Written evidence submitted by Cogito Epistemology Research Centre, University of Glasgow

White Paper

Trustworthiness, Disinformation, and Evidence Resistance

Professor Mona Simion & Professor Christoph Kelp, Chair in Epistemology
Cogito Epistemology Research Centre, University of Glasgow

1. About the Cogito Epistemology Research Centre

The Cogito Epistemology Research Centre at the University of Glasgow (https://www.cogito-glasgow.com) is a leading research centre internationally in epistemology (theory of knowledge) research, i.e. research on e.g. rational belief, rational trust and trustworthiness, the nature of evidence, knowledge, understanding, evidence resistance, irrationality, information/misinformation/disinformation, digital knowledge, norms for scientific communication, disagreement, epistemic risk, and group knowledge. Cogito is leading the European Consortium for Knowledge and Information Research (ECKIR http://www.knowledge-consortium.com), it is hosting the British Society for the Theory of Knowledge, it is part of the world-wide Social Epistemology Network, and the European Epistemology Network.

Cogito currently hosts several prestigiously funded major research projects in epistemology, including:

- *Knowledge-Lab: Knowledge-First Social Epistemology* (https://www.knowledgelab-research.com funded by the European Research Council; PI Professor Mona Simion (Glasgow)): investigating ways of acquiring knowledge and rational belief from social sources (the testimony of others, disagreement, groups, mass media).
- *A Virtue Epistemology of Trust* (http://trust-well.com funded by the Leverhulme Trust; PI Dr. Adam Carter, Co-Is Professor Mona Simion and Professor Christoph Kelp (Glasgow)): investigating the nature of rational trust and trustworthiness.

- *Evidence: Knowledge and Understanding* (https://www.cogito-glasgow.com/evidence-knowledge-understanding funded by the Swiss National Science Foundation; PIs: Professor Christoph Kelp (Glasgow), Professor Anne Meylan (Zurich): investigating the nature and normativity of evidence.
- *Digital Knowledge* (https://www.digital-knowledge.org funded by the Arts and Humanities Research Council, UKRI; PI: Dr. Adam Carter (Glasgow), Co-I Professor Jesper Kallestrup (Aberdeen): investigating knowledge acquisition in digital environments.
- *Expanding Autonomy: Scaffolded, Embedded, and Distributed* (funded by the Arts and Humanities Research Council, UKRI; PI: Professor Neil Levy (Oxford) and Dr. Adam Carter (Glasgow): investigating autonomous belief formation in digital environments.
- *KnowSoc* (funded by the European Research Council; PI Professor Christoph Kelp (Glasgow), Co-Is Professor Mona Simion (Glasgow), Dr. Adam Carter (Glasgow), Professor Esa Diaz Leon (Barcelona), Professor Thomas Grundman (Cologne), Professor Sven Bernecker (Cologne and UC Irvine), Professor Anandi Hattiangadi (Stockholm), Professor Frank Hofmann (Luxembourg), Professor Rene Van Woundenberg (Amsterdam), Professor Mikkel Gerken (Southern Denmark), Professor Jesper Kallestrup (Aberdeen): investigating knowledge acquisition in social settings.

In what follows, we outline our recent research results on the nature of disinformation, rational trust and trustworthiness, and evidence resistance, together with research-informed recommendation for policy and practice. The aim is to inform policy makers on the best strategies to adopt for fighting disinformation campaigns and evidence resistance via trustworthy voices.

2. Introduction

Disinformation is widespread and harmful, epistemically and practically. We are currently facing a global information crisis that the Secretary-General of the World Health Organization (WHO) has declared an 'infodemic' https://www.who.int/health-topics/infodemic#tab=tab_1 Furthermore, crucially, there are two key faces to this crisis, i.e. two ways in which disinformation spreads societal ignorance: One concerns the widespread sharing of disinformation (e.g., fake cures, health superstitions, conspiracy theories, political propaganda etc.) especially online and via social media, which contribute to dangerous and risky political and social behaviour. Separately, though at least as critical to the wider infodemic we face, is the prevalence of resistance to evidence: even when the relevant information available is reliably sourced and accurate, many information consumers fail to take it on board or otherwise resist or discredit it, due to the rise in lack of trust and scepticism generated by the polluted epistemic environment – i.e. by the ubiquity of disinformation. What we need, then, is an understanding of how to help build and sustain more resilient trust networks in the face of disinformation. To this effect, we need a better understanding of: the nature of rational trust and trustworthiness, the nature and mechanisms of disinformation, and the triggers of evidence resistance.

3. Evidence Resistance

We have increasingly sophisticated ways of acquiring and communicating knowledge, but efforts to spread this knowledge often encounter resistance to evidence. Resistance to evidence consists in a disposition to reject evidence coming from highly reliable sources. This disposition deprives us of knowledge and understanding and comes with dire practical consequences; recent high-stakes examples include climate change denial and vaccine scepticism.[10,11,16,18,22,33, 34, 41]

Until very recently, the predominant hypothesis in social epistemology and social psychology[13, 14, 24, 40] principally explained evidence resistance with reference to politically motivated reasoning: on this view, a thinker's prior political convictions (including politically directed desires and attitudes about political group-membership) best explain why they are inclined to reject expert consensus when they do. Typically, epistemologists who have explored the consequences of this empirical hypothesis take its merits at face value.[2,7]

However, on closer and recent inspection, the hypothesis is both empirically and epistemically problematic. Empirically, there are worries that in extant studies political group identity is often confounded with prior (often rationally justified) beliefs about the issue in question; and, crucially, reasoning can be affected by such beliefs in the absence of any political group motivation. This renders much existing evidence for the hypothesis ambiguous[43]. Epistemologically, the worry is that the hypothesis is ineffective in making crucial distinctions among a number of phenomena, such as: (1) Concerning epistemic status: between irrational resistance to evidence and rationally justified evidence rejection; (2) Concerning triggers: between instances of motivated reasoning on one hand, and epistemically deficient reasoning featuring cognitive biases and unjustified premise beliefs on the other; (3) Concerning strategies for addressing the phenomenon of evidence resistance: between targeting widely spread individual irrationality on one hand, and targeting an unhealthy epistemic environment on the other.[11,33,34,41]

Furthermore, difficulties in answering the question as to what triggers resistance to evidence have very significant negative impact on our prospects of identifying the best ways to address resistance to evidence. If resistance to evidence has one main source - for instance, a particular type of mistake in reasoning, such as motivated reasoning - the strategy to address this problem will be unidirectional and targeted mostly at the individual-level. In contrast, should we discover that a pluralistic picture is more plausible when it comes to what triggers resistance to evidence - whereby this phenomenon is, e.g., the result of a complex interaction of social, emotive, and cognitive phenomena - we would have to develop much more complex interventions, at both individual and societal levels.[33]

Our results show[11,33,34] that the widespread irrationality hypothesis assumed by the politically motivated reasoning account of evidence resistance is incorrect: humans are very reliable cognitive machines, in spite of relatively isolated instances of biased cognitive processing or heuristics-based reasoning. [26,27,28] Irrational resistance to evidence is rare, and is an instance of input-level epistemic malfunctioning, often encountered in biological traits the proper function of which is input-dependent: just like our respiratory systems are biologically malfunctioning when failing to take up easily available oxygen from the environment, our cognitive systems are epistemically malfunctioning when failing to take up easily available evidence from the environment.[33, 34]

Irrational evidence resistance, sourced in cognitive malfunction, is rare: what is often encountered in the population, however, is rationally justified evidence rejection, due to overwhelming (misleading) evidence present in the (epistemically polluted) environment of the agent.[3,17,19,21,33,34] When agents rationally reject reliable scientific testimony, they do so in virtue of two types of epistemic phenomena: rebutting epistemic defeat, and undercutting epistemic defeat.[3,17,19] Rebutting epistemic defeat is a technical term for evidence against a particular belief: for instance, testimony from sources one is rational to trust that contradicts scientific testimony on the issue. These sources will be rationally trusted by the agent because of an excellent track record of testimony: they are overall reliable testifiers in the cognitive agent's community (which is why it is rational for the agent to trust them), but who are mistaken about the matter at hand: reliability is not infallibility, it admits for failure.[3,17,18,19, 27,33,37,38]

The most often encountered trigger for rational evidence resistance is undercutting epistemic defeat. Undercutting epistemic defeat refers to evidence that suggests a particular testimonial source is not trustworthy: relevant examples include misleading evidence against the reliability of a particular source of scientific testimony, a particular media outlet, or against the trustworthiness of a particular public body, for instance, is an undercutting defeater. In vaccine sceptic communities, for instance, we often encounter worries that the scientific community, or the NHS, do not have the relevant communities' interests in mind when they recommend vaccine uptake. These worries, in turn, are, once more, often rationally sourced in otherwise reliable testimony (testifiers within the agent's community, that the agent trusts due to excellent track record, but that are wrong on this particular occasion).[3,17,19,21,33,37,38]

These results, in turn, illuminate the best strategies to address the phenomenon of evidence resistance. Two major types of interventions are required:

(1) For combatting rational evidence rejection:[33] engineering enhanced social epistemic environments. This requires: (1.1) combatting rebutting defeaters via evidence flooding: evidence resistant communities, inhabiting polluted epistemic environments, cannot be reached via the average communication strategies designed to reach the mainstream population, inhabiting a friendly epistemic environment (with little to no misleading evidence). What is required is (1.1) quantitatively enhanced reliable evidence flow: this is a purely quantitative measure, aimed to outweigh rebutting defeaters in the agent's environment. More evidence in favour of the scientifically well supported facts will, in rational agents, work to outweigh the misleading evidence they have against the facts; (1.2) qualitatively enhanced reliable evidence flow: this is a qualitative measure, that aims to outweigh misleading evidence via evidence from sources that the agent trusts – that are trustworthy vis-à-vis the agent's environment (see below on context-variant trustworthiness); (1.3) quantitatively and qualitatively enhanced evidence aimed at combatting undercutting defeat (misleading evidence against the trustworthiness of reliable sources): flooding evidence resistant communities with evidence from sources they trust in favour of the trustworthiness of sources they fail to trust due to misleading undercutting defeaters.

(2) For combatting (relatively isolated) cases of irrational evidence resistance due uptake cognitive malfunction: increasing availability of cognitive flexibility training[9,31,33] (e.g. in workplaces, schools, alongside anti-bias training). Cognitive flexibility training helps with enhancing open-mindedness to evidence that runs against one's held beliefs, and to alternative decision pathways.

4. Disinformation

Our results[35], show that disinformation need not come in the form of false content, but rather consists of content with a disposition to generate ignorance in normal conditions at the context at stake. This predicts that disinformation is much more ubiquitous and harder to track than it is currently taken to be in policy and practice: mere FactCheckers[15, 23, 30, 32, 41] just won't be able to adequately protect us against disinformation, because disinforming does not require making false claims. Disinformation is ignorance generating content: Content X is disinformation in a context C iff X is a content unit communicated at C that has a disposition to generate ignorance at C in normal conditions.[35] The same communicated content will act differently depending on contextual factors such as: the evidential backgrounds of the audience members, the shared presuppositions, extant social relations, and social norms.[25,35]

Generating ignorance can be done in a variety of ways – which means that disinformation will come in diverse incarnations:[35]

(1) Disinforming via spreading content that has the capacity of generating false belief: The paradigmatic case of this is the traditionally recognised species of disinformation: intentionally spread false assertions with the capacity to generate false beliefs in hearers. Our results show that this simple way of disinforming is both the least efficient and the least dangerous (because easiest to detect by average cognizers).

(2) Disinforming via misleading defeat[1,3,17,19,35]: This category of disinformation has the capacity of stripping the audience of held knowledge via defeating justification for belief. A classic example is communicating the content: 'There is disagreement about climate change in the scientific community'. This claim is strictly speaking true – there is some very minor disagreement in the community – but it generates the false belief (via implicating it) that there is substantive disagreement, which makes it rational to suspend belief on the issue.

(3) Disinforming via content that has the capacity of inducing epistemic anxiety*:* this category of disinformation has the capacity of stripping the audience of knowledge via belief defeat – i.e. via triggering belief loss.[3,17,19,37] The paradigmatic way to do this is via artificially raising the stakes at the context/introducing irrelevant alternatives as being relevant (e.g. 'Are you really sure the vaccine is safe? After all, scientists do sometimes make mistakes'). The way this variety of disinforming works is via falsely implicating that these error possibilities are relevant at the context, when in fact they are not – since the scientific community has an excellent track record, the possibility of error is rationally negligeable – and, indeed, rationally neglected whenever we take an aspirin or use toothpaste. In this, the audience's body of evidence is changed to include misleading justification defeaters.

(4) Confidence-defeating disinformation: introducing misleading (justification/doxastic) defeaters, which gets you to lower your confidence: you now take/are justified to take the probability of climate change happening to be much lower than it actually is.

(5) Disinforming via exploiting pragmatic phenomena:[1,21,35,37,40] Pragmatic phenomena can be easily exploited, and very often are, to the end of disinforming in all ways above: True assertions carrying false implicatures or false presuppositions will display this capacity to generate false beliefs in the audience. An important example in the public sphere is generated by ethical codes for journalism, that ubiquitously feature two often conflicting principles: the reliability principle and the impartiality principle. In cases of conflict, giving primacy to impartiality over reliability carries a high risk of generating ignorance in the audience and widely spread distrust in expertise: via giving two viewpoints on an issue equal exposure, the implicature triggered for the audience is that they are equally well supported by evidence. Several studies show that this generates widely spread false beliefs in the audience.[39] There are two ways to address this problem: either (1) the impartiality principle needs to be explicitly limited to expressing opinions on matters on which we have limited access to facts; or (2) the impartiality principle needs to be constrained by the

reliability principle: opposing voices need to be explicitly given the weight corresponding to their reliability.

What all of these ways of disinforming have in common is that they generate ignorance – either by generating false beliefs, or by generating knowledge loss. Importantly, this capacity to generate ignorance will heavily depend on the audience's background evidence/knowledge. A signal r carries disinformation for an audience A wrt *p* iff A's evidential probability that *p* conditional on r is less than A's unconditional evidential probability that *p*, and p is true.[25,35]

Some of the best disinformation detection tools at our disposal will fail to capture most types of disinformation. To give but a few examples: the PHEME-project[30] aims to algorithmically detect and categorize rumours in social network structures (such as Twitter and Facebook), and to do so, impressively, in near real time. The rumours are mapped according to four categories, which include "disinformation, where something untrue is spread with malicious intent."[41]. Similarly, Kumar and Geethakumari's project[23] develops an algorithm which ventures to detect and flag whether a tweet is misinformation or disinformation. In their framework "Misinformation is false or inaccurate information, especially that which is deliberately intended to deceive [and] [d]isinformation is false information that is intended to mislead, especially propaganda issued by a government organization to a rival power or the media."[23] In Karlova and Fisher's[15] diffusion model, disinformation is taken to be deceptive information. Shao et al's Hoaxy[32] is "a platform for the collection, detection, and analysis of online misinformation, defined as "false or inaccurate information"[32].

It becomes clear that these otherwise excellent tools are just the beginning of a much wider effort that is needed in order to capture disinformation in all of its facets, rather than mere paradigmatic instances thereof, which involve false assertions. At a minimum, we need to build Fact Checkers that track pragmatic deception mechanisms, as well as evidential probability lowering potentials against an assumed (common) evidential background of the audience.

5. Rational Trust and Trustworthiness

Rational trust is trust that is sourced in good (contextually-appropriate) evidence of trustworthiness.[4,5,6,8,12,38] In turn, trustworthiness amounts to a disposition to comply with one's contextually-determined obligations: maximal trustworthiness amounts to a maximally strong disposition to meet all of one's contextually determined obligations.[20,36] Degrees of trustworthiness of an entity (e.g. individuals, corporations, public institutions, Artificial Intelligence[36]) can be measured against a contextually-set threshold by measuring the disposition of norm-compliance of the relevant entities, across two dimensions: breath (number of contextually-determined obligations that the entity has a disposition to comply with) and depth (strength of disposition for norm-compliance).[20]

This explains why often agents and communities rationally distrust reliable sources: at the relevant context, the obligations shouldered by the relevant agents (or organisations, policy makers, science communicators, AIs) amount to more than merely being reliable informants on the issues at hand. A paradigmatic case will be one in which a public body fails to offer enough evidence of self-reliability for the context: (see above) in communities in which may defeaters exist about the reliability of the relevant expert testimony, the testifiers shoulders the obligation to defeat those defeaters (via offering quantitively and qualitatively superior evidence of source reliability, and in support of the communicated content, well adapted/as required for the epistemic environment present at the context).[3,8,17,19,20,31,38] This result suggests, for instance, that mere testimony form e.g. NHS that vaccines are safe will not be sufficient for rational trust in communities that inhabit an environment heavily polluted by misleading evidence (e.g. from reliable and rationally trusted informants that are wrong on this particular issue) against the reliability of NHS testimony, or against the safety of vaccines. In these environments, in order for the NHS to qualify as a trustworthy testifier (and thereby to be likely to be rationally trusted), the NHS shoulders the obligation of defeating misleading evidence present in the communities that it aims to reach (see above for strategies).

References

1. Alfano, M., Cheong, M., and Carter, J.A. (2018). Technological Seduction and Self Radicalization. *Journal of the American Philosophical Association* 4(3): 298-322
2. Ancell, A. (2019). The fact of unreasonable pluralism. *Journal of the American Philosophical Association* 5(4): 410-428.
3. Brown, J. and Simion, M. (2021). *Reasons, Justification, and Defeat* (with J. Brown, eds.). Oxford University Press.
4. Carter, J.A. (Forthcominga). *A Telic Theory of Trust*. Oxford: Oxford University Press.
5. Carter, J.A. (Forthcomingb). Trust and Trustworthiness. *Philosophy and Phenomenological Research.*
6. Carter, J.A. (Forthcomingc). Trust as Performance. *Philosophical Issues.*
7. Carter, J.A. and McKenna, R. (2020) Skepticism Motivated: On the Skeptical Import of Motivated Reasoning. *Canadian Journal of Philosophy*, 50 (6). 702 - 718.
8. Carter, J.A. and Simion, M. (2020). The Ethics and Epistemology of Trust. (with A. Carter). *Internet Encyclopaedia of Philosophy* (J. Matheson ed.).

9. Chaby LE, Karavidha K, Lisieski MJ, Perrine SA, Liberzon I. (2019). Cognitive Flexibility Training Improves Extinction Retention Memory and Enhances Cortical Dopamine With and Without Traumatic Stress Exposure. Front Behav Neurosci. 2019 Mar 1;13:24. doi: 10.3389/fnbeh.2019.00024. PMID: 30881293; PMCID: PMC6406056.
10. Chrisman, Matthew (2008). Ought to Believe. *Journal of Philosophy*, 105/7: 346-370
11. Gluer, K. and Wikforss, A. (2022). What is Knowledge Resistance?. In Stromback et al. 2022.
12. Gordon, E.C. (2022). When Monitoring Facilitates Trust. *Ethical Theory and Moral Practice*. Online First.
13. Kahan, D. (2013). Ideology, Motivated Reasoning, and Cognitive Reflection. Judgement and Decision Making 8: 407-424.
14. Kahan, D. (2016). The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It. In *Emerging Trends in the Social and Behavioral Sciences*, 1–16.
15. Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disin- formation for understanding human information behavior. *Information Research, 18*(1).
16. Klintman, M. (2019). *Knowledge Resistance: How We Avoid Insight from Others*. Manchester University Press.
17. Kelp, C. (Forthcoming). *The Nature and Normativity of Defeat*. Cambridge: Cambridge University Press.
18. Kelp, C. (2021). *Inquiry, Knowledge, and Understanding*. Oxford: Oxford University Press.
19. Kelp, C. (2020). 'Internalism, Phenomenal Conservatism, and Defeat.' Philosophical Issues 30, 192-204.
20. Kelp, C. and Simion, M. (Forthcoming). What is Trustworthiness?. *Nous*.
21. Kelp, C. and Simion, M. (2021). *Sharing Knowledge: A Functionalist Account of Assertion*. Cambridge University Press.
22. Kornblith, H. (2001). Epistemic Obligation and the Possibility of Internalism, in A. Fairweather and L. Zagzebski, eds., Virtue Epistemology: Essays on Epistemic Virtue and Responsibility (New York: Oxford), pp. 231-48.
23. Kumar, K.P.K. and Geethakumari, G. (2014), Detecting misinformation in online social networks using cognitive psychology, Human-centric Computing and Information Sciences, Vol. 4 No. 1, 22pp.x
24. Lord, Charles G., Lee Ross, and Mark R. Lepper. (1979). Biased Assimiliation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. *Journal of Personality and Social Psychology* 37 (11), 2098–2109.

25. Lyons, J. C. (2022) Cognitive diversity and the contingency of evidence. *Synthese*, 200(3), 202.
26. Lyons, J. C. (2009) *Perception and Basic Beliefs*. Oxford: Oxford University Press.
27. Lyons, J. C. (2016) Inferentialism and cognitive penetration of perception. *Episteme*, 13(1), pp. 1-28.
28. Lyons, J. (2011) Circularity, reliability, and the cognitive penetrability of perception. *Philosophical Issues*, 21(1), pp. 289-311.
29. Molden, D. C., & Higgins, E. T. (2012). Motivated thinking. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 390–409). Oxford University Press.
30. PHEME (2014), "About Pheme", available at: www.pheme.eu (accessed September 14, 2022).
31. Sassenberg, K., Winter, K., Becker, D., Ditrich, L., Scholl, A., and Moskowitz, G. (2022). Flexibility Mindsets: Reducing Biases that Result From Spontaneous Processing. European Review of Social Psychology, Vol. 33/1: 171-213.
32. Shao, C.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2016. Hoaxy: A platform for tracking online misinformation. In Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion, 745–750. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.
33. Simion, M. (Forthcominga). *Resistance to Evidence*. Cambridge: Cambridge University Press.
34. Simion, M. (Forthcomingb). Resistance to Evidence and the Duty to Believe. Philosophy and Phenomenological Research.
35. Simion, M. (Forthcomingc). Knowledge and Disinformation. *Episteme*.
36. Simion, M. and Kelp, C. (Forthcoming). Trustworthy Artificial Intelligence. *Asian Journal of Philosophy*, Special Issue, ed. N. Pedersen.
37. Simion, M. (2021). *Shifty Speech and Independent Thought: Epistemic Normativity in Context*. Oxford University Press.
38. Simion, M. (2021). Testimonial Contractarianism: A Knowledge-First Social Epistemology. *Nous*. 55/4: 891-916.
39. Simion, M. (2017). Epistemic Norms and He Said/She Said Reporting. *Episteme,* 14/4: 413-422.
40. Søe, S.O. (2016), The urge to detect, the need to clarify. Gricean perspectives on information, misinformation, and disinformation, PhD thesis, Faculty of Humanities, University of Copenhagen.
41. Stromback, J., Wikforss, A., Gluer, K, Lindholm, Oscarsson, H. (eds) (2022). *Knowledge Resistance in High-Choice Information Environments*. New York: Routledge.

42. Taber, C. S., and Lodge, M.. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science* 50 (3): 755–69.
43. Tappin, B. M., Pennycook, G., & Rand, D. G. (2021). Rethinking the link between cognitive sophistication and politically motivated reasoning. *Journal of Experimental Psychology: General, 150*(6), 1095–1114.