Does it Harm Science to Suppress Dissenting Evidence?

Matthew Coates

Abstract

There has been increased attention on how scientific communities should respond to spurious dissent. One proposed method is to hide such dissent by preventing its publication. To investigate this, I computationally model the epistemic effects of hiding dissenting evidence on scientific communities. I find that it is typically epistemically harmful to hide dissent, even when there exists an agent purposefully producing biased dissent. However, hiding dissent also allows for quicker correct epistemic consensus among scientists. Quicker consensus may be important when policy decisions must be made quickly, such as during a pandemic, suggesting times when hiding dissent may be useful.

1 Introduction

Whilst dissent in scientific inquiry can be valuable, certain dissent may be problematic. One exemplar case was the tobacco industry's propagation of dissenting evidence against a causal link between smoking and cancer, to obscure the truth about scientific issues and introduce widespread ignorance and unwarranted doubt (Oreskes and Conway 2010). This is often called "epistemically detrimental" dissent. There is deep disagreement over how we should respond to epistemically detrimental dissent. One strategy is to hide at least some of the dissenting evidence (Oreskes, 2017; Cook, 2017; Nash, 2018). However, this is controversial, for example because it may lead to hiding "good" dissent (de Melo-Martín and Intemann, 2012; 2018).

To compliment these existing arguments, I investigate the effects of hiding dissent by modeling scientists preventing dissenting evidence from being published in journals. Whilst I only model this as a possible mechanism for how hiding dissent may occur, journals do sometimes hide dissenting evidence in this way (van Niekerk, 2003). Empirical evidence also shows that reviewers show a strong bias against manuscripts reporting results contrary to their own position (Mahoney, 1977).

I model this with a two-armed bandit problem, as used by Zollman (2009), which models different publication strategies of scientific journals. Zollman (2009) does not consider publication strategies which react to the beliefs of the community of researchers, something that my model does do.

I find evidence that it may be epistemically harmful to hide evidence supporting dissenting theories. When evidence supporting dissenting theories is hidden, the scientific community is typically less likely to come to a correct consensus. Furthermore, the epistemic harm caused by hiding dissent holds even when there exists an agent purposefully producing biased evidence supporting a dissenting theory. Nevertheless, I also find that it may be epistemically beneficial to hide dissent when scientists are already sufficiently far along the path to reaching consensus. In this case hiding dissent is as successful at coming to the correct consensus, whilst also doing so more quickly. This may, to some extent, justify decisions to hide dissent once the mainstream theory has significant evidential support.

In section two I motivate my model by discussing the debate about hiding dissent and provide examples where hiding dissent occurred. In section three I introduce my basic model, and in section four I describe the results. In section five I introduce certain modifications to account for scientists aiming to only hide epistemically detrimental dissent. Finally, in section six, I provide discussion about my results, as well as discuss limitations of the modeling decisions that I have used.

2 Background

Dissent plays an important role in the generation of scientific knowledge. There have been many cases where a scientific consensus has been false, and only through dissent could this false consensus be replaced by a new theory. Thus, it seems that scientists should increase opportunities for dissenting views to be heard and take them seriously when they do.

However, there are situations where dissent may be considered detrimental to scientific advancement. This is so called: "epistemically detrimental" (Biddle and Leuschner, 2015); "normatively inappropriate" (NID) (de Melo-Martín and Intemann, 2018); or "manufactured" (Oreskes and Conway, 2010) dissent. It may involve industry or think tanks purposefully funding research with the aim to generate studies calling into question widely received scientific views, as occurred in climate science and on the hazards of smoking. Alternatively, it may involve seemingly respectable scientists

defending claims that are entirely inconsistent with the widely accepted scientific view, such as on AIDS (de Melo-Martín and Intemann, 2014, 3). This dissent is problematic because it may impact scientific progress by preventing closure of scientific controversies when warranted, and by leading research and argumentation efforts astray in unfruitful directions (Miller, 2021). It may also affect policy-makers' and the public's views on science.

As a result, strategies to target certain dissent have been endorsed (Oreskes, 2017; Cook, 2017; Nash, 2018). De Melo-Martín and Intemann (2014, 596-599) have categorized three groups of strategies. The first is that dissent can be *masked*. This includes not reporting the full range of opinions on some matter, or by emphasizing some consensus position. The second is that dissent can be *silenced*, i.e. hidden such as during the peer-review process (De Melo-Martín and Intemann. 2014, 597). The third is that dissenters may be *discredited*.

I look at dissent being hidden through the peer review process. An example of this occurred when the *South African Medical Journal (SAMJ)* announced that they would no longer publish articles containing certain dissenting views on AIDS (van Niekerk, 2003), including denials of the claim that HIV was the cause of AIDS, because it served no useful purpose and may be harmful. This dissenting evidence had previously been used by the South African Government to justify policies like limiting the use of HIV antiretroviral drugs (de Melo-Martín and Intemann, 2018, 1-3). Whilst the journal did recognize that dissent is usually good, they argued that given the amount of evidence already demonstrated, printing and refuting dissenting arguments was taking resources away from solving the AIDS pandemic (van Niekerk, 2003). As a result they decided that hiding dissent would serve science better than publishing it.

In this case it is easy to recognize epistemically detrimental dissent. However, in many cases it may be more difficult. As a result, attempting to hide dissent may lead to the suppression of legitimate dissent that would have helped the advancement of science, hindering scientific practice. This is particularly important in sciences where consensus is difficult to challenge, including clinical practice in biomedical sciences (de Melo-Martín and Intemann, 2014, 609).

We can see that the blocking of potentially non-epistemically detrimental dissent does occur in empirical studies. For example, both Mahoney (1977) and Ernst and Resch (1994) find that reviewers may show a preference for evidence that supports their preferred theory, and they are more likely to reject manuscripts against it. This is likely to lead to greater, albeit less organized, hiding of dissenting evidence - as the reviewer is less likely to hold the minority view - and is not

limited to cases where dissenting evidence is known to be epistemically detrimental.

This is why de Melo-Martín and Intemann claim that a philosophically satisfying characterization of epistemically detrimental dissent, "must be able to successfully identify NID as such when the dissent in question is in fact normatively inappropriate and be able to exclude scientific dissent that is actually legitimate" (de Melo-Martín and Intemann, 2018, 8). Attempts at such characterizations have been put forth, for example by Biddle and Leuschner (2015), although de Melo-Martín and Intemann (2018) claim that these characterizations do not work.

Whilst other characterizations have since been suggested – for example Miller (2021) – in this paper I do not make a claim that a characterization of epistemically detrimental dissent exists, nor do I provide one. Even if one existed, empirical studies like Mahoney (1977) show that scientists do not necessarily follow one when hiding dissent, and instead are just less likely to publish evidence against their preferred theory. Therefore, I provide a computational model of hiding dissent to model the epistemic effect of hiding dissent on scientists without such a characterization.

3 The Base Model

My model is based on Zollman (2009), who models scientific inquiry as a network of scientists engaging in a two-armed bandit problem. In this model, each scientist is trying to decide which arm of a bandit (arms may represent actions, theories, medical procedures etc.) is better. The scientists have a choice between two arms, arm A and arm B. Arm A is the slightly worse option, with a success rate of p_A, whilst arm B has a success rate of $p_B = p_A + \delta$, however they do not know this success rate. At each time step the scientist repeatedly pulls one of the arms T times, producing their results (their evidence). Utility is the success they gain from pulling the arm.

Zollman (2009) focuses on the effects of different strategies that a journal may use to decide which results to publish. There exists a journal which has some strategy to choose which results to disseminate to the rest of the scientists. After each round of pulls, some results are chosen for publication. They are disseminated and each scientist updates their beliefs based on both their own results and the other results published by the journal. Because they also update on their own results, not every scientist updates on exactly the same information.

Each agent's beliefs about the probability of the success of each arm is represented by a beta distribution. A beta distribution is a continuous probability distribution whose shape is determined

by two parameters, α and β. An agent with a beta distribution thinks that all probability values, p, for each arm are possible, but that some values are more likely than others[1]. The initial beta distributions are created by assigning each agent an $\alpha, \beta \in [0, 4]$ for arm A and then for arm B randomly through a uniform distribution. α may be interpreted as tracking the number of successes of that theory, and β the number of failures of that theory. These α and β determine the shape of the initial beta distributions. Each agent has an expected value for each arm, given by $\mathbb{E}[X] = \alpha_X/(\alpha_X + \beta_X)$, and thinks the arm with the highest expected value is better.

The reason for initial $\alpha, \beta \in [0, 4]$, which are selected through a uniform distribution, is to follow Zollman (2009; 2010), where it is chosen so that initial beliefs do not swamp even a single experimental result. This also means that it reflects the case where there is likely no initial consensus among scientists.

The updating takes place as follows; if an agent currently has a beta distribution with parameters α, β, and they, or another scientist whose work has been published, perform T tests, where S is the number of successes, their new, updated, beta distribution has parameters $\alpha_{new} = \alpha + S$ and $\beta_{new} = \beta + T - S$. This is done for each set of results the agent receives.

How should agents choose which arm to pull? They may *exploit*, pulling the arm they currently think is better, or *explore*, pulling an arm they currently think is worse to find out more information about it. There is a tradeoff between getting a perceived better payoff (higher utility) immediately by exploiting, or exploring with the possibility of getting a better payoff later. There is also an incentive to let others explore and free-ride on their exploration whilst exploiting yourself. Even in relatively simple problems, finding the optimal strategy can become very complex (Berry and Fristedt, 1985).

One strategy is to be *myopic* by only exploiting. This is the standard assumption in models of this kind (Bala and Goyal, 1998, 596; Zollman, 2010, 27). There are reasons for assuming myopia. First, real agents may not possess the computational capacity to make the complex calculations needed to optimally choose between exploitation or exploration (Bala and Goyal, 1998, 596). Additionally, real scientific communities may sometimes be myopic. For example, myopia is optimal when an individual cares significantly more about their current payoff, rather than the future pay-

[1] **Definition: Beta Distribution:** A function $f(\cdot)$ on $[0, 1]$ is a beta distribution if and only if for some $\alpha, \beta > 0$,

$$f(x) = \frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha, \beta)}$$

where $B(\alpha, \beta) = \int_0^1 u^{(\alpha-1)}(1-u)^{(\beta-1)}\, du$

off. Given scientists are often rewarded for current success, through tenure, awards, and grants, this may cause myopia in large scale decisions like research methodology (Zollman, 2010, 27). There may also be cases where myopia is enforced in science. One example is pharmaceutical and biomedical research, where using a worse drug simply to explore whether it is actually worse may be considered unethical. This is pertinent because consensus is particularly difficult to challenge in clinical practice in biomedical sciences (de Melo Martin and Intemann, 2014, 609).

There are also modeling reasons for this assumption. Firstly, it simplifies the model to help better understand the findings. Additionally, adding exploration has a large effect on the results of the model, washing out many of the interesting findings produced with myopia. Because adding exploration has such a large effect, it is useful to look at myopia first, to see whether the findings are a result of exploration or some other component of the model. Finally, whilst it would limit their domain of inquiry, it may be that these bandit models are predominately applicable to cases where scientists are myopic.

However, myopia is a problematic assumption. Not only is myopia usually not the optimal strategy in bandit problems, it is typically not even a good strategy (Berry and Fristedt, 1985, 4). It may also be unrealistic to assume scientists are myopic. They are often aware of the epistemic benefits that come from exploring alternatives and are unlikely to completely ignore them. For example, in the early 1900's T. H. Morgan was opposed to Boveri–Sutton chromosome theory, but his lab tested it anyway, eventually producing results that caused him to abandon his opposition (Mayr, 1982, 769).

Therefore, I also consider *ε-greedy strategies* for two armed models, (Kummerfeld and Zollman 2016), which allow for a scientist to test perceived worse alternatives. In ε-greedy strategies, each agent exploits most of the time, however, each round each agent may instead explore with probability ε. Higher ε's correspond to higher exploration rates. For $\varepsilon > 0$, in the limit all agents will come to prefer the actually better arm and each will pull it with probability (1-ε) (Sutton and Barto, 2018, 28), leading to a consensus about the correct arm. Therefore, I assume that each agent has the same ε, to simplify the model, as assuming that agents have different ε does not significantly affect the results.

3.1 Journal Publication Strategies

Zollman (2009) considers five possible strategies for journals to choose which results are shared. These are *Random*, *Random Hero*, *Best*, *Best from Each* and *Best and Worst*.

Random selects a certain number of results without considering any feature of them and distributes them. *Random Hero* selects a certain number of scientists at the start of each run, again considering no feature of them, whose results are always distributed. *Best* selects the results that are most successful - the results that show the highest number of successful pulls - without regard to which arm produced them. *Best from Each* also selects the results that are most successful, but specifically selects the best from each of the competing arms. Finally, *Best and Worst* publishes the most successful and the most unsuccessful attempts to apply an arm in equal amounts.

However, Zollman (2009) does not consider strategies for sharing evidence which react to the changing beliefs of the agents. I introduce the *Hiding Dissent* strategy which does this to hide evidence that supports dissenting theories.

Each round a certain number of scientists have their evidence randomly chosen to be considered for publication. This evidence is then reviewed, where each scientist chosen is assigned r number of "Reviewers". These are scientists randomly chosen from the entire network, excluding the scientist who produced the evidence. These reviewers choose whether the evidence should be published. Only the evidence they decide to publish is shared with the community.

I define dissenting evidence as follows: Take two arbitrary theories (arms) X and Y, and a scientist 1 who currently has a higher expected value of Theory X to Theory Y ($\mathbb{E}_1[X] > \mathbb{E}_1[Y]$). Scientist 2 produces some evidence. This is dissenting evidence for scientist 1 in two possible cases. The first is evidence produced for Theory Y, which shows a greater success of Theory Y than scientist 1's expected value for Theory X. The second is evidence produced for Theory X which shows a lower success of Theory X than scientist 1's expected value for Theory Y. Evidence greater than expected for Theory X and evidence worse than expected for Theory Y would not be considered dissenting evidence for scientist 1. If a piece of evidence is considered dissenting by all reviewers, then they hide the evidence. This means it is not published and will not be shared with the other scientists.

I have modeled hiding dissent as something indexed to individual scientists, rather than directly tracking wider community opinion. This is more similar to the empirical findings that reviewers prefer papers supporting similar ideas to their own, rather than the organized hiding of dissent by

the *SAMJ*. It also means my model is similar to models studying self-preferential biases, e.g. Akerlof and Michaillat (2018). However, because evidence needs to be considered dissenting by more than one scientist to be considered dissent, this still tracks wider community opinion, particularly with a large proportion of reviewers. As more scientists view one theory as being better than the other, it is more likely that all reviewers chosen will judge evidence in favor of the other theory to be dissenting, therefore hiding the evidence.

Additionally, the *SAMJ* example is a case of work being desk rejected rather than being reviewed and rejected. However, because my model already does not include reviewers updating on the evidence they review, nor does it include the time it takes to review, modeling it this way will not affect my results.

My model also makes no distinction between standard dissent and epistemically detrimental dissent. Though partially a consequence of the modeling paradigm, this also both represents that scientists may not have a reliable characterization of whether the dissent is epistemically detrimental or not, and echoes the results of the empirical findings about self-preferential bias by reviewers without such a reliable characterization (Mahoney, 1977; Ernst and Resch, 1994). Nonetheless, I do also test the case where there does exist a biased agent purposefully producing detrimental dissent.

Finally, consensus in the model only refers to what Miller (2013) calls mere agreement, rather than knowledge-based consensus. Knowledge-based consensus is a stronger criterion than mere agreement and requires the satisfaction of additional conditions, such as the consensus being socially diverse, beyond just scientists agreeing on a position. Mere agreement for a statement p does not carry additional credence for p. Again, this is partially for modeling reasons. However, it makes sense in my model because what counts as dissent is indexed to individual beliefs, rather than community shared knowledge.

4 Base Model Results

All my results represent an average of 3000 runs of the model, each run consisting of 2000 timesteps (rounds). In each result shown, I set $p_A = 0.50$, $p_B = 0.51$. Each simulation presented has N = 10 scientists. 10 scientists was chosen because I found a larger number did not affect the results of the model, whilst taking longer to run. I use the *Random* strategy as a baseline to compare to *Hiding Dissent*. The parameters I alter are:

- T: The number of Tests performed each round.
- k: The number of results considered for publication.
- r: The number of scientists selected as reviewers per article.

The effects I looked at are:

- Percentage of times that the community comes to a correct consensus.
- Time taken for the community to come to the correct consensus.

I define the time taken for the community to come to the correct consensus in Theory B to be the first round where all of scientists prefer Theory B to Theory A and none switch back to preferring Theory A in any later round[2].

4.1 Analytic Analysis

I first analyse my model analytically. Using *Random*, the agents reach consensus in the limit when both myopic (Bala and Goyal, 1998) and ε-greedy (Sutton and Barto, 2018, 28). Additionally, as a corollary of the Martingale Convergence Theorem, each agents expected values of the arm they reach a consensus around will converge to the actual success of that arm with probability one (Bala and Goyal, 1998). With myopia, this consensus can be for either arm. Two conditions are required for ending with a consensus in the worse arm:

- A large enough series of tests (of arm A, B or both) which give results pushing the expected value of arm B:
  - Lower than the expected value of the arm A in enough agents in the community.
  - Lower than the actual success of arm A.
- These tests give enough evidence so that a subsequent series of tests giving worse results than expected from arm A does not change the expected values in a way that leads to enough agents testing arm B again.

[2]Specifically the correct consensus because I am looking at the time taken for good theory choice.

What counts as "enough" for these conditions is highly dependent on the state of the network, however there is a probability greater than 0 of the needed tests occurring. If this occurs it is possible for the community to have accurate beliefs about the success of the worse arm, but, as they never test the better arm (because they are myopic) they continuously believe that the worse arm is better than the actually better arm as they never get accurate information about the actually better arm.

With ε-greedy strategies, when $\varepsilon > 0$, the community will come to a consensus around the better arm, and each agent will play the better arm with probability (1-ε) (Sutton and Barto, 2018). This means smaller ε's are better for each individual agent in the limit, as they will then pull the better arm more often.

I next consider *Hiding Dissent*. With myopia, the community will still come to a consensus, correct or not, in the limit. However, the amount of results needed to come to that consensus should typically be lower than in *Random*, as, once they get close to a consensus, evidence that might move them away from that consensus is likely hidden from the community. We would therefore expect the incorrect consensus to be reached more often. The reason for this is because the series of tests needed in the conditions required for a consensus in the worse arm to be reached will be lower. This means the probability of it occurring would be higher.

These results hold in the limit, but there is no guarantee of them holding in the finite time-frames applicable for studying scientific communities. Therefore, I also provide simulations. Simulations also allow the exploration of other possible factors.

4.2 Myopic Agents

In simulations, when agents are myopic, *Hiding Dissent* performs worse than *Random* at reaching the correct consensus, as can be seen in Figure 1(a). This occurs across each parameter tested, with the only exceptions being when scientists start more "steadfast". *Hiding Dissent* performs worse because it destroys diversity in scientific practice more quickly. At the beginning of a run, each scientist has very little information. As a result, it does not take many outlier results showing Theory A to be more successful or Theory B less successful than they actually are, to lead to more scientists preferring Theory A. If more scientists prefer Theory A then the likelihood of hiding evidence in favor of Theory B, or against Theory A, increases. This reduces the likelihood of results supporting the better theory being shared with the population. Additionally, the proportion of scientists testing Theory B themselves also reduces, meaning they do not get evidence of its true

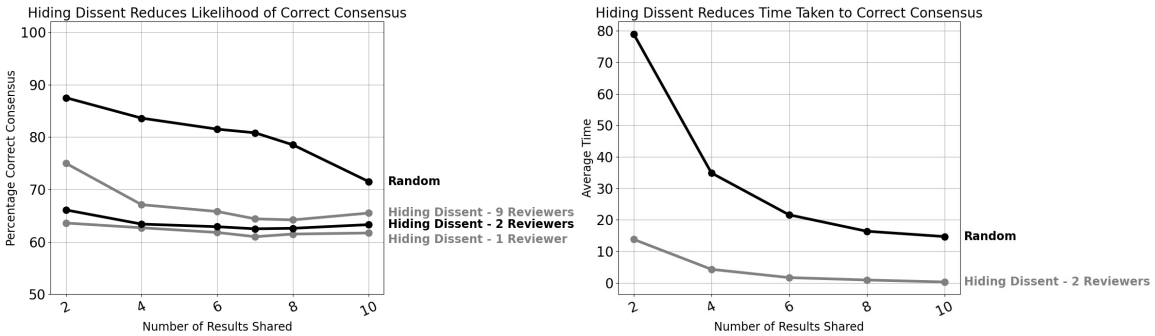

Figure 1: As the amount of evidence considered for publication increases (a) Percentage of runs correct consensus is reached and (b) Average time until correct consensus. T = 100

success. This eventually leads to the entire population preferring the worse theory.

Destruction of diversity also leads to *Random* performing worse as the number of results being shared each round increases. As more results are communicated, it is more likely that short term outlier results cause the entire community to adopt or abandon a theory. This reflects results from Zollman (2010), which found that reducing the amount of information shared could improve the long run reliability of the group, often called the "Zollman Effect".

Hiding Dissent also begins by displaying the Zollman Effect. However, once so many results are shared that consensus is typically reached within a single round, this effect reverses. After that point, it begins improving as more results are shared. This is because diversity is destroyed almost immediately anyway, so it becomes better to share more evidence quickly to potentially cancel out a small number of outliers before that destruction.

These results hold no matter the number of reviewers. However, *Hiding Dissent* does perform better with more reviewers than fewer. With more reviewers, it is less likely that all reviewers will agree on the better theory early in the run, reducing evidence hidden and keeping diversity for longer.

Whilst *Random* is better for simply reaching the correct consensus, *Hiding Dissent* is better for reaching the correct consensus quickly, as shown in Figure 1(b). This is because once a large number of agents have started preferring Theory B to Theory A, this quickly makes it likely the remaining agents will only hear evidence in favor of Theory B (besides their own in favor of Theory A if they produce it) because work is more likely to be reviewed by Theory B preferring agents. This makes the switch from disagreement to consensus much quicker.

Partially because *Hiding Dissent* is faster, it can do better than *Random* when agents start

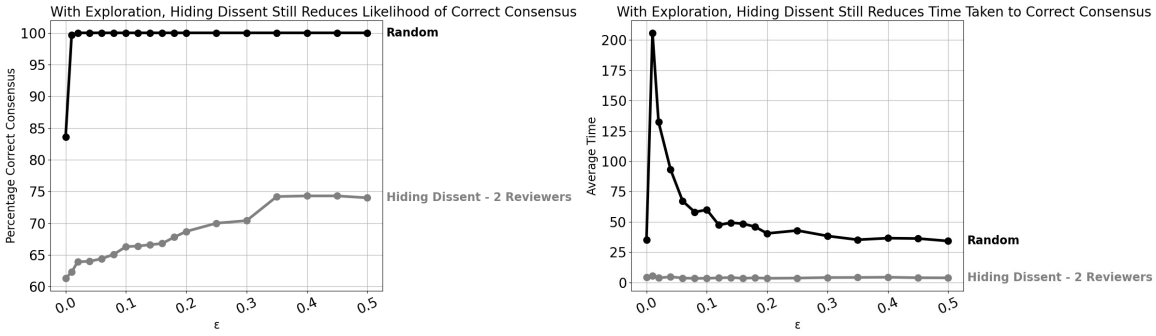

Figure 2: As ε increases (a) Percentage of times correct consensus is reached, (b) Average time until correct consensus is reached. k = 4, r = 2, T = 100

more "steadfast". Following Zollman (2010), steadfast means that scientist's initial α, β have a larger maximum value, meaning it typically takes more information to change from their initial beliefs. An example of such steadfast belief may be the discovery of mRNA vaccines by Katalin Karikó. If she had been quicker to give up her beliefs in the face of seemingly failed experiments then she may not have made this discovery.

As the possible range for α, β increases, the success of both publication strategies at reaching correct consensus increases, matching both the results from Zollman (2010), and the Karikó case. For lower values of α, β, *Random* does better than *Hiding Dissent* at reaching correct consensus. However, at higher values this can reverse. Exactly how big α, β need to be before this reversal does happen is heavily dependent on how many rounds each run goes on for. This result occurs because starting with more extreme priors means it takes more evidence to change their preferences. *Hiding Dissent* does better when scientists are more steadfast because a small number of outliers are unlikely to switch the majority of opinions, making incorrect consensus less likely. They also learn more quickly, so can overcome the extra time that the scientists need to change their views. In contrast scientists learn slower when *Random* is used, so will likely not have converged to a consensus before the number of rounds has finished. This means the simulation ends without a final state being reached.

4.3 ε-greedy Agents

I now look at ε-greedy agents. As previously, each simulation presented in this section has 10 scientists, and $p_A = 0.50$, $p_B = 0.51$.

As expected analytically, I find that both publication strategies do better at coming to the correct consensus with ε-greedy agents than they did with myopic agents. However, Figure 2(a) shows that *Hiding Dissent* still hinders correct consensus. *Random* leads to the community always reaching the correct consensus, approximating the limit, even with very low ε. Adding exploration allows for the community to permanently keep the diversity that myopia loses.

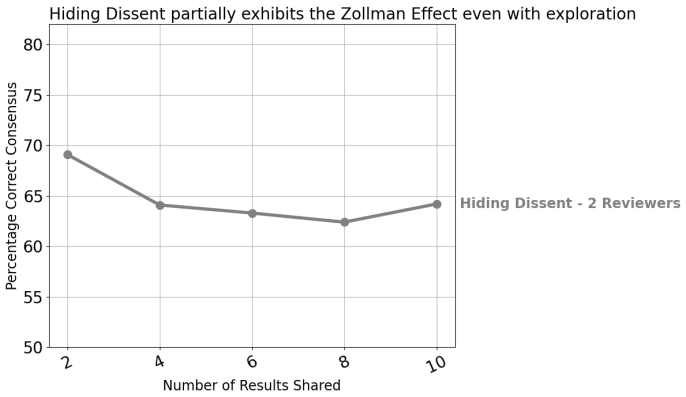

Figure 3: The percentage of runs correct consensus is reached as the amount of evidence considered for publication increases, with exploration. T = 100, r = 2, $\varepsilon = 0.05$

In contrast, *Hiding Dissent* does not always reach the correct consensus, no matter the ε. This is because exploration showing the success of the alternative theory is likely to be blocked from being shared, so the other agents would not update on the evidence. Nevertheless, it still does better with as ε increases, because a larger epsilon allows for increased diversity in a short time span – as agents are more likely to test their worse theory each round. Because *Hiding Dissent* very quickly reaches a point where opposing views are hidden, a greater likelihood of early exploration allows the sharing of more evidence for opposing views before this point is reached.

Therefore, the results with ε-greedy agents qualitatively replicate those with myopic agents. In both cases *Random* comes to the correct consensus more often than *Hiding Dissent*, whilst *Hiding Dissent* comes to the correct consensus more quickly. This may indicate robustness to my results, given they are the same under different arm selection strategies.

These results are also interesting from a network epistemology perspective. Many of the findings that have come out of the Bala-Goyal modeling framework, particularly the Zollman effect (Zollman, 2009; 2010), require the assumption of myopia. With exploration, the effects are very quickly washed out, even with only low levels of exploration. However, as I have already noted, myopia may be

an unrealistic assumption in many cases of science. Realistically, scientists may test alternative hypotheses. This may limit the applicability of such findings.

My results show the Zollman effect can hold even with with ε-greedy agents, provided the community hides dissent, or members of that community show a self-preferential bias, when considering what research should be published. With *Hiding Dissent*, when more papers are considered for sharing the community becomes less successful at coming to the correct consensus, until the point where consensus is reached within 1 round. This can be seen in Figure 3.

The reason that such results are usually washed out by exploration is because they are caused by the community losing diversity of opinion too quickly, meaning they can lock into the worse arm and never hear about the better arm with very few results. With exploration this is not possible, as agents will occasionally test the alternative arm and not lock into the worse arm. Self-preferential bias provides a way that diversity of opinion may still be lost despite agents occasionally exploring. Given evidence of such self-preferential bias has been found to various extents (Mahoney, 1977; Ernst and Resch, 1994), it provides another way that the Zollman effect may apply to real communities.

5 Modifications

So far, I have assumed scientists hide all dissenting evidence. This may be uncharitable to those who advocate hiding dissent. They could claim that they are not advocating hiding all dissent, just certain forms of dissent. Whilst we may not have a satisfactory characterization of epistemically detrimental dissent capturing all cases, there do appear to be cases where dissent is obviously epistemically detrimental and could be more safely hidden. Therefore, I now consider three modifications to the model to attempt to capture this.

The first modification I make is adding scientists having tolerance for some dissent. They do not try to hide all dissent, just the most extreme evidence for dissenting theories. Scientists may be fine with dissenting evidence if it is not too far from their expectations but be more skeptical if it is.

I define dissenting evidence with tolerance as follows: Take two arbitrary theories (arms) X and Y, and a scientist 1 who currently has a higher expected value of Theory X to Theory Y ($\mathbb{E}_1[X] > \mathbb{E}_1[Y]$). Scientist 2 produces some evidence. This is dissenting evidence for scientist 1 in two cases. The first is evidence produced for Theory Y, which shows a greater success of Theory Y than $\mathbb{E}_1[X] + t$. The second is evidence produced for Theory X which shows a lower success of

Theory X than $\mathbb{E}_1[Y] - t$. The value t is the tolerance for dissent and represents how far from a scientist's own expected value another scientist's evidence needs to be to be considered dissenting by them.

In my second modification, I consider scientists not hiding dissent immediately, but instead waiting until later in the scientific process. This represents that scientists may only hide dissenting evidence against more established scientific theories. This was explicitly stated by the *SAMJ*, when they recognized that whilst there previously was value in allowing the dissenting views, now there is not. This may also better represent the views of those advocating for hiding dissenting evidence, who may claim scientists should only hide dissenting evidence against an established consensus. Thus they should only begin hiding the evidence once the consensus has been established. I model this by having the *Random* strategy used until a certain number of rounds have passed. Once those number of rounds have passed, the community switches to using *Hiding Dissent* instead.

The third modification is when there exists a "biased agent", as in Holman and Bruner (2015). A biased agent is an agent who is interested in convincing the group of a view irrespective of that view's truth. An example is a pharmaceutical company representative trying to convince doctors to administer that company's drug regardless of its efficacy.

A biased agent only pulls their favored arm, and the results they obtain are produced by a biased distribution. In this case it is a binomial distribution with a mean of $p_A + b$, where b is the strength of their bias (Holman and Bruner, 2015). This represents the subtle ways that those producing the evidence can find to bias their results. I introduce the biased agent because, typically, epistemically detrimental dissent is considered to come from such agents, and an argument in favor of hiding dissent is that specifically epistemically detrimental dissent is what has negative effects.

It is difficult to comparably measure the success of both *Hiding Dissent* and *Random* with a biased agent. For *Random*, there is no guarantee that a scientific community will come to a stable correct consensus. This is because, once the correct consensus is reached, the only information that the scientists will hear about the worse arm is from the biased agent (assuming myopia). This then influences the agent's perceptions about the worse arm until one or more of them have a higher expectation of the worse arm than the better arm. They will then test the worse arm themselves, producing evidence that brings them back to having a higher expectation of the better arm. Therefore, I cannot use the percentage of times that the community comes to a correct consensus and the time taken for the community to come to the correct consensus.

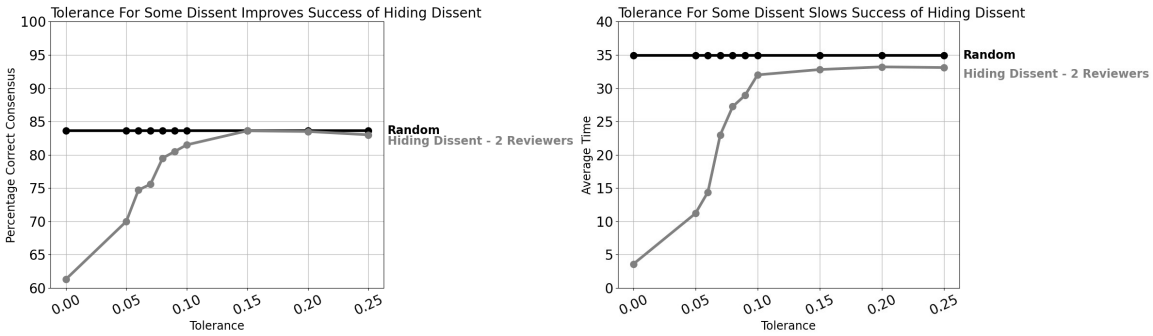

Figure 4: As Tolerance for Dissent increases (a) Percentage of times correct consensus is reached, (b) Average time until correct consensus is reached. k = 4, T = 100

Holman and Bruner (2015) look at the last 1,000 rounds of a 2,000-round simulation and then determine the frequency that the scientists perform arm B during that time. This would not work alone for my model. When using *Hiding Dissent* the community does still come to consensus (either correctly or incorrectly). This means that the frequency that the scientists perform arm B during that time during each run is always going to be 0% or 100%, so does not provide useful comparable results.

Therefore, I provide two measures. The first is Holman and Bruner's (2015) average frequency that the scientists perform arm B in the last 1,000 rounds. The second is the the percentage of runs that the agents pull the better arm over 50% of the time during the last 1,000 rounds. The latter is measuring the times they pull the better arm more often than not. The reason for this is that if agents are using the better arm more often than not, then it seems realistic to say that they think that arm is better.

5.1 Modification Results

I now present results for these modifications. As previously, unless explicitly stated each simulation presented in this section has 10 scientists, $p_A = 0.50$, $p_B = 0.51$, and I assume myopia.

I first test the addition of tolerance for dissent. As Figure 4(a) shows, adding tolerance for dissent does make *Hiding Dissent* more successful at coming to the correct consensus, and as the tolerance increases, the group becomes more successful. This is because they are less likely to hide dissent, and therefore diversity of practice will not be destroyed as quickly. However, it only does as well as *Random* when very little dissent is hidden because there is so much tolerance for dissent.

Figure 5: As the round where the scientists begin hiding dissent increases (a) Percentage of times correct consensus is reached. (b) Average time until correct consensus. k = 4, r= 2, T = 100

It also becomes as slow as *Random* in this case too. It therefore seems as if adding a tolerance of dissent to *Hiding Dissent* has no advantages over *Random*.

Secondly, I test the case where scientists only start to hide dissent after some number of rounds have passed. We should expect that it would allow diversity of practice to last longer, leading to greater success at reaching the correct consensus for *Hiding Dissent*. This is the case, as seen in Figure 5(a), and the longer that they wait the greater this success becomes until eventually it becomes as successful as *Random*.

Whilst waiting leads to *Hiding Dissent* becoming as successful as *Random*, it also reaches the correct consensus on average more quickly than *Random* does, as can be seen in Figure 5(b). The reason for this is because as soon as *Hiding Dissent* begins, consensus is reached very quickly, eliminating the outlier cases, whereas *Random* may take a very long time to reach consensus in certain outlier cases.

Finally, I test the case where there exists a biased agent. This agent publishes in the same journal as the other scientists and is subject to the same selection strategies. For the results shown, the biased agent has a bias of $p_A + 0.03$, i.e. they are biased towards the worse arm.

As Figure 6(a) shows, *Random* does better than *Hiding Dissent* at using the better arm more often than not in the last 1000 rounds. *Hiding Dissent* performs worse because the results of the biased agent affect the consensus that is formed. The biased agent may lead to more members of the community preferring the worse arm, increasing the likelihood that evidence in favor of the better arm is hidden, and eventually leading to the community coming to the incorrect consensus which they cannot then leave. In contrast, with *Random*, whilst the evidence produced by the biased agent may temporarily cause other agents to prefer the worse arm, they will then test the worse arm and

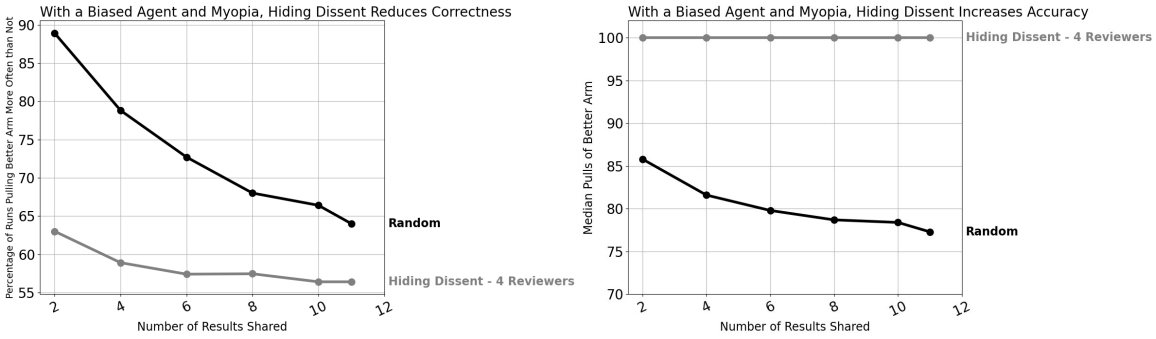

Figure 6: With myopic agents and a biased agent (a) Average percentage of runs agents use better arm over 50% of the time during the last 1000 rounds. (b) Median usage of better arm over last 1000 rounds. T = 100, r = 4

produce accurate evidence, which is shared with the community. This leads to them coming to prefer the better arm again.

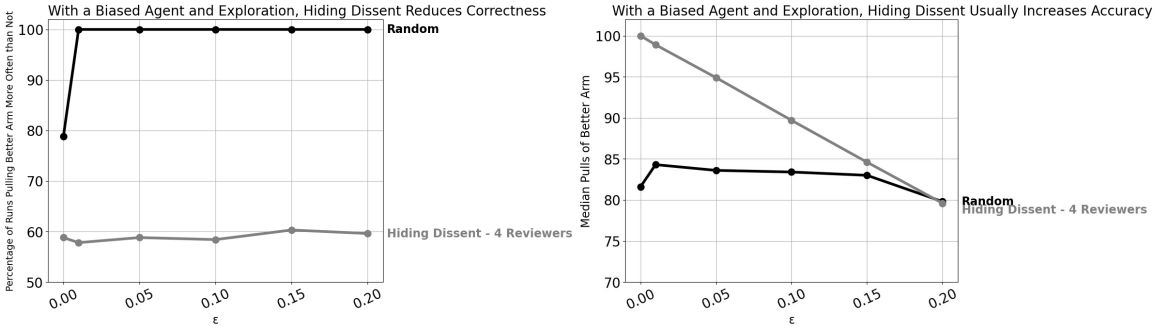

Figure 7: With ε-greedy agents and a biased agent (a) Average percentage of runs agents use better arm over 50% of the time during the last 1000 rounds. (b) Median usage of better arm over last 1000 rounds. T = 100, k = 4, r = 4

However, I find that the average frequency of scientists using the better arm in the last 1000 rounds is very similar for both strategies, and under certain parameters *Hiding Dissent* actually does better than *Random*. This is because, in a run, *Hiding Dissent* either always pulls or never pulls the better arm, and the former occurs more often. This is also shown by the median frequency of scientists using the better arm in the last 1000 rounds, seen in Figure 6(b), which is always 100% for *Hiding Dissent*. In contrast, *Random* may occasionally stop using the better arm to test the worse arm before coming back to using the better arm.

Therefore, these two measures come apart. Taking a random run, *Random* is more likely to have scientists pulling the better arm more often than not, but it never reaches a stable consensus.

However, in the specific individual runs where *Hiding Dissent* does pull the better arm more often than not, it will perform better than if they had used *Random*, because every scientist will always pull the better arm and actually reach a stable consensus. There appears to be a trade-off here between being correct in less runs but being more accurate in those runs, versus being correct in more runs but being less accurate in those runs, where correctness is taken to be the community believing the actually better arm is indeed the better arm, and accuracy is taken to be closeness to the real success of that better arm. These results may suggest that if a case exists where having a consensus is more useful, then *Hiding Dissent* may be better.

These findings continue to hold with an ε-greedy community, suggesting robustness. As Figure 7(a) shows, a ε-greedy strategy still leads to *Random* doing better than *Hiding Dissent* at using the better arm more often than not in the last 1000 rounds. It also increases the average amount of times that *Random* pulls the better arm over the last 1000 rounds. This increased success happens because *Random* agents will never lock into the worse arm, and the evidence produced by the biased agent is less likely to briefly shift agents away from the better arm as they already have preemptive counter-evidence.

In contrast, *Hiding Dissent* with an ε-greedy strategy will have to occasionally pull the worse arm, leading to a lower average usage. However, even excluding the enforced pulling of the worse arm due to an ε-greedy strategy, *Random* does better than *Hiding Dissent* at pulling the better arm more often on average over the last 1000 rounds.

Nonetheless, we still see the same trade-off between correctness and accuracy. When *Hiding Dissent* does lead to the community pulling the better arm, the community still ends up pulling the better arm in every round (except the rounds with exploration due to being ε-greedy) because they have reached a consensus of belief. This is unlike *Random*. This can again be seen by looking at the median amount of time pulling the better arm over the last 1000 rounds, in Figure 7(b). For *Hiding Dissent* it is always roughly $1 - \varepsilon$ and usually higher than that of *Random*. So the results are robust under different strategies to choose which arm to pull.

Finally, I test what happens when there exists more than one biased agent. As the number of biased agents increases, both publication strategies become much less successful across both measures. Nonetheless, I find that eventually *Hiding Dissent* does do better than *Random*. In a community with 10 honest scientists with myopia, k=4 and r = 4, this point is at 3 biased agents. With the same parameters, but ε-greedy honest scientists with $\varepsilon = 0.05$ this point becomes 4

biased agents. This suggests that with many biased agents, *Hiding Dissent* may better protect the community, though both strategies are highly unsuccessful. However, this may just be a modeling artifact. Bandit problems struggle to deal with cases where many agents are pulling from different bandit arm distributions. It also seems unrealistic to suggest a community consisting of such a large number of biased agents would act with such simple dynamics.

6 Discussion

The model presented is very idealized, so we cannot draw detailed recommendations for how journals should, or should not, publish evidence in the real world. Despite this, we may draw some broad conclusions.

Firstly, journals hiding dissenting evidence may itself be epistemically detrimental if no reliable characterization of epistemically detrimental dissent is available, complimenting the views of de Melo-Martín and Intemann (2018). Additionally, these results hold even when there exists a biased agent purposefully producing epistemically detrimental dissent. Though keeping in mind the limitations and idealizations of the model, these results would seem to support a policy of not hiding dissent even in cases where dissent is more obviously epistemically detrimental.

However, in contrast to de Melo-Martín and Intemann (2018), my models suggest a reliable characterization of epistemically detrimental dissent may not be needed if the community is already sufficiently far along the path to consensus. If enough evidence has already been gathered, hiding all dissent can quickly promote consensus in the better theory. This may justify hiding dissent in areas where large amounts of evidence pointing towards a certain theory already exists. For example, it may justify the *SAMJ's* decision to no longer accept evidence for certain dissenting views about AIDS. However, my model is very idealized and cannot support claims about how much evidence would be enough before hiding dissent becomes preferable.

One way my model is idealized is the use of the *Random* strategy. The *Random* strategy does not capture the selectivity used by real journals, which may affect the actual success of not hiding dissent. However, *Hiding Dissent* also does worse at reaching correct consensus than all of the other strategies considered in Zollman (2009). That these results hold with many different journal strategies suggest the results may be robust.

My results also accord with those from models created using very different modeling paradigms.

I modeled hiding dissent as a bandit model where scientists express self-preferential bias. Therefore, I can compare my results with evolutionary models which also look at self-preferential bias, such as Akerlof and Michaillat (2018), and Smaldino and O'Connor (2020). These models find that self-preferential biases in reviewing can lead to the stabilization of false paradigms. My results accord and also show that self-preferential biases promote consensus in the worse theory, again suggesting robustness in my results.

Even though hiding dissent reduces consensus in the better theory, my results do highlight situations where it may be beneficial. The first is when speed until consensus is important, even if there is a higher risk of being incorrect. This may be advantageous if advice needs to be given quickly because if decisions are not made quickly the situation may be more difficult to deal with, leading to greater harms, for example, during an epidemic.

Secondly, some empirical studies suggest consensus increases public acceptance of science, and even modest dissent undermines it (Lewandowsky et al, 2013; Aklin and Urpelainen, 2014). If these results are true, hiding dissent may have benefits not captured by the models presented. As my results show, when there exists a biased agent and dissent is not hidden, the scientific community is prevented from reaching a consensus, potentially reducing public acceptance of science and pointing to a function for hiding dissent. However, these studies are controversial. Findings also suggest that public awareness of suppressed dissent undermines public trust in science, so hiding dissent in order to increase public acceptance of science may end up having the opposite effect (Ryghaug and Skjølsvold, 2010).

These issues with public trust indicate a further limitation of my models. The models only consider limited measures of epistemic success: likelihood and speed of correct consensus. Whilst these are standard measures of success in network epistemology, there are far more epistemic values that scientists may wish to promote, such as justification, or public trust in science. My model cannot inform whether hiding dissent helps or hinders these other values. However, if it is the case that public trust is eroded by awareness of suppressed dissent, then my model may provide a complimentary reason against hiding dissent. Hiding dissent may be negative even when you exclude other negatives such as undermining public trust in science, because it may just be bad for forming true beliefs in scientists.

Away from the debate over epistemically detrimental dissent, my findings are also useful for network epistemology. Zollman (2009; 2010) showed that scientific communities may do worse when

more information is shared, as diversity is lost too quickly. These findings rely on an assumption of myopia, which may be an unrealistic assumption for many scientific communities. My model, using the same modeling paradigm, has shown that these results can still hold with non-myopic communities, provided that there exists some form of hiding dissent, or self-preferential bias for one's own theories.

Overall, my model compliments and supports more traditional arguments against hiding dissent. However, it has only looked at one specific form of hiding dissent, and a limited measure of epistemic success. There are many more dynamics that have not been considered that could be modeled further.

References

Akerlof, G. and Michaillat, P., (2018). "Persistence of false paradigms in low-power sciences". *Proceedings of the National Academy of Sciences*, 115(52), 13228-13233.

Aklin, M. and Urpelainen, J., (2014). "Perceptions of scientific dissent undermine public support for environmental policy". *Environmental Science & Policy*, 38(2014), 173-177

Berry, D. A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. London: Chapman and Hall

Biddle, J.B. and Leuschner, A. (2015). "Climate skepticism and the manufacture of doubt: can dissent in science be epistemically detrimental?" *European Journal for Philosophy of Science*, 5(3), 261–278.

Cook, J. (2017). "Response by Cook to 'Beyond Counting Climate Consensus'." *Environmental Communication* 11(6):, 733–735.

de Melo-Martín, I. and Intemann, K. (2012). "Scientific dissent and public policy". *EMBO reports*, 14(3), 231–235.
———— (2014). "Who's Afraid of Dissent? Addressing Concerns about Undermining Scientific Consensus in Public Policy Developments". *Perspectives on Science*, 22(4), 593–615.
———— (2018). *The fight against doubt: how to bridge the gap between scientists and the public*. New York: Oxford University Press.

Ernst, E., and Resch, K.L. (1994). "Reviewer bias: A blinded experimental study". *The Journal of Laboratory and Clinical Medicine*, 124(2), 178-182

Holman, B and Bruner, J.P. (2015). "The Problem of Intransigently Biased Agents." *Philosophy of Science*, 82 (5), 956-968

Kummerfeld, E and Zollman, K. J. S. (2016). "Conservatism and the Scientific State of Nature". *The British Journal for the Philosophy of Science*, 67(4), 1057–1076

Lewandowsky, S., Gignac, G.E, and Vaughan, S. (2013). "The Pivotal Role of Perceived Scientific Consensus in Acceptance of Science." *Nature Climate Change* (4):399–404

Mahoney, M. J. (1977). "Publication prejudices: An experimental study of confirmatory bias in the peer review system". *Cognitive Therapy and Research*, 1(2):161–175

Mayr, E. (1982). *The Growth of Biological Thought: Diversity, Evolution, and Inheritance.* Cambridge Massachusetts: Belknap Press

Miller, B. (2013). "When is consensus knowledge based? Distinguishing shared knowledge from mere agreement." *Synthese*, 190(7), 1293–1316.
——— (2021). "When Is Scientific Dissent Epistemically Inappropriate?". *Philosophy of Science*, 88(5), 918-928.

Nash, E. J. (2018). "In Defense of "Targeting" Some Dissent about Science". *Perspectives on Science*, 26(3), 325–359.

Oreskes, N. (2017). "Response by Oreskes to 'Beyond Counting Climate Consensus'." *Environmental Communication* 11(6), 731–732

Oreskes, N., and Conway. E.M (2010). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming.* London: Bloomsbury

Ryghaug, M., and Skjølsvold, T.M. (2010) The Global Warming of Climate Science: Climategate and the Construction of Scientific Facts, *International Studies in the Philosophy of Science*, 24:3, 287-307.

Smaldino, P and O'Connor, C. (2020). "Interdisciplinarity can aid the spread of better methods between scientific communities.". Forthcoming in *Collabra*

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement learning: An introduction Second Edition.* Cambridge: The MIT Press.

van Niekerk, J.P. (2003). "Politics Must Move Mainstream On AIDS." *South African Medical Journal* 93(3):154.

Zollman, K. J. S. (2007) "The Communication Structure of Epistemic Communities." *Philosophy of Science* 74(5), 574–87.
——— (2009). "Optimal Publishing Strategies". *Episteme*, 6(2), 185–199.
——— (2010). "The Epistemic Benefit of Transient Diversity". *Erkenntnis*, 72(1), 17–35.