within-person information for contemporaneous and lagged effects (Epskamp, Borsboom, & Fried, 2018).

We find evidence for both hypotheses. Our within-person lagged effects models suggest a shorter lag time and that short-term effects can build up reciprocally and the high standard deviations show that there are interindividual differences (i.e., some persons show stronger feedback loops). In conclusion, we argue that conspiracy beliefs form a spiral in which a person is triggered by others' content, resulting in a lower volitional barrier to becoming more radicalised. Not only repetition of information sources but also repetition of the behaviour of social others might be causally relevant for the observed self-reinforcement loops. As our results suggest the short-lived nature of self-reinforcement in the online context this renders top-down moderation attempts as interventions to be in a too wide time interval. The importance of the credibility of peers in the online sphere seems much more a promising path.

References

Barron, D., Furnham, A., Weis, L., Morgan, K. D., Towell, T., & Swami, V. (2018). The relationship between schizotypal facets and conspiracist beliefs via cognitive processes. *Psychiatry Research*, 259, 15–20. https://doi.org/10.1016/j.psychres.2017.10.001

Bertin, P., Nera, K., & Delouvée, S. (2020). Conspiracy beliefs, rejection of vaccination, and support for hydroxychloroquine: A conceptual replication-extension in the COVID-19 pandemic context. *Frontiers in Psychology*, *11*, 2471.

Cho, J., Ahmed, S., Keum, H., Choi, Y. J., & Lee, J. H. (2018). Influencing myself: Self-reinforcement through online political expression. *Communication Research*, *45*(1), 83-111. https://doi.org/10.1177%2F0093650216644020

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50, 195–212. http://dx.doi.org/10.3758/s13428-017-0862-1

Hughes, S., & Machan, L. (2021). It's a conspiracy: Covid-19 conspiracies link to psychopathy, Machiavellianism and collective narcissism. *Personality and Individual Differences*, *171*, 110559. https://dx.doi.org/10.1016%2Fj.paid.2020.110559

Jordan, D. G., Winer, E. S., & Salem, T. (2020). The current status of temporal network analysis for clinical science: Considerations as the paradigm shifts?. *Journal of clinical psychology*, *76*(9), 1591-1612. https://doi.org/10.1002/jclp.22957

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Stoltz, D. S., & Taylor, M. A. (2019). Concept Mover's Distance: Measuring concept engagement via word embeddings in texts. *Journal of Computational Social Science*, *2*(2), 293–313. https://doi.org/10.1007/s42001-019-00048-6

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures. Journal of Political Philosophy, *17*(2), 202–227. https://doi.org/10.1111/j.1467-9760.2008.00325.x

Van Prooijen, J.-W., & Van Vugt, M. (2018). Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on Psychological Science*, *13*(6), 770–788. https://doi.org/10.1177%2F1745691618774270

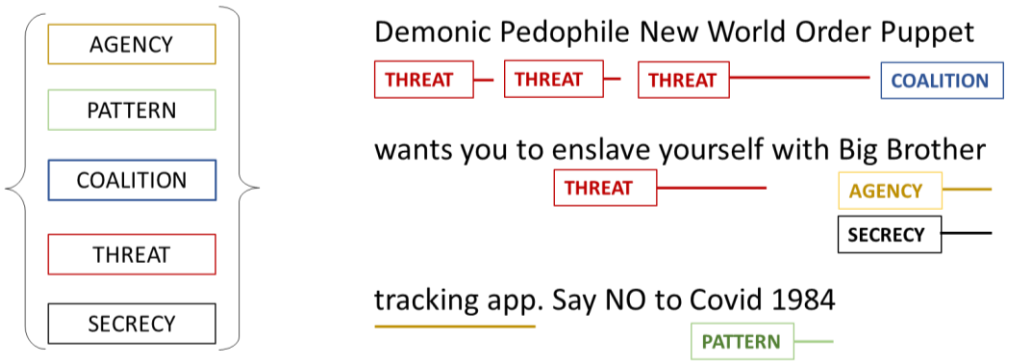

Figure 1. Manual annotation of tweets as potential conspiracy if containing five features (agency, pattern, coalition, threat, secrecy).

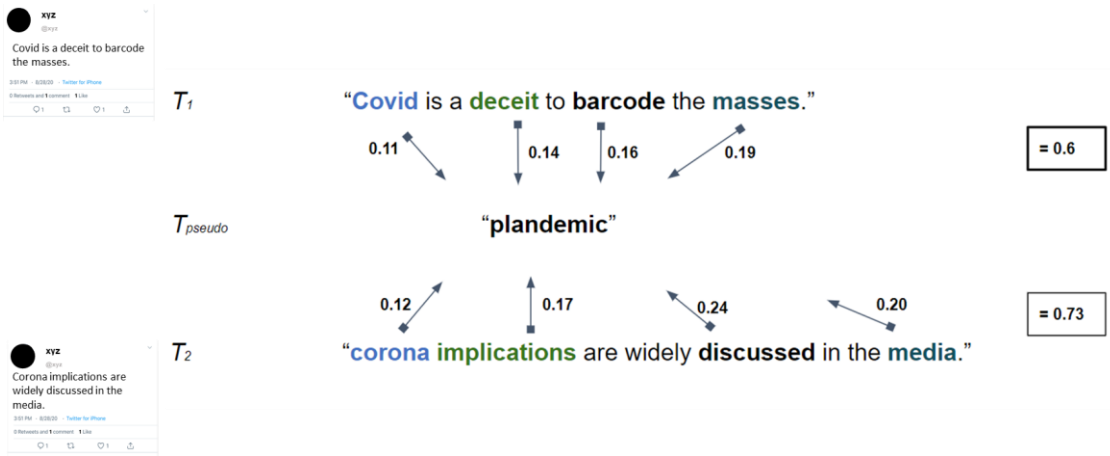

Figure 2. Relative engagement (i.e., cosine similarity) with target constructs (T_{pseudo}) of different word vectors (T_1, T_2) by means of concept mover's distance.

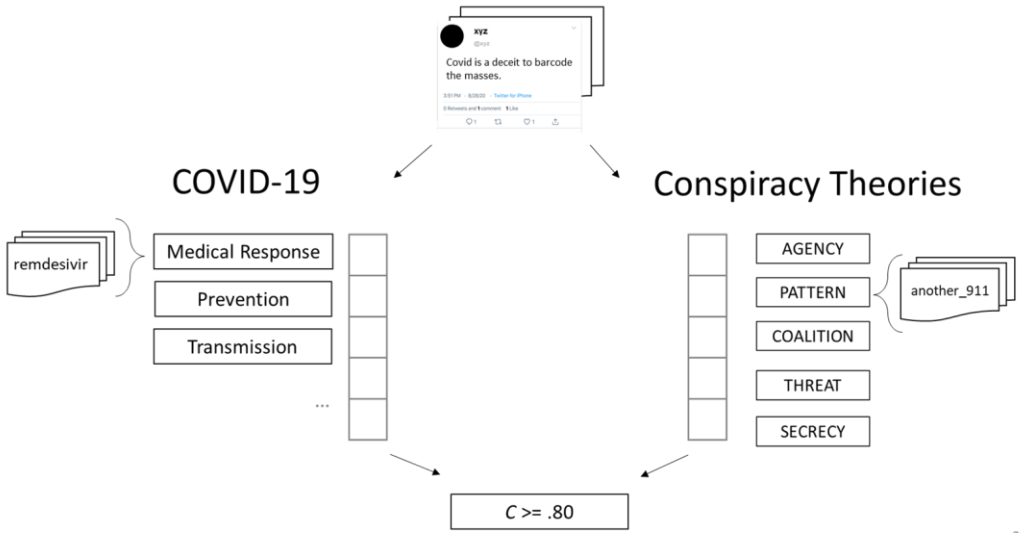

Figure 3. Differentiating conspiracy (i.e., seed terms for each of the five conspiracy features such as pattern) from non-conspiracy (i.e., based on a general COVID-19 related seed dictionary) word vectors by means of cosine similarity.