Do Prebunking and Debunking Protect Against Novel Misinformation?

Li Qian Tay[1,2], Mark Hurlstone[3], Tim Kurz[2], & Ullrich K. H. Ecker[2,4]

[1] School of Medicine and Psychology, Australian National University

[2] School of Psychological Science, University of Western Australia

[3] Department of Psychology, Lancaster University

[4] Public Policy Institute, University of Western Australia

Abstract

Misinformation can negatively influence belief formation and reasoning. To counter these influences, a range of countermeasures have been developed, the two dominant approaches being (proactive) prebunking and (reactive) debunking. While psychological research has revealed important insights into the efficacy of these interventions, the literature (1) lacks comparative analysis contrasting the two approaches, (2) has mostly used time-constrained designs that provide conducive intervention contexts, and (3) has focussed predominantly on cognitive measures, neglecting behavioural outcomes. We addressed these limitations in an experiment ($N = 793$) that provided two misleading articles on separate topics. We contrasted prebunking and debunking interventions that targeted the initial misinformation (article 1), and tested to what extent the interventions conferred protection against novel misinformation (article 2) either immediately or after a one-week delay. We included standard questionnaire measures of inferential reasoning as well as a text-generation task (writing a social-media post). With no delay, prebunking effectively protected against novel misinformation on both measures, while debunking effects were statistically significant only on the cognitive measure. Neither intervention was effective after a seven-day delay.

Public significance statement: This study demonstrates that both pre-emptive and retroactive interventions against misinformation can provide some protection against future misdirection when encountering

new misinformation. However, these protective effects did not survive a one-week delay between the intervention and exposure to the new misinformation.

Do Prebunking and Debunking Protect Against Novel Misinformation?

Misinformation has the potential to incur substantial costs on society (e.g., Lewandowsky et al., 2023; Treen et al., 2020; Tay et al., 2024). Although not a new phenomenon, recent technological advances and societal trends have likely exacerbated misinformation impacts (e.g., Lewandowsky et al., 2017). Importantly, from a psychological perspective, misinformation is particularly pernicious because of its potential to influence cognition even after clear corrections—a phenomenon known as the continued influence effect (Chan et al., 2017; Ecker et al., 2022). Much research has thus focussed on designing effective interventions (Kozyreva et al., 2022; also see Tay et al., 2023).

Two dominant intervention approaches are prebunking and debunking. Prebunking refers to interventions that seek to pre-emptively reduce individuals' susceptibility to misinformation. This includes interventions inspired by inoculation theory (Compton, 2013; van der Linden et al., 2020), which posits that misinformation susceptibility can be reduced via a warning about an impending threat to one's existing beliefs, combined with presentation of a specific piece of misinformation that is "weakened" by a refutation highlighting the misleading argumentation strategy applied. For example, illustrating how the tobacco industry used fake experts to challenge the scientific consensus on the adverse impacts of smoking can confer a protective effect that can help individuals resist climate misinformation that employs the same strategy (e.g., Cook et al., 2017).

Debunking refers to the retroactive correction of misinformation post-exposure. It is theorized that debunking leads to the coactivation of misinformation and correction in memory, which in turn facilitates updating and knowledge revision processes (Kendeou & O'Brien, 2014). To be maximally effective, corrections should (a) highlight any inconsistency between true and false information provided (Kendeou et al., 2019), (b) explain why the misinformation is misleading (MacFarlane et al., 2020), and (c) come from a trustworthy source (Ecker & Antonio, 2021; also see Lewandowsky et al., 2020; Paynter et al., 2019).

One limitation of the existing literature is that the two dominant approaches have largely been examined in isolation. Studies that have compared the two have produced mixed findings, with some favouring prebunking (Bolsen & Druckman, 2015; Jolley & Douglas, 2017) but others suggesting debunking to have a marginal advantage (Brashier et al., 2021; Bruns et al., 2023; Tay et al., 2022). Importantly, no research has directly contrasted the two approaches' ability to protect individuals against subsequent encounters with novel misinformation. This is the explicit aim of prebunking but is not usually considered in the debunking domain. Yet, given that both prebunking and debunking can involve detailed refutations and exposure of misleading strategies, the question of whether debunking may also offer protection against novel misinformation seems a natural one to ask. This is not a given—prebunking may only offer prospective protection to the extent that individuals encode the intervention with future-oriented expectations known to enhance memory encoding and recall (e.g., Klein et al., 2010).

By contrast, given the retrospective nature of debunking, it may lead individuals to focus on updating and revision of existing memory representations, which may result in shallow learning of the misleading strategies and limited transfer to future situations.

The aforementioned limitation is exacerbated by the fact that prior research has relied mostly on time-constrained designs, with studies evaluating misinformation and intervention impacts within a single experimental session (e.g., Cook et al., 2017; Rich & Zaragoza, 2016). Obviously, outside the laboratory, there can often be large time gaps between misinformation exposure and intervention, or between intervention and the potential implementation of the learnings by the recipient of the intervention. This means that protective effects observed in experimental studies may not transfer to real-world scenarios. Indeed, to the best of our knowledge, only two studies thus far have demonstrated delayed effects of prebunking (Maertens et al., 2020; Maertens et al., 2021). Likewise, although there are numerous studies that have investigated the delayed effects of debunking, they have explored how specific beliefs in debunked misinformation change over time (e.g., Swire et al., 2017), not whether the intervention provides lasting protection against novel misinformation.

Moreover, studies have mostly assessed misinformation reliance using questionnaire measures. For instance, given misinformation on a certain event (e.g., a fire), typical outcome measures include inferential-reasoning questions (e.g., “What might be a good headline for a report about the fire?”) or rating scales (e.g., “Negligence contributed to the

fire”; ‘strongly agree’ to ‘strongly disagree’; Connor Desai & Reimers, 2019; Johnson & Seifert, 1994). Some studies have used questions regarding behavioural intentions (e.g., the intention to share a particular piece of misinformation or engage in a relevant behaviour, such as getting a vaccine that has been the target of a misinformation campaign). Such measures provide valuable insights into individuals’ cognition but may not translate to the behavioural level (e.g., Kormos & Gifford, 2014; McEachan et al., 2011). Consideration of the behavioural level is important because misinformed behaviours can arguably be more harmful than misinformed beliefs.

In the current study, we sought to offer incremental progress addressing the above limitations. We presented participants with an article containing misinformation about fair trade—a topic selected to be vaguely familiar to most people, without many having strong opinions or detailed insights into its mechanisms and efficacy. We randomly provided participants with either no intervention, a pre-exposure prebunking, or a post-exposure debunking. We then provided an article featuring novel misinformation on nanotechnology and its use in sunscreen production (another topic selected to be only vaguely familiar). There was either no delay between the intervention and the novel misinformation or a delay of seven days. A no-misinformation control condition was also included, in which participants received only neutral, descriptive information about nanotechnology. In addition to a standard inference questionnaire, we assessed participants’ behaviour by asking them to compose a social-media post.

We tested three hypotheses: (1) misinformation exposure would increase participants' reliance on misinformation; (2) both prebunking and debunking would reduce novel-misinformation reliance in the absence of a delay between intervention and novel misinformation; and (3) prebunking would be more effective than debunking after a delay.

Method

The experiment adopted a 3 (condition: misinformation-only, prebunking, debunking) × 2 (delay: no-delay, delay) plus (no-delay) control between-subjects design. Ethics approval was granted by the University of Western Australia's Human Research Ethics Office.

Participants

We set a target of 120 participants per cell based on prior research (Tay et al., 2022) and a simulation-based power analysis using the R package *Superpower* (Lakens & Caldwell, 2021), assuming an ordinal interaction with $\eta_p^2 = .027$ ($1 - \beta \geq .80$, $\alpha = .05$). We thus aimed to recruit 840 participants via Prolific (U.S. residents with platform approval rate ≥ 98%). Excluding only participants that did not complete the entire study, our final sample size was $N = 793$ (female: 387; male: 394; non-binary: 12; $M_{age} = 43.63$; $SD_{age} = 14.34$). Participants received GBP 1.50.

Materials

Target Articles

Two target articles were presented in a fixed order: the first (adapted from Tay et al., 2022) presented initial misinformation falsely claiming that fair trade was a corrupt and ineffective movement; the

second introduced misinformation that was novel (in the context of the experiment), falsely claiming health risks associated with nanoparticles in sunscreen (the control condition used a version that contained only basic descriptive information about sunscreen nanoparticles). The two topics were chosen to balance relevance and experimental control, allowing us to construct credible yet misleading content with real-world implications, while by-and-large avoiding people having detailed knowledge or strong opinions. Articles used misleading strategies commonly employed by real-world fraudsters (e.g., use of fake experts and emotive anecdotes). See Supplement for both articles.

Intervention Article

The intervention article targeted the initial fair-trade misinformation, incorporating best-practice recommendations from the inoculation and continued-influence literatures. It warned participants of impending threat to their beliefs, exposed misleading strategies via examples, highlighted a trustworthy source, and included infographics adapted from Cook et al. (2017). It was provided before or after the fair-trade article in the prebunking and debunking conditions, respectively. The article is provided in the Supplement.

Outcome Measures

Misinformation Reliance. A five-item point-allocation questionnaire (e.g., Connor Desai & Reimers, 2019; Tay et al., 2022) was used to measure reliance on the nanotechnology misinformation. An example question was "*With regards to sunscreen regulations, which of the following changes would you support? Sunscreen manufacturers*

should (a) *increase the use of nanoparticles*; (b) *lower the price of sunscreens*; (c) *reduce or eliminate the use of nanoparticles*; (d) *offer a wider range of finishes (e.g., matte, dewy, etc.)*". Participants were told to allocate up to 10 points to the option(s) that best represented their opinion. Order of questions and response options was randomised. Points allocated to the misinformation-consistent options were summed across questions to create a misinformation-reliance score (range: 0-50, with greater scores indicating greater misinformation reliance).

Social-Media Post. Participants were asked to compose a short social-media post on nanoparticles, with friends, family, and followers as the intended audience. A minimum character count of 100 was enforced.

Procedure

After providing informed consent, participants read the target and intervention articles in the order defined by their randomly allocated condition (prebunking: prebunking, initial misinformation, novel misinformation; debunking: initial misinformation, debunking, novel misinformation). Depending on delay condition, there was or was not a seven-day delay prior to the novel misinformation (there was no delay in the control condition). The novel misinformation was followed by the misinformation-reliance questionnaire and the social-media task. Following best practice (Greene et al., 2023), a debriefing phase highlighted all false claims within the presented articles and provided web-links to additional resources for participants wanting to learn more about fair trade and nanotechnology. The study took approximately 11 min.

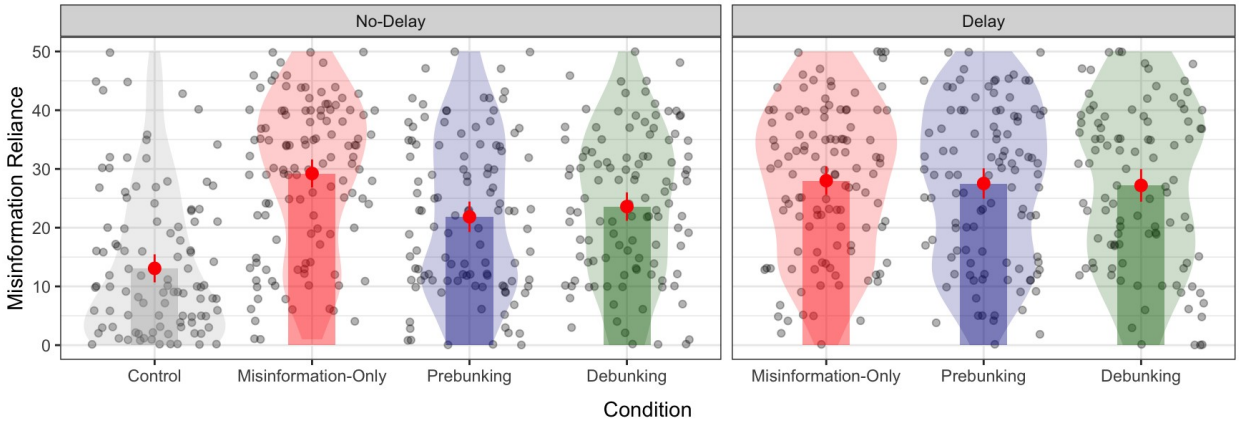
Results

Misinformation Reliance

Misinformation reliance across conditions is shown in Figure 1. First, as a "sanity check" and an initial test of our hypotheses, we focussed on the experimental conditions in a 3 (condition) × 2 (delay) between-subjects ANOVA. We found main effects of condition, $F(2, 671) = 5.49$, $p = .004$, $\eta_p^2 = .016$, and delay, $F(1, 671) = 6.66$, $p = .010$, $\eta_p^2 = .010$, and a significant interaction, $F(2, 671) = 3.91$, $p = .021$, $\eta_p^2 = .012$.

Figure 1

Novel-Misinformation (Nanoparticle Misinformation) Reliance Across Conditions



Note. Error bars represent 95% confidence intervals.

Pairwise contrasts are presented in Table 1. To test for misinformation effects, control was contrasted against misinformation-only in both delay conditions; both contrasts were significant, indicating an impact of nanotechnology misinformation on participants' reasoning both immediately (Contrast A) and after a delay (B). Both prebunking and debunking reliably reduced misinformation reliance when there was no

delay (C; E). However, when there was a delay, neither prebunking nor debunking conditions reliably differed from the misinformation-only condition (D; F). Next, to test for continued influence effects, control was contrasted against prebunking and debunking conditions. Notably, all contrasts were significant, suggesting that nanotechnology misinformation influenced participants' reasoning despite intervention both immediately (I; K) and after a delay (J; L), and that neither prebunking (I; J) nor debunking (K; L) could eliminate this influence. Additional Bayesian analyses lending support for the null (where no condition differences were observed) are reported in the Supplement.

Table 1

Planned Contrasts

		Misinformation Reliance		Social-Media Posts	
Contrast	Contrasted Conditions	*t*(786)	*p*	*z*	*p*
A	Misinformation-Only (No-Delay) - Control	9.217	<.001*	7.347	<.001*
B	Misinformation-Only (Delay) - Control	8.415	<.001*	5.894	<.001*
C	Prebunking (No-Delay) - Misinformation-Only (No-Delay)	4.198	<.001*	3.308	<.001*
D	Prebunking (Delay) - Misinformation-Only (Delay)	0.269	.788	0.337	.736
E	Debunking (No-Delay) - Misinformation-Only (No-Delay)	3.208	.001*	2.017	.044
F	Debunking (Delay) - Misinformation-Only (Delay)	0.444	.657	0.454	.650
G	Prebunking (No-Delay) - Debunking (No-Delay)	0.980	.327	1.369	.171
H	Prebunking (Delay) - Debunking (Delay)	0.175	.861	0.777	.437

I	Prebunking (No-Delay) - Control	4.957	<.001*	4.629	<.001*
J	Prebunking (Delay) - Control	8.030	<.001*	6.036	<.001*
K	Debunking (No-Delay) - Control	5.939	<.001*	5.849	<.001*
L	Debunking (Delay) - Control	7.793	<.001*	5.403	<.001*

Note. Contrasts for misinformation reliance relied on the full linear regression model, and contrasts for social-media posts relied on asymptotic *z*-tests over the full cumulative-link ordinal regression model. *** indicates post-correction significance with Holm-Bonferroni adjustments.

Social-Media Posts

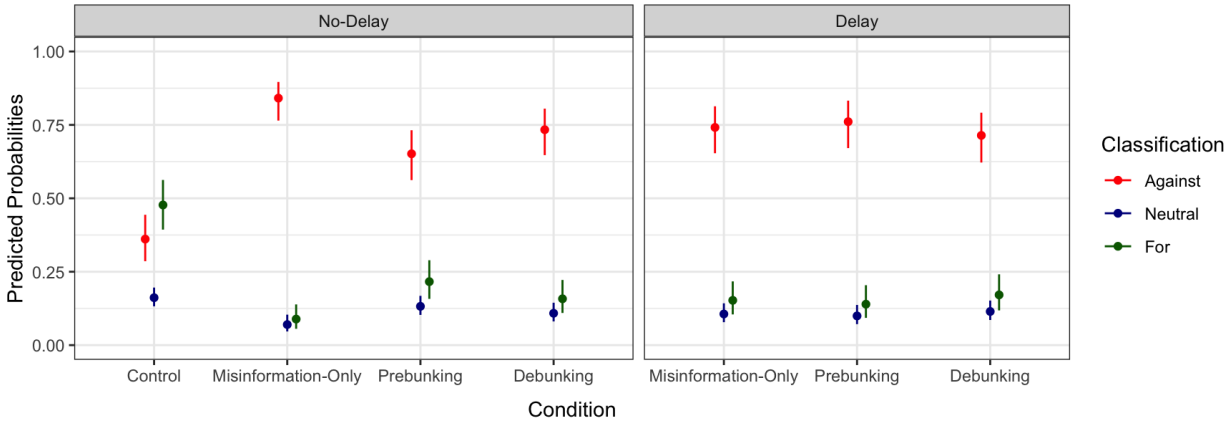
The raw text data of the social-media posts were analysed using large language models trained on large text corpora. We used two models, both implemented via the R package *text* (Kjell et al., 2023): a variant of Facebook's bidirectional autoregressive transformer (Lewis et al., 2020; Yin et al., 2019) and a decoding-enhanced version of BERT (He et al., 2019). We gave the models the natural-language-inference task of classifying posts into the categories "neutral", "for", and "against" use of nanotechnology.

Predicted classification probabilities across conditions are shown in Figure 2. Cumulative-link ordinal regression analysis revealed a significant main effect of condition, $\chi^2(2) = 10.78$, $p = .005$, but no significant effect of delay, $\chi^2(1) = 3.46$, $p = .063$. There was a significant interaction, $\chi^2\ (2) = 6.69$, $p = .035$. The full set of contrasts is presented in Table 1. Results revealed that nanotechnology misinformation had an impact on participants' social-media posts immediately (Contrast A) and

after a delay (B). The misinformation influenced participants despite intervention both immediately (I; K) and after a delay (J; L), with neither prebunking (I; J) nor debunking (K; L) eliminating this influence. In addition, only prebunking without delay reliably reduced misinformation reliance (C). We note that results were broadly consistent across the two models. See Supplement for further details and additional Bayesian analyses.

Figure 2

Predicted Probabilities for Social-Media Post Classification Across Conditions



Note. Against/neutral/for are in reference to the use of nanoparticles; error bars represent 95% confidence intervals.

Discussion

The present paper used both cognitive and behavioural measures to test for protective effects of prebunking versus debunking against novel misinformation across time. As expected, misinformation impacted participants' inferential reasoning as well as the social-media content they wrote; thus, there was some consistency across the two measures. When there was no delay between intervention and exposure to novel

misinformation, prebunking was effective in reducing misinformation reliance on both measures, although debunking was also effective for the cognitive measure. This is in line with prior studies showing the effectiveness of prebunking in conferring a general protective effect (e.g., Cook et al., 2017; van der Linden et al., 2020). In addition, our results offer the first evidence showing that debunking might similarly protect against novel misinformation, thus paving the way for theoretical integration between the two dominant intervention approaches. In this vein, further research could explore whether recommendations pertaining to prebunking based on inoculation theory may be applied to debunking (e.g., more focus on refutations of logical flaws may provide more general protection) and recommendations pertaining to debunking based on memory and learning theories may be applied to prebunking (e.g., incorporating techniques that may reduce forgetting of a prebunking intervention or information that may facilitate future intervention retrieval, for example by anticipating likely retrieval cues).

However, when a one-week delay was introduced prior to the novel misinformation, neither intervention was effective. This is inconsistent with studies showing lasting prebunking effects (Maertens et al., 2020; Maertens et al., 2021). There are at least three plausible reasons for such inconsistency. First, in Maertens et al. (2020), both prebunking and misinformation articles were on the topic of climate change. By contrast, our study investigated generalization to a distinct misinformation topic, which may have reduced ease of intervention retrieval. Moreover, Maertens et al. (2021) adopted a gamified approach and employed

within-participants repeated testing, both of which may have boosted participant engagement and facilitated encoding and content retention. Although speculative, these differences suggest possible boundary conditions for the efficacy of interventions after delays.

Regarding practical implications, our results suggest that prior research may have focussed on relatively conducive intervention contexts, such that findings may reflect the upper bounds of potential efficacy. This underscores the value of consistent reinforcement and tracking of interventions over time to ensure their effectiveness. Both researchers and practitioners should also work to explore and employ stronger or more diverse intervention methodologies, for example, based on the broader memory and learning literature, to identify conditions under which individuals might show more effective intervention encoding and retrieval strategies (see also Hennessee et al., 2019). This could also involve “booster” or combined interventions that build on earlier interventions (e.g., accuracy nudges used in conjunction with prebunking; Pennycook et al., 2023), the use of varied communication channels and innovative content formats, as well as tailored strategies based on specific target demographics (Whitehead et al., 2023). Above all, the crucial takeaway may be the need to stay proactive and guard against complacency.

Constraints on Generality

Several additional factors related to the generality of results may be investigated in future research. First, the current study relied on a sample of U.S. participants, and it is unclear whether the observed

patterns would apply in non-Western contexts with differences in perceived (and real) levels of misinformation in the environment and institutional trust, given that our topics related to the operations of commercial organisations (see also Tay et al., 2024). Second, as people tend to tailor contents of communication to the intended audience (e.g., Rudat et al., 2014), future research using open-ended content-generation tasks could vary the target of communication. In our case, participants were asked to think of friends and family, who may be associated with different thresholds for accuracy compared to a general audience. Third, our social-media measure was not "incentivised" in the sense that participants' posts were not actually posted online. Future research could attempt to include more realistic social-media simulations (e.g., Butler et al., 2023) or tasks that involve actual communication or shared decision-making processes between participants (e.g., Fay et al., 2021; Hurlstone et al., 2017). Fourth, one could systematically vary not only the interval between intervention and novel misinformation, but also the interval between the intervention and initial misinformation, to better capture real-world variation (also see Ecker et al., 2015). Finally, we focused on long-format, persuasive misinformation in the form of news articles (vs. short-format misinformation such as news headline). Our design thus precludes an analysis of the interventions' potential impacts on true information (e.g., Modirrousta-Galian & Higham, 2022), as asking participants to engage with both true and false news article in the large sets required to calculate discernment can be impractical and unreflective of real-world contexts. Future research could consider

replicating current design with novel but true information instead of misinformation to test for such ancillary impacts (see also Tay et al., 2023).

To conclude, our findings suggest both reason for optimism and cause for concern. Although the existing literature has already offered important insights into the psychological underpinnings of misinformation's influence and the efficacy of interventions, there remains an urgent need for both researchers and practitioners to further refine and test, and improve our understanding of, proposed interventions.

References

Bolsen, T., & Druckman, J. N. (2015). Counteracting the politicization of science. *Journal of Communication, 65*(5), 745-769. https://doi.org/10.1111/jcom.12171

Brashier, N., Pennycook, G., Berinsky, A., & Rand, D. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences, 118*(5). https://doi.org/10.1073/pnas.2020043118

Bruns, H., Dessart, F. J., Krawczyk, M. W., Lewandowsky, S., Pantazi, M., Pennycook, G., … Smillie, L. (2023, July 27). *The role of (trust in) the source of prebunks and debunks of misinformation. Evidence from online experiments in four EU countries.* PsyArXiv. https://doi.org/10.31219/osf.io/vd5qt

Butler, L. H., Lamont, P., Wan, D. L. Y., Prike, T., Nasim, M., Walker, B., Fay, N., & Ecker, U. K. H. (2023). The (Mis)Information Game: A social media simulator. *Behavior Research Methods.* https://doi.org/10.3758/s13428-023-02153-x

Chan, M. S., Jones, C. R., Jamieson, K. H., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science, 28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Compton, J. (2013). Inoculation theory. *The Sage Handbook of Persuasion: Developments in Theory and Practice*, 2, 220-237.

Connor Desai, S., & Reimers, S. (2019). Comparing the use of open and closed questions for Web-based measures of the continued-

influence effect. *Behavior Research Methods, 51*(3), 1426–1440. https://doi.org/10.3758/s13428-018-1066-z

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE, 12*(5), e0175799–e0175799. https://doi.org/10.1371/journal.pone.0175799

Ecker, U. K. H., & Antonio, L. (2021). Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition, 49*(4), 631–644. https://doi.org/10.3758/s13421-020-01129-y

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology, 1(*1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Fay, N., Walker, B., Kashima, Y., & Perfors, A. (2021). Socially Situated Transmission: The Bias to Transmit Negative Information is Moderated by the Social Context. *Cognitive Science, 45(*9), e13033–n/a. https://doi.org/10.1111/cogs.13033

Greene, C. M., de Saint Laurent, C., Murphy, G., Prike, T., Hegarty, K., & Ecker, U. K. H. (2023). Best practices for ethical conduct of misinformation research: A scoping review and critical commentary. *European Psychologist, 28*(3), 139–150. https://doi.org/10.1027/1016-9040/a000491

Haghighi, F., & Omranpour, H. (2021). Stacking ensemble model of deep learning and its application to Persian/Arabic handwritten digits recognition. *Knowledge-Based Systems, 220*. https://doi.org/10.1016/j.knosys.2021.106940

Hoes, E., Altay, S., & Bermeo, J. (2023). *Using ChatGPT to fight misinformation: ChatGPT nails 72% of 12,000 verified claims.* PsyArXiv https://psyarxiv.com/qnjkf/

Hurlstone, M. J., Wang, S., Price, A., Leviston, Z., & Walker, I. (2017). Cooperation studies of catastrophe avoidance: implications for climate negotiations*. Climatic Change, 140*(2), 119–133. https://doi.org/10.1007/s10584-016-1838-3

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Kendeou, P., & O'Brien, E. J. (2014). The Knowledge Revision Components (KReC) framework: Processes and mechanisms. In D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 353-377). MIT Press.

Kendeou, P., Butterfuss, R., Kim, J., & Van Boekel, M. (2019). Knowledge revision through the lenses of the three-pronged approach. *Memory & Cognition, 47,* 33-46. https://doi.org/10.3758/s13421-018-0848-y

Khered, A. S., Abdelhalim, I. Y. H. A., & Batista-Navarro, R. T. (2022, December). Building an ensemble of transformer models for Arabic dialect classification and sentiment analysis. *In Proceedings of the The Seventh Arabic Natural Language Processing.* http://dx.doi.org/10.18653/v1/2022.wanlp-1.53

Kjell, O. N. E., Giorgi, S., & Schwartz, H. A. (2023). The text-package: An R-package for analyzing and visualizing human language using natural language processing and transformers. *Psychological Methods.* https://doi.org/10.1037/met0000542

Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports, 12*(1), 3918–3918. https://doi.org/10.1038/s41598-022-07520-w

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the post-truth era. *Journal of Applied Research in Memory and Cognition, 6*, 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008

Klein, S. B., Robertson, T. E., & Delton, A. W. (2010). Facing the future: Memory as an evolved system for planning future acts. *Memory & Cognition, 38*(1), 13–22. https://doi.org/10.3758/MC.38.1.13

Kormos, C., & Gifford, R. (2014). The validity of self-report measures of proenvironmental behavior: A meta-analytic review. *Journal of Environmental Psychology, 40,* 359–371. https://doi.org/10.1016/j.jenvp.2014.09.003

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., … Wineburg, S. (2022, December 16). *Toolbox of interventions against online misinformation and manipulation.* PsyArXiv. https://doi.org/10.31234/osf.io/x8ejt

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920951503

Lewandowsky, S., Ecker, U. K., Cook, J., Van Der Linden, S., Roozenbeek, J., & Oreskes, N. (2023). Misinformation and the epistemic integrity of democracy. *Current opinion in psychology, 101711*. https://doi.org/10.1016/j.copsyc.2023.101711

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. arXiv. https://doi.org/10.48550/arXiv.1910.13461

MacFarlane, D., Tay, L. Q., Hurlstone, M.J., & Ecker, U. K. H. (2020). Refuting spurious COVID-19 treatment claims reduces demand and misinformation sharing. Journal of *Applied Research in Memory and Cognition. 10*(2), 248–258. https://doi.org/10.1016/j.jarmac.2020.12.005

Maertens, R., Anseel, F., & van der Linden, S. (2020). Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *Journal of*

Environmental Psychology, 70, 101455–. https://doi.org/10.1016/j.jenvp.2020.101455

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology. Applied, 27*(1), 1–16. https://doi.org/10.1037/xap0000315

McEachan, R. R. C., Conner, M., Taylor, N. J., & Lawton, R. J. (2011). Prospective prediction of health-related behaviours with the Theory of Planned Behaviour: a meta-analysis. *Health Psychology Review, 5*(2), 97–144. https://doi.org/10.1080/17437199.2010.521684

Modirrousta-Galian, A., & Higham, P. A. (2022, August 18). How effective are gamified fake news interventions? Reanalyzing existing research with signal detection theory. *PsyArXiv*. https://doi.org/10.31234/osf.io/4bgkd

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications, 13*(1), 2333–2333. https://doi.org/10.1038/s41467-022-30073-5

Pennycook, G., Berinsky, A., Bhargava, P., Cole, R., Goldberg, B., Lewandowsky, S., & Rand, D. G. (2023, August 21). *Misinformation inoculations must be boosted by accuracy prompts to improve judgments of truth.* https://doi.org/10.31234/osf.io/5a9xq

Rich, P. R., & Zaragoza, M. S. (2016). The continued influence of implied and explicitly stated misinformation in news reports. *Journal of*

Experimental Psychology. Learning, Memory, and Cognition, 42(1), 62–74. https://doi.org/10.1037/xlm0000155

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Rudat, A., Buder, J., & Hesse, F. W. (2014). Audience design in Twitter: Retweeting behavior between informational value and followers' interests. *Computers in Human Behavior, 35,* 132–139. https://doi.org/10.1016/j.chb.2014.03.006

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: comprehending the Trump phenomenon. *Royal Society Open Science, 4*(3), 160802–160802. https://doi.org/10.1098/rsos.160802

Tay, L. Q., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2022). A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *The British Journal of Psychology, 113*(3), 591–607. https://doi.org/10.1111/bjop.12551

Tay, L. Q., Lewandowsky, S., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2023). A focus shift in the evaluation of misinformation interventions. *Harvard Kennedy School (HKS) Misinformation Review.* https://doi.org/10.37016/mr-2020-124 https://doi.org/10.37016/mr-2020-124

Tay, L. Q., Lewandowsky, S., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2024). Thinking Clearly About Misinformation. *Communications Psychology.*

Treen, K. M. d'I., Williams, H. T. P., & O'Neill, S. J. (2020). Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change, 11*(5), e665–n/a. https://doi.org/10.1002/wcc.665

Umair, A., & Masciari, E. (2023). Sentimental and spatial analysis of covid-19 vaccines tweets. *Journal of Intelligent Information Systems, 60*(1), 1-21. https://doi.org/10.1007/s10844-022-00699-4

van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.566790

Whitehead, H. S., French, C. E., Caldwell, D. M., Letley, L., & Mounier-Jack, S. (2023). A systematic review of communication interventions for countering vaccine misinformation. *Vaccine, 41*(5), 1018–1034. https://doi.org/10.1016/j.vaccine.2022.12.059

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. arXiv. https://doi.org/10.48550/arXiv.1909.00161