Social meta-inference and the evidentiary value of consensus

Keith J. Ransom (keith.ransom@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne

Andrew Perfors (andrew.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne

Rachel Stephens (rachel.stephens@adelaide.edu.au)
School of Psychology, University of Adelaide

Abstract

Reasoning beyond available data is a ubiquitous feature of human cognition. But while the availability of first-hand data typically diminishes as the concepts we reason about become more complex, our ability to draw inferences seems not to. We may offset the sparsity of direct evidence by observing the statements of others, but such social meta-inference comes with challenges of its own. The strength of socially-provided evidence depends on multiple factors which themselves must be inferred, like the knowledge, social goals, and independence of the people providing the data. Here, we present the results of an experiment aimed at examining how people draw conclusions from information provided by others in the context of social media posts. By systematically varying the degree of consensus along with the number of people and distinct arguments involved we are able to assess how much each factor affects the conclusions reasoners draw. Across a range of topics we find that while people are influenced by the number of people on each side of an argument, the number of posts is the dominant factor driving belief revision. In contrast to well established findings in simpler domains, we find that people are largely insensitive to the diversity of the arguments made.
Keywords: reasoning; social meta-inference; consensus; induction; diversity; explanation.

Introduction

Learning is hard. Humans are constantly faced with situations where we have to make complicated inferences based on very little data, like acquiring new concepts given only a few examples or extrapolating intelligently about patterns based on a handful of instances. This kind of induction based on sparse data is something humans excel at: for instance, given a few examples of animals with some property *P*, people draw reliable and sensible conclusions about what other animals also have that property. In these relatively simple situations, there is robust evidence that people's generalisations are highly sensitive to the structure of the underlying conceptual space as well as the diversity and number of the examples provided (e.g., Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Sloman, 1993).

These regularities are explained, at least in part, by how people assume the data were generated. If they believe the examples were strongly or helpfully sampled, people generalise less from additional similar examples (Hayes, Navarro, Stephens, Ransom, & Dilevski, 2019) or repeated instances (Perfors, Ransom, & Navarro, 2014; Xie, Hayes, & Navarro, 2020) but generalise more when they see more diverse ones (Voorspoels, Navarro, Perfors, Ransom, & Storms, 2015; Ransom, Perfors, & Navarro, 2016). This behaviour is consistent with normative statistical inference, which is sensitive to assumptions about sampling and independence.

Unfortunately, most of the problems people are faced with in the real world are far more complex than these simple situations. Consider somebody who hears the claim that "perfect avocados are getting harder to find." The problem of determining whether that claim is true somewhat resembles a more complex category-induction problem in which the examples are analogous to arguments or premises supporting the claim and the property being generalised is gettingHarderToFind.[1] However, there are many differences that make the situation sufficiently more complicated that it is still unclear not only empirically *what* people do in this kind of situation, but also what a normative standard would say they *should* do.

Multiple factors make this problem more complex. First, unlike the simple domains that most categorisation and category induction experiments operate over, we cannot fully or accurately characterise the shape of the space (either as scientists or as reasoners). What *is* the shape of "argument space" around the above claim about perfect avocados? Intuitively, some arguments offer better support for the truth of that claim than others and hence are closer in that space: if true, an argument that "avocado farms have been destroyed due to climate change" is better support for the claim than one like "the price of oranges is skyrocketing." But how much better is it, and why? In order for people to reason about how to generalise from arguments, they must be able to figure out how to calculate these distances on the fly for any arbitrary argument. Very little is known about how we do this.

Second, the generative process for the data (the arguments) goes far beyond strong or helpful sampling. Individuals making arguments in support of a claim (e.g., on social media) are embedded in a rich social system and have complicated and often unknown goals. They may or may not be helpful or knowledgeable; they might be sharing information in order to be deceptive, to troll, or to signal their identity. All of these possibilities have different implications for how a reasoner should reason about the arguments they offer.

Finally, when multiple arguments are present, it is often unclear how independent they are. Normative statistical models

[1] Of course, reformulating category induction problems in this more general way brings in additional issues surrounding explanation and meta-inference (Sloman, 1994). We return to this in the discussion, although space precludes a detailed treatment of this idea.

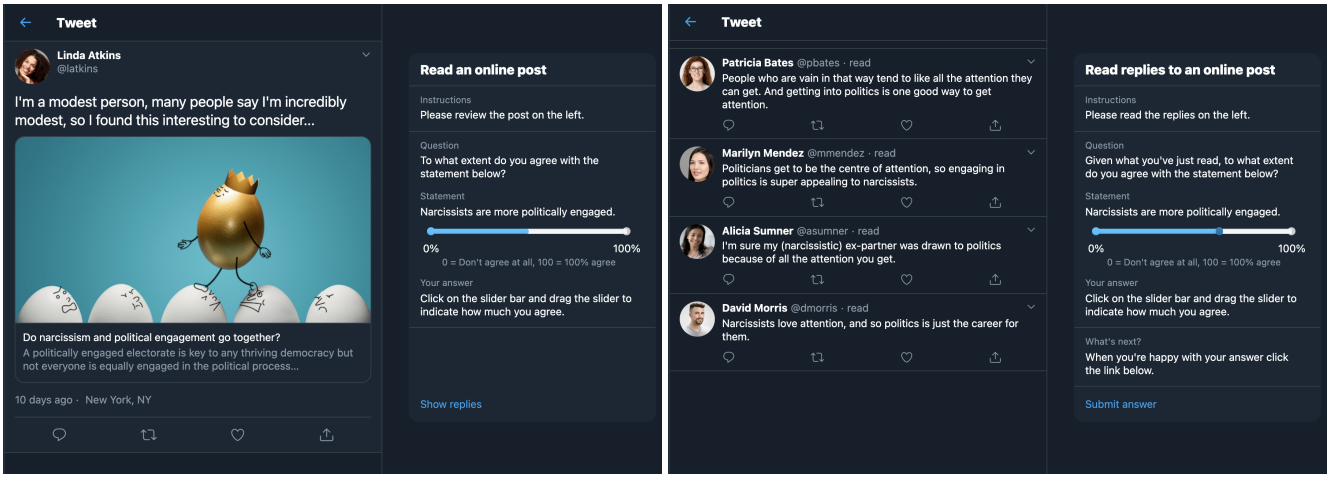


Figure 1: **Example trial.** [Left panel] On each trial participants view a brief neutral post on a given topic (here, about whether narcissists are more political). They provide their prior beliefs by rating their agreement with that statement. [Right panel] Participants then view a number of tweets replying to the original post, after which they update their agreement rating. Conditions vary in the number of distinct people and arguments provided in support of or against the claim. In this example, four distinct people are all making essentially the same argument in support, with very slight variations in wording (i.e., that narcissists like the attention they get in politics).

that assume independence almost certainly do not apply in social situations. If the same person offers multiple arguments, was evidence for those arguments acquired and verified independently and then used to support the conclusion, or derived *post hoc* from the conclusion? If multiple people offer the same argument, did they independently conclude that it was the most compelling, or did all of them view the same information? Standard treatments of evidence aggregation recognise the difficulty of integrating information from multiple sources but are not designed for this level of social complexity (e.g., Budescu & Yu, 2007).

Although our ultimate goal is to build on existing models of category-based induction, social sampling, and evidence aggregation to determine a normative standard for how people *should* reason in this sort of situation, in this work we seek to obtain empirical evidence about how people *do* reason. There is substantial research on important aspects of this topic, of course. It suggests that people sometimes (Whalen, Griffiths, & Buchsbaum, 2018) but not always (Yousif, Aboody, & Keil, 2019) reason appropriately when given sources that are not independent. People are also sensitive to factors such as the perceived confidence (Sah, Moore, & MacCoun, 2013) or prestige of the source (Atkisson, O'Brien, & Mesoudi, 2012), the number of people supporting a claim (Efferson, Lalive, Richerson, McElreath, & Lubell, 2008; Lewandowsky, Cook, Fay, & Gignac, 2019), or the complexity of explanatory arguments (Zemla, Sloman, Bechlivanidis, & Lagnado, 2017).

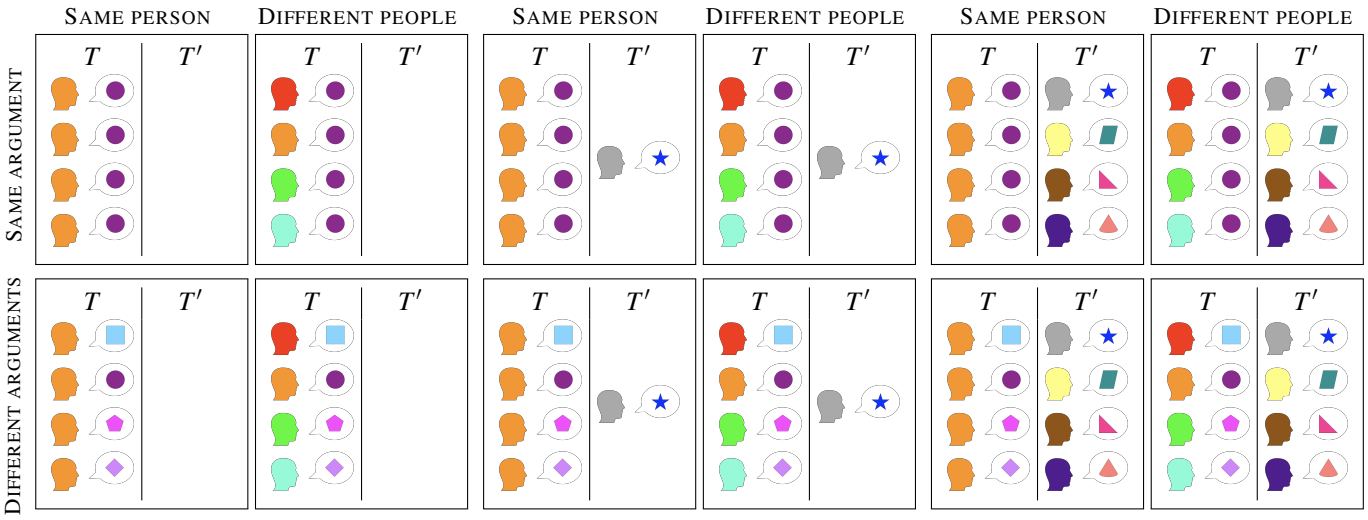
However, to our knowledge, there is nothing that systematically varies several of these factors at once, especially with a wide variety of stimuli in a realistic social context. This paper offers preliminary work in that direction. To be precise, the question we consider in the present study is how people integrate evidence when reasoning about propositions in which the basis for induction is unclear, the conceptual space is complex and uncertain, and people vary substantially in their prior beliefs and access to information. When what little data the reasoner has to go on is social in origin and may be unreliable, which cues provide better support for a given claim? Does the number of people providing arguments matter more than the content of what they say? Does repetition (of individuals or of statements) improve or decrease support?

Method

We investigate these questions in an experiment where participants viewed arguments for and against a variety of claims, presented as tweets via a mock twitter interface (see Figure 1). After reading a brief post introducing a claim people rated their support for it, both before and after seeing arguments in favour or against it. By manipulating the diversity of the arguments and the people making them, as well as the raw number of tweets on either side of a claim, we investigated how the strength of people's inferences are impacted by each of these factors.

Design and procedure

Our experiment employed a $3 \times 2 \times 2$ factorial design, illustrated in Figure 2. Two factors (source diversity and argument diversity) were varied within subjects, while a third (consensus level) was manipulated between subjects. People were allocated to one of three consensus level groups, and thus participated in 4 out of 12 experimental conditions. The experimental procedure was consistent across all groups and conditions. Each trial began with the presentation of a social media post depicting a proposition. After reading it, people were asked to rate their agreement with the claim on a sliding scale from 0 ("Don't agree at all") to 100 ("100% agree"). Following this initial rating, a number of reply tweets were presented and people were given the opportunity to update their rating. To ensure that all material was read, people had to click on each tweet before the rating could be updated.



(a) FULL (b) MAJORITY (c) CONTESTED

Figure 2: **Experiment design**. Our $3 \times 2 \times 2$ design varied the relative quantity of information in favour or against the target statement (between subjects), and the perceived quality of the consensus (within subjects). On each trial participants were shown a combination of target tweets (T) and opposing tweets (T'). All trials involved four target tweets (either all in favour or all against the target statement). The number of opposing tweets varied by condition, creating the appearance of: (a) a FULL consensus; (b) a MAJORITY consensus; or (c) a CONTESTED consensus. Consensus quality was manipulated via two factors: source diversity – whether the target tweets were written by the same person or different people; and argument diversity – whether each of the target tweets represented the same argument or different arguments. All participants saw the full set of 20 scenarios (five trials in each of the four consensus quality conditions). The order of tweets for a given trial, and whether the target tweets were in favour or against the target statement, was randomised across trials.

Consensus level In order to examine the degree to which a numerical consensus among data points (tweets) drives belief revision independent of other factors, we systematically varied the proportion of tweets on either side of a claim in this between-subjects manipulation. For people in the FULL group, the four target tweets were always unopposed; for each claim, people randomly saw either four Pro tweets arguing in favour or four Con tweets arguing against, with no dissenters. For the MAJORITY group the numerical advantage was 4 : 1 in favour of the target side of the argument (which was randomly set to either Pro or Con for each person). Lastly, for people in the CONTESTED group, there was no such advantage – tweets were split 4 : 4 across all trials.

Source diversity A potentially important factor in how people might evaluate the quality of an apparent numerical consensus is the number of *unique* people making the argument. In this within-subjects manipulation, we therefore varied whether the four target tweets appeared to be written by the same person or different people. Each distinct person was randomly assigned to a distinct user icon and name, which were prominently displayed alongside each tweet to enhance the salience of this source information (see Figure 1).

Argument diversity A second within-subjects factor varied whether the tweets communicated the same or different arguments. In the DIFFERENT ARGUMENTS condition each of the tweets advanced a different reason, while in the SAME ARGUMENT condition the tweets were differently worded variants of the same underlying idea (sharing key words and phrases to enhance similarity).

Stimuli

The full set of stimulus material for each trial consisted of a social media post containing a proposition statement on a variety of topics (see Figure 3), as well as a set of associated "reply" tweets. There were 20 different posts/topics in all, and all participants saw all posts in random order. Posts were designed to elicit a range of prior beliefs (cf. Figure 3), and varied in degree of subjectivity as well as the extent of prior knowledge that people were likely to have about the topic.

While each initial post was neutral with respect to the associated proposition, the reply tweets were either for or against it. For each topic there were seven tweets of each kind (Pro or Con), made up of four non-diverse tweets making the same argument with slightly different words, and three additional diverse tweets (making different arguments). On each trial, between four and eight tweets were selected according to the condition design (Figure 2), and presented in random order. Each participant saw five trials from each of the four conditions to which they had been assigned, with a randomised mapping between topic and condition. Whether the four target tweets were Pro or Con was randomised across trials.

Participants

A total of 345 participants were recruited via Amazon Mechanical Turk, with one excluded because of a failure to save all of the data. The remaining 344 participants ranged in age between 20 and 71 (median: 39.5 years), comprised 42% females, and were drawn predominately from the U.S. and Canada (86%). 78% of people identified as native English speakers but all had been screened for English language competency prior to recruitment. Other demographic variables

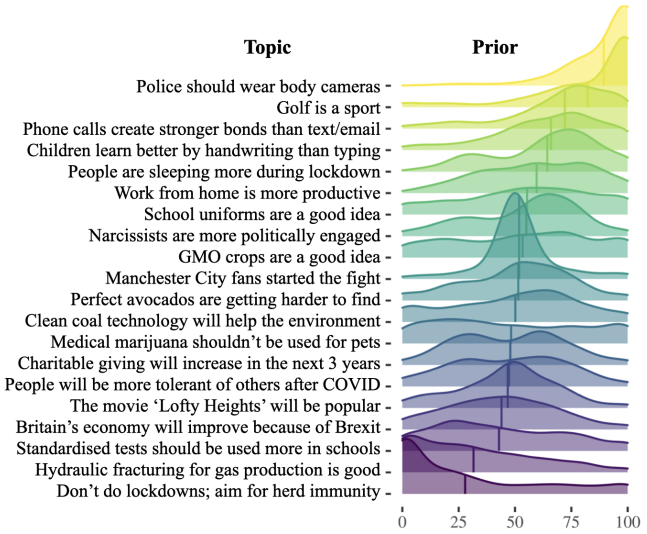


Figure 3: **Prior ratings by topic.** Choices on a rating agreement task for each of the 20 topics used in the study, aggregated across participants and conditions. Distributions show participant agreement with each topic on a scale from 0–100 (where 0 is "not at all") before having seen any arguments for or against, based only on a neutrally worded tweet placing the proposition in context. Prior beliefs varied widely across topics and people, allowing us to assess the impact of our experimental factors across a range of circumstances.

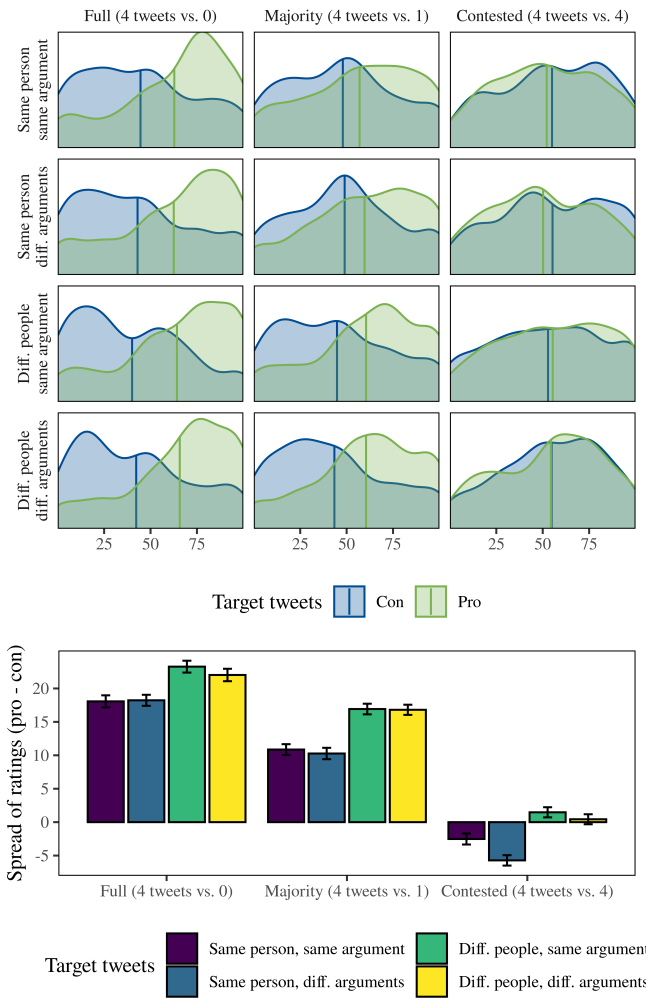


Figure 4: **Main results.** Ratings of agreement with post propositions as a function of the information provided in follow-up tweets across 12 conditions. [Upper panel] Posterior density plots of agreement ratings according to whether the target tweets were in favour of (Pro) or against (Con) the proposition, collapsed across all topics and participants. Vertical lines show distribution means. [Lower panel] The mean degree of separation between Pro and Con distributions after controlling for people's prior beliefs. Overall, posterior beliefs were affected most strongly by the numerical advantage of Pro versus Con tweets, with people shifting the most in the FULL condition and the least in the CONTESTED condition. Target tweets also had a greater impact when they appeared to come from different people, rather than the same person. Whether or not each of target tweets made the same point had little impact on belief revision.

which do not affect the analyses are not reported here. Due to differences in experiment duration across the different consensus level conditions (which varied the number of tweets that people were asked to read), participants were recruited and compensated separately: 117 people in the CONTESTED condition were paid $5.00USD for 25–30 minutes participation, while 117 and 110 people in the FULL and MAJORITY groups respectively were paid $4.00USD for 20–25 minutes.

Results

Our work is focused on determining what cues people rely on when presented arguments in a social situation. Are people sensitive to the quantity of information on either side, irrespective of the number of unique sources or the number of unique arguments? Or do these cues to consensus quality mediate people's reasoning, indicating that people are making more nuanced assumptions about how opinions are generated and what it means to express them?

To address these questions, we first examined people's levels of support for the propositions presented. People's prior beliefs for each topic, collected before any of the reply tweets were viewed, are shown in Figure 3. As intended, the different topics elicit different degrees of prior support. Topics vary in overall support (most people agree that golf is a sport and that standardised tests should not be used more widely, but have no strong opinions about clean coal technology) as well as the distribution of beliefs (e.g., views on charitable giving are somewhat bi-modal).

To analyse people's posterior beliefs after having seen a mixture of target and opposing tweets, we aggregated responses across topics separately according to whether the target tweets were in favour (Pro) or against (Con) the proposition. The degree of separation between the distributions, shown in the upper panel of Figure 4, reflects how much opinions differed based on whether people received primarily Pro or primarily Con arguments. The difference in means ($\mu_{\text{Pro}} - \mu_{\text{Con}}$) for each condition after controlling for prior ratings is shown in the lower panel of Figure 4. These results suggest that the main factor affecting belief revision is how many more target tweets there are than opposing tweets. When the number of target tweets is evenly matched by op-

Model	Consensus indicators	LOOIC	SE
1. PRIOR	—	-14965	337
2. PRIOR +	TWEETS	-15799	332
3. PRIOR +	TWEETS + PEOPLE	-15860	331
4. PRIOR +	TWEETS + PEOPLE + ARGUMENTS	**-15898**	332

Table 1: Comparison of how well four different regression models capture people's support for a proposition after viewing a number of tweets for or against (lower LOOIC indicates better fit). The best model contained predictors for the number of tweets, unique authors, and unique arguments, suggesting that all three factors affect people's posterior beliefs.

posing tweets (as in the CONTESTED conditions) there is little separation between the Pro and Con distributions, despite the fact that the opposing tweets always contain four distinct arguments from four distinct people while the target tweets varied in their diversity.

Although this numerical advantage is the largest factor affecting people's beliefs, it is not the only one. People are also sensitive to source diversity: when the target tweets came from different people, reasoners drew stronger inferences than when a single person created all of them. In contrast, there is little to no effect of argument diversity: the same person making essentially the same point four times was as effective (if not slightly more so) than when the same person offered four different arguments in support of their position.

To quantitatively assess the strength of evidence for these findings we compared four nested generalised linear models whose outcome variable is the posterior rating of agreement with the proposition made by each person for each topic. For the basis of comparison, our first model is a baseline which assumes that people's posterior support for a given proposition is affected only by their prior beliefs, not by any of the tweets they saw. Our second model adds a predictor that represents the numerical advantage (or disadvantage) of Pro tweets over Con tweets; it captures the idea that people do update their prior beliefs in light of such numerical imbalance, but in a way that is insensitive to potential cues about the quality of consensus that the imbalance reflects. The third model assumes that people are additionally sensitive to whether arguments are provided by many people rather than a single individual. Lastly, the fourth model reflects the possibility that people are also sensitive to the number of unique arguments provided.[2]

Table 1 shows leave-one-out cross-validation information criteria (LOOIC) for each of the models considered. A comparison of LOOIC reveals that the third model which captures

[2]Models were fit using the `brms` package (version 2.14.4) in R (version 4.0.3). Posterior ratings were scaled onto the range $(0, 1)$ and modelled as draws from a beta distribution via a logistic-link function. All predictors were scaled and centered on the range $[-.5, .5]$. To capture variability in the strength of (or reliance on) prior beliefs across the different topics as well as between native and non-native English speakers, all models included a random slope for the prior that varied accordingly. To capture individual response variability, the error term was modelled explicitly using a random intercept term for each participant. For space reasons we omit a discussion of the random effects here.

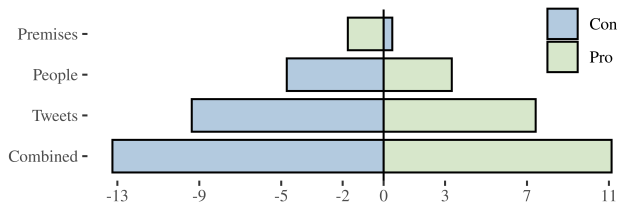


Figure 5: **Simulated marginal effects.** Simulated shift in agreement ratings (on a scale from -100 to 100) for a novel proposition based on fits obtained from the highest-performing (fourth) model. The first three bars (top to bottom) represent the marginal effect of seeing a full consensus (either Pro or Con) in terms of distinct arguments, distinct people, or number of tweets respectively. For comparison, the lower bar represents the combined effect of a full consensus on the basis of distinct people and number of tweets (but assuming an equal number of distinct arguments on both sides). The number of tweets on either side of an argument has the greatest impact on belief revision, shifting agreement nearly twice as much in both Pro and Con directions as the number of distinct people. Sharing distinct arguments rather than repeating them has a small negative impact.

sensitivity to source diversity, significantly out-performs both the preceding models. However, the best-performing model (with the lowest LOOIC) was the fourth one, suggesting that argument diversity also had a small effect on people's beliefs. To visualise the relative strength of the three factors we simulated draws from the posterior predictive distribution obtained from our fourth model and used them to make predictions for a novel topic given a nominal prior (set to the overall prior mean). The simulation results, shown in Figure 5, reveal that, holding all else constant, the numerical advantage of target tweets has the greatest impact on belief revision (approximately double the magnitude of the next-largest factor, which is the number of unique people authoring the tweets). Interestingly, the effect of argument diversity is in the opposite direction than might be expected: seeing the same argument four times led to a slightly greater degree of belief revision than seeing four different arguments once each.

Discussion

There is a well established literature that documents the many ways in which people appear to depart from normative statistical principles when reasoning probabilistically. For instance, people appear to be blind to matters of sample construction (Fiedler, 2012), exhibiting (among other things) an insensitivity to the interdependence of information sources (Yousif et al., 2019). Indeed, models of stimulus generalisation (Shepard, 1987), property induction (Osherson et al., 1990; Sloman, 1993), and category learning (Nosofsky, 1986) have enjoyed considerable success despite containing no explicit mechanism for capturing different sampling assumptions. However, even in these relatively simple situations, a sensitivity to sampling can explain otherwise puzzling aspects of generalisation (e.g., Ransom et al., 2016; Hendrickson, Perfors, Navarro, & Ransom, 2019). Moreover, this sensitivity may become even more important as the concepts we reason about become more complex: the psychological "space" that we reason within is more complicated and un-

certain, and more of the data is socially generated rather than directly observed (consisting of arguments, explanations, or ideas that were themselves the product of inference).

In this paper we took a first step toward methodically mapping out how people reason in this situation. Using a socially realistic context, we systematically varied three factors in which sampling assumptions are relevant: the total number of arguments, the number of unique individuals offering that data, and the number of unique arguments made. The first offers a measure of data quantity, while the other two potentially reflect aspects of data quality. We found that data quantity was the largest factor affecting generalisation, followed by the number of unique individuals.

Are these results consistent with normative standards of statistical reasoning? While it is certainly appropriate to be alert to sample size in the way our participants were, it is less clear whether people "should" weight the same information more strongly if it comes from more people. Some have argued that if those people arrived at their conclusions non-independently, e.g., based on the same source, then multiple people should be treated the same as one distinct person (Yousif et al., 2019). However, in our experiments – as is common in real life – it was not obvious whether the distinct individuals were operating from the same information or not; and even if they were, there still may be some benefit in knowing that multiple people drew the same conclusions from that information. As such, it is perhaps sensible that people drew stronger conclusions when a more diverse set of people were making the arguments.

Less straightforward is the finding that people were either unaffected by the diversity of arguments, or (if anything) drew slightly *weaker* conclusions when offered multiple distinct arguments rather than the same one repeated several times. This contradicts robust evidence in the category induction literature finding that premise diversity yields stronger conclusions (e.g., Osherson et al., 1990). One possibility is that people paid very little attention to the content of the tweets, focusing instead on more readily available cues like the number of tweets and the tweets' authors. However, while we did not explicitly test comprehension, there is reasonable evidence to suggest that people were reading and thinking about the content of the tweets. Firstly, the only indication that a tweet argued for or against a given claim was the content of the tweet itself. The qualitative pattern of our results (see Figure 4), whereby people's belief in a claim was strengthened by arguments in favour but weakened by arguments against, could only be obtained if people were indeed reading and comprehending the tweets. Secondly, an analysis of trial durations indicated reading rates within the normal adult range of 175–300 words per minute for silent reading of English non-fiction (Brysbaert, 2019).[3]

Another possibility is that the lack of an argument diversity effect arose because the need to provide four distinct arguments meant including weaker ones.[4] If so, then the value of additional arguments may have been diminished or reversed due to a form of "weak evidence effect". For example, Fernbach, Darlow, and Sloman (2011) found that people judge arguments with weak reasons to be poorer than arguments with no reasons. This happens, they argue, because people focus on the weak reasons mentioned and fail to think of alternatives.

Along similar lines, studies of inductive reasoning suggest that people interpret the absence of certain data differently depending on what they assume about the data generation process. For example, under a strong sampling assumption (but not a weak one) the weight of positive evidence may be diminished (Hayes et al., 2019) or reversed altogether (Ransom et al., 2016). A weak evidence effect might thus have arisen in our experiment due to people's assumptions about the way that arguments are selected and what the absence of stronger arguments signifies. Such absence may have led reasoners to assume that there were no stronger arguments available (why else would the weaker argument have been included?), or that there was no single strong reason to believe the claim. Conversely, seeing the same or similar argument multiple times may have been taken as an indication that the argument represents a strong reason to believe the claim and/or is based on reliable evidence. In future work, we plan to explore the environmental conditions under which different assumptions may be justified, and to investigate people's sensitivity to such underlying conditions.

Repeating an argument may also increase its evidentiary weight by improving recall when evidence is weighed at decision time, or by enhancing processing fluency (see Reber & Unkelbach, 2010, for a review of some effects of processing fluency and why it may constitute a reliable cue to truth). Exploring these issues further is the subject of ongoing work. In particular, in an attempt to replicate the effect of repetition to determine its robustness, we plan to run a study where the total number of tweets presented is fixed across different consensus levels to allow a cleaner comparison than was possible in the current study.[5]

Our work makes clear that there are many sources of variation in this area. Individuals varied considerably in their prior beliefs as well as how willing they were to change their mind in light of new data. Topics varied strongly in how much consensus there was beforehand as well as how easily that consensus could be changed. Topics also varied along a number of important dimensions (albeit not systematically in this preliminary work): matters of fact versus matters of opinion, fa-

[3] Based on the mean tweet length of 21.8 words, and an estimated 15 seconds decision time per trial. A re-analysis of our data after excluding all participants who appeared to have read too quickly by this measure yielded the same qualitative pattern of results and the same conclusions as our original analyses.

[4] In future work, we plan to collect argument diversity ratings for each topic, as well as argument strength ratings for single arguments.

[5] Such a design is not without its challenges. The number of tweets in the FULL condition is somewhat constrained by the extent to which distinct but meaningful arguments can be generated for each topic. But keeping this number manageable (i.e. low) potentially reduces sensitivity between consensus level conditions.

miliarity of topic matter, whether evidence given represented first-hand experience or second-hand knowledge, and so on. We have not had the space to delve into these details (nor the statistical power at the topic-level to do so), but a full understanding of reasoning in these situations requires a comprehension of how and why these topic-based and person-based differences matter. However, by measuring both prior and posterior beliefs, over a wide variety of topics which differ in meaningful ways, we were able to make more robust generalisations about how reasoning in these situations works in general, without overfitting to any one topic or message type. We believe that this sort of methodological generality is an important characteristic of scientific work in this area, given the level of variation that exists. That being said, in future research we plan to pursue the interaction between certain topic dimensions (such as the epistemic beliefs/values distinction) and the more general factors that we have explored.

Given the complexity of the issues involved in how people weigh the evidence of consensus, much work remains in order to understand the issue in its full generality. Nonetheless, our work represents a first step towards a systematic exploration of the factors that affect reasoning based on data from other people, as well as what this reveals about the assumptions people are making about how that information was generated. We see this as an important real-world extension of a rich tradition studying generalisation and explanation in simpler contexts.

Acknowledgments

This work was funded in part by the Department of Defence and the Office of National Intelligence under the AI for Decision Making Program, delivered in partnership with the Defence Innovation Partnership in South Australia.

References

Atkisson, C., O'Brien, M., & Mesoudi, A. (2012). Adult learners in a novel environment use prestige-biased social learning. *Evolut Psychol*, *10*(3), 519-537.

Brysbaert, M. (2019). How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, 104047.

Budescu, D., & Yu, H.-T. (2007). Aggregation of opinions based on correlated cues and advisors. *Jn. Beh. Decision Making*, *20*(2), 153–177.

Efferson, C., Lalive, R., Richerson, P., McElreath, R., & Lubell, M. (2008). Conformists and mavericks: The empirics of frequency-dependent cultural transmission. *Evol & Hum Beh*, *29*(1), 56-64.

Fernbach, P., Darlow, A., & Sloman, S. (2011). When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition*, *119*(3), 459–467.

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In *Psychology of learning and motivation* (Vol. 57, pp. 1–55). Elsevier.

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K. J., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, 1–8.

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. J. (2019). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Cog. Psych.*, *111*, 80–102.

Lewandowsky, S., Cook, J., Fay, N., & Gignac, G. (2019). Science by social media: Attitudes towards climate change are mediated by perceived social consensus. *Mem & Cog*, *47*(8), 1445-1456.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Jn. of Exp. Psych: Gen.*, *115*(1), 39–57.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psych. Review*, *97*(2), 185.

Perfors, A., Ransom, K. J., & Navarro, D. J. (2014). People ignore token frequency when deciding how widely to generalize. In *36th Conf. Cog. Sci. Soc.* (pp. 2759–2764).

Ransom, K. J., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40*(7), 1775–1796.

Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of philosophy and psychology*, *1*(4), 563–581.

Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Org. Behav. and Human Decision Processes*, *121*(2), 246–255.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Sloman, S. A. (1993). Feature-based induction. *Cog. Psych.*, *25*, 213–280.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition*, *52*(1), 1–21.

Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cog. Psych.*, *81*, 1–25.

Whalen, A., Griffiths, T. L., & Buchsbaum, D. (2018). Sensitivity to shared information in social learning. *Cognitive Science*, *42*(1), 168-187.

Xie, B., Hayes, B., & Navarro, D. J. (2020). Adding types, but not tokens, affects property induction. *Cognitive Science*, *44*.

Yousif, S., Aboody, R., & Keil, F. (2019). The illusion of consensus: A failure to distinguish between true and false consensus. *Psychological Science*, *30*(8), 1195-1204.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic bulletin & review*, *24*(5), 1488–1500.