Why Bayesian Agents Polarize

Olsson, Erik J

Published in:
The Epistemology of Group Disagreement

2020

Document Version:
Early version, also known as pre-print

Why Bayesian Agents Polarize

Erik J. Olsson

Abstract: A number of studies have concluded that polarization may be rational in the sense that even ideal Bayesian agents can end up seriously divided on an issue given exactly the same evidence. In this spirit, Pallavicini, Hallsson and Kappel (2018) demonstrate that group polarization is a very robust phenomenon in the Bayesian so-called Laputa model of social network deliberation. However, in their view polarization arises due to a failure of Laputa to take into account higher-order information in a particular way, making the model incapable of capturing full rationality. I show that taking into account higher-order information in the way proposed by Pallavicini *et al.* fails to block polarization. Rather, what drives polarization is expectation-based updating in combination with a modelling of trust in a source that recognizes the possibility that the source is systematically biased. Finally, I show that polarization may be rational in a further sense: group deliberations that lead to polarization can be, and often are, associated with increased epistemic value at the group level. The upshot is a strengthened case for the rationality of polarization.

1. Introduction

Many societal debates are polarized in the sense that a substantial proportion of the population hold one view while the remaining part is of the diametrically opposite opinion. Abortion, climate change, immigration and the merits of Donald Trump's presidency come to mind as issues upon which, at the time of writing, opinions are seriously divided in many Western societies. A somewhat comforting thought is that this only means that one party must be not only wrong but wrong because irrational. If people end up on the wrong side in a dispute because they are irrational, that would suggest that we could avoid or even eradicate polarization by educating people in the normatively correct way of reasoning and weighing evidence.[1]

[1] Acknowledgement: I would like to thank Fernando Broncano-Berrocal and Adam Carter for their valuable comments on an earlier version of this article.

But what if polarization is not irrational, but even rational? More carefully put: what if even people who carefully consider their evidence in conformity with impeccable principles of rationality may still end up divided on the issues at hand, and what if this happens, not only once in a while, but frequently?

In fact, a number of studies have concluded that polarization may result from rational processes (Cook and Lewandowsky 2016, Easwaran, Fenton-Glynn, Hitchcock, and Velasco 2016, Jern, Chang and Kemp 2014, Kelly 2008). One such body of work stems from the Bayesian community and, in particular, from research exploring the Laputa simulation model for social network communication developed by Staffan Angere and Erik J. Olsson (see Olsson, 2011, for an introduction and overview). To convey the main ideas, in Laputa two or more inquirers are concerned with the same question whether a factual proposition p ("Climate change is man-made", "Trump will be re-elected" …) is true or false. Their inquiry is an on-going process that takes place over time in a network of connected inquirers. Each inquirer can at any time consult her own outside source as to whether p is true. The inquirers can at any time ask other inquirers in the social network to which they are connected whether p. The outside sources are somewhat but (typically) not fully reliable. The inquirers (typically) do not fully trust their outside sources, nor do they fully trust each other; rather they update trust dynamically as they receive information from their outside source and/or their network peers. The situation allows for different social practices to be implemented (e.g. much, little or no communication) and the question arises which practice is most beneficial in the interest of inquirers' arriving at the true answer to the underlying question.

Olsson (2013) showed by computer simulation how communication in Laputa among ideally rational agents leads to polarization under various plausible conditions. He concluded (p. 130): "[t]o the extent that Bayesian reasoning is normatively correct, the perhaps most surprising, and disturbing, results of this study are that polarization and divergence are not necessarily the result of mere

irrational 'group thinking' but that even ideally rational inquirers will predictably polarize or diverge under realistic conditions."

How robust is polarization in Laputa? This question is thoroughly investigated in an extensive (55 page) study by Pallavicini, Hallsson and Kappel (2018). The authors also consider and eventually rule out a number of intriguing hypotheses about what causes polarization in Laputa networks. Vindicating Olsson's study, the authors find that "groups of Bayesian agents show group polarization behavior under a broad range of circumstances" (p. 1).[2] However, rather than concluding that polarization may be rational, they argue that the results are, in the end, explained by an alleged failure of Laputa to capture rationality in its full sense. In particular, they notice that agents in Laputa lack the ability to respond to "higher-order evidence". This lack is what, according to the authors, ultimately explains the fact that agents polarize. As a remedy, they sketch a revised updating mechanism that they think does justice to higher-order evidence. Pallavicini *et al.* do not provide a detailed rule, nor do they demonstrate analytically or by computer simulation that their proposal would prevent groups from polarizing.

In this paper, I show that incorporating higher-order evidence in the way Pallavicini and her colleagues suggest fails to block polarization in Laputa. Thus, failure to comply with the revised updating rule cannot be the root cause of polarization. Rather, what drives polarization, on closer scrutiny, is expectation-based updating in combination with a modelling of trust that recognizes the possibility that the source is biased, i.e. gives systematically false information. Finally, I demonstrate that polarization is rational in a further sense: epistemic practices that lead to polarization can be, and often are, associated with increased epistemic value at the group level. I conclude that the case for the rationality of polarization has been significantly strengthened rather than weakened.

[2] All references to page numbers in Palavicini *et al.* (2018) are to the online version; no printed version had appeared at the time of writing.

In section 2, having given a brief snapshot of relevant parts of the Laputa model, I summarize the findings in Pallavicini *et al.* (2018) concerning the ubiquity of polarization in the model. In section 3, I consider the authors' argument for a revised rule intended to take higher-order evidence into account in order to block polarization among deliberating agents. Section 4 is devoted to an investigation into polarization and epistemic value. In the final section, I summarize the results and draw some additional conclusions.

2. Background on Laputa and polarization

The Laputa framework for studying epistemological aspects of deliberation in social networks is in many ways a Bayesian model. For instance, an agent's belief state is represented by a probability distribution corresponding to the agent's degree of belief in the proposition in question. Moreover, updating of degrees of belief (credence) takes place through conditionalization on the evidence. The evidence here means either evidence coming from inquiry (a personal outside source not part of the network) or from a source in the network. While the model is generally complex, the messages that can be sent and received by agents are only of two kinds: p or not-p, for a proposition p. Thus, Laputa models network activity in response to a binary issue: guilty or not guilty, climate change is man-made or not, and so on. At any step in a deliberation, agents can communicate with other agents to whom they are connected, or they can conduct inquiry in the sense of receiving information from their outside source. The distributions that determine the chance of communication, of conducting inquiry and so on at a given point in the deliberation are parameters in the model. The information obtained leads to a new credence through conditionalization on the evidence. An important point here is that the evidence will be of the kind "Source S reported that p" rather than p itself. This opens up for the possibility of not taking what a source says at face value.

A novel feature is that the Laputa framework incorporates a Bayesian mechanism for representing the degree to which an agent trusts her own inquiry (outside source) as well as her network peers. Trust here means perceived reliability and is represented as a "trust function" over all possible

reliability profiles – from being systematically biased/anti-reliable to being systematically truth-telling – representing how likely those profiles are taken to be at a given stage in the deliberation. It turns out that for some purposes trust can be represented by a single number: the expected value of the trust function. An inquirer's new trust function after having received information is obtained via conditionalization on the evidence.[3] In the simple case in which the inquirer has a normally distributed trust function with expected value 0.5 and assigns p a degree of belief exceeding 0.5, the inquirer will, upon receiving repeated confirming messages from one source source, update her trust function so that it approaches a function having expected value 1, representing full trust in the source. Interestingly, representing trust by a function rather than a single number allows for complex interactions between different parameters. Two agents who assign the same degree of belief to p, have trust function with the same expected value and receive exactly the same information (say, from inquiry) may nevertheless, depending on their initial trust functions, end up with very different degrees of belief and new trust functions.

Updating in Laputa is quite complex, especially the updating of trust. Fortunately, there exists a computer implementation that does the computations automatically which, as we will see, greatly facilitates investigation into the model and its consequences (see Olsson, 2011, for an overview). For the purposes of this paper, there is very little the reader needs to know about Laputa in addition to what has already been explained. An exception is the "Laputa table" (Table 1) containing the derived updating rules for belief (credence) and (expected trust) (see Olsson, 2013, for derivations).

[3] Derivations of the credence and trust updating functions can be found in Angere and Olsson (2017). Pallavicini *et al.* (2018) is also rich in background information on Laputa.

	Message expected	Neither nor	Message unexpected
Source trusted	+ (↑)	↑ ( )	- (↓)
Neither nor	0 (↑)	0 ( )	0 (↓)
Source distrusted	- (↑)	↓ ( )	+ (↓)

Table 1: Derived updating rules for belief (credence) and trust.

Table 1 is a condensed representation of how updating in Laputa works. Consider, for example, the upper left-most cell in the table. This is the case in which an agent receives an expected message from a trusted source. That the message, let us say p, is expected means that the receiving agent assigns p a credence higher than 0.5. That the source is trusted means that the receiving agent assigns a trust function to the source such that the expected value of that function is higher than 0.5. What should happen in this case? The plus sign here means that the receiving agent will strengthen her current belief. In our example, it means that she will believe even more strongly that p is the case. The up-arrow means that the receiving agent will trust the source even more. Similarly, the minus sign in Table 1 means that the receiving agents weakens her current belief, and the down-arrow means that she trusts the source less than she did before. It is important to understand that the rules described in Table 1 are derived rules in the sense that they follow from the underlying Bayesian machinery.[4] [5]

In the following, I will use the terminology in Pallavicini *et al.* (2018) regarding polarization and related concepts. Thus, "polarization", as the term is often used in social epistemology to denote the

[4] Since the Laputa model was first described (Olsson, 2011) it has been applied to a number of issues in epistemology, such as norms of assertion (Olsson and Vallinder, 2013, Angere and Olsson, 2017), the argument from disagreement (Vallinder and Olsson, 2013a), the epistemic value of overconfidence (Vallinder and Olsson, 2013b), the problem of jury size in law (Angere, Olsson and Genot, 2015), peer disagreement (Olsson, 2018) and the epistemic effect of network structure (Angere and Olsson, 2017, Hahn, Hansson and Olsson, 2018).
[5] Collins *et al*. (2018) examined, theoretically and empirically, the implications of using message content as a cue to source reliability in the spirit of Laputa. They presented a set of experiments examining the relationship between source information and message content in people's responses to simple communications. The results showed that people spontaneously revise their beliefs in the reliability of the source on the basis of the expectedness of a source's claim and, conversely, adjust message impact by perceived reliability, much like how updating works in Laputa. Specifically, people were happy downgrading the reliability of a source when presented with an unexpected message from that source.

tendency of deliberation to strengthen the pre-existing attitudes in a group of like-minded people, will instead be called “escalation” (Isenberg, 1986, Sunstein, 2002). Escalation will not play any major role in this paper. The term “group polarization”, or simply “polarization”, will be reserved for the phenomenon of a group being seriously divided on an issue. In the extreme case, half the group believes p and the other half not-p.

The degree of polarization for a given network of agents can be computed as the average (root mean square of the) deviation of individual credence from the mean. Thus, a network in which every agent has the same credence in p has polarization 0 (minimum). One in which half the inquirers are certain that p and half are certain that not-p has polarization 0.5 (maximum). We are often interested in the extent to which polarization has increased or decreased following deliberation. This we can find out by simply computing the final (post-deliberation) degree of polarization minus the initial (pre-deliberation) degree of polarization. A positive value means that agents in the network have become more polarized as the result of deliberation. A negative value means that they have become less so.

Olsson (2013), in section 5, studied polarization by means of computer simulation for what he called a “closed room” debate without anyone in the network undertaking inquiry (inquiry chance set to 0). Rather, all the activity consisted in communication between mutually trusting agents in the network, where the initial trust values were drawn from a (truncated) normal distribution with a mean value of 0.75. The initial degree of belief (credence) in p was assumed to be normally distributed with a mean value just above 0.5. Laputa was instructed to generate 1,000 networks satisfying these and some other reasonable constraints, allowing each network to evolve ten time steps (“rounds”). The simulation was run in “batch mode” with Laputa collecting average results over all the network runs. The result was belief escalation towards degree of belief 1 in p. In fact, after 10 rounds, virtually all agents believed fully that p, with very few exceptions.

Olsson (2013) also studied some conditions under which agents in Laputa polarize. The three remaining cases were considered: people trust but are biased to give false reports, people distrust

but tell the truth, and people distrust and are biased to give false reports. Olsson concluded that the first two cases, characterized by a lack of social calibration in the sense that there is a mismatch between trust and actual reliability, give rise to polarization (p. 128). Additionally, Olsson walked through a simplified case involving just two communicating agents under similar circumstances to see how polarization arises, step by step.

Pallavicini *et al.* (2018) contains a much more detailed analysis of the conditions under which agents in Laputa polarize and, in particular, how the agents' initial beliefs in p affect subsequent polarization. Trust is also varied, but in a different way than in Olsson's study, making the two studies non-trivial to compare. Pallavicini *et al.* look at five different cases concerning initial beliefs (Figure 1). In the first case, initial credence are drawn from a normal distribution with mean 0.5. (undecided group). They also consider two cases in which agents are already polarized, in one case more so than in the other. Finally, in two of the examples, the distributions of initial belief are tilted towards believing p, thus corresponding to the setting in Olsson (2013), in one case more so than in the other (called the "suspecting groups").

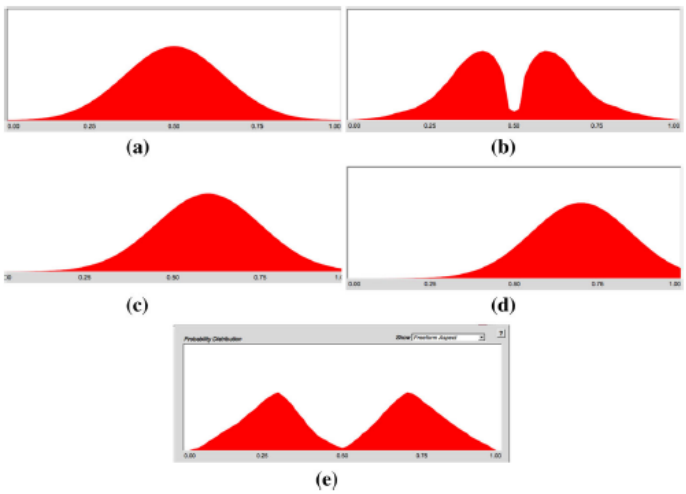

Figure 1: Initial distributions of degrees of belief for the five different groups: the undecided group (a), the polarized group (b), the suspecting group (c), the very suspecting group (d) and the very polarized groups (e). Adapted from Pallavicini *et al*. (2018), p. 15.

The other parameters studied by Pallavicini and her colleagues involve are inquiry trust and communication trust, i.e. peer trust. It should be noted that both communication chance and inquiry chance are set to be governed by a uniform distribution over [0, 1] in the study (see their Appendix B.2). Thus, there will usually be both communication and inquiry going on in a particular network generated by Laputa in batch mode. The cases considered regarding trust are the following (p. 16): agents generally trust themselves (inquiry) as much as they trust others (i.e. the inquiry trust distribution is the same as for the communication trust distribution), agents generally trust themselves (inquiry) more than they trust others (i.e. the inquiry trust distribution has a higher mean than the communication trust distribution), and agents generally trust others more than they trust themselves (inquiry) (i.e. the communication trust distribution has a higher mean than the inquiry trust distribution). Combined with the five belief distributions, this leaves the authors with 15 (5*3) possibilities to consider.

The striking result is that groups polarize under all conditions. In the undecided and polarized cases, agents end up divided into two equally large camps: one camp believing assigning credence 1 to p and the other credence 1 to not-p (credence 0 to p). The suspecting cases (in which inquirers are initially statistically inclined to believe p) lead to polarization as well, although here the camp assigning credence 1 to p is bigger than the camp assigning credence 1 to not-p. The different trust conditions studied basically do not affect these results. Pallavicini and her colleagues also perform a very extensive robustness study by varying trust in more fine-grained ways, which effectively means that they simulate 285 different groups. Their conclusions are noteworthy (p. 22): “The very surprising result of this simulation is all of the groups polarized to some degree. In fact, most groups polarized to the maximum level. There were no conditions under which depolarization occurred.”

Thus, “the observed polarization behavior is a very stable phenomenon for these Bayesian agents” (p. 20).

At this point, it would seem that Pallavicini *et al.* have achieved a striking vindication of the “disturbing” conclusion of Olsson’s 2013 study to the effect that even ideally rational Bayesian agents polarize under a broad range of conditions. Surprisingly, however, this is not the moral Pallavicini and her colleagues draw from their investigation.

3. Pallavicini, Hallsson and Kappel on the cause of polarization

Pallavicini and her co-authors devote several sections of their paper to inquiring into the root causes of group polarization in Laputa. For instance, one might think that it is the way trust is updated in Laputa that ultimately causes polarization. However, the authors show that polarization occurs even if trust updating is turned off in the Laputa simulation program, finding that “deactivating trust-updating does not stop the polarization behavior” (p. 26); rather “the trust-updating speeds up an already existing process” (p. 27).

Another possible explanation of polarization considered by Pallavicini *et al.* is the fact that once Bayesian agents research a credence of 1 in a proposition, they cannot – for familiar reasons – change their mind. However, this, too, does not explain polarization (p. 32): “The aspect of Bayesian agents that they cannot change their minds after reaching an extreme credence is not the cause either, it just means that the results will be stable after a certain point.” They also look into the possibility that “double-counting the evidence” might be causing polarization. In Laputa, an agent A updates belief and trust every time a network peer B sends a message to A, regardless of whether B has already asserted the same message before without performing any inquiry in-between. Pallavicini *et al*. test the hypothesis that double counting causes polarization by turning on a feature of Laputa which prevents agents from sending messages without having received an intermediate message from inquiry, observing that “[i]n all cases the groups reached maximal polarization within the 30 time steps of the original simulation” (p. 31). Hence, double-counting is not the root cause of

polarization either. Finally, they manage to exclude the possibility that polarization results from networks having a certain density, i.e. a high proportion of communication connections between agents.

So what, then, is responsible for the ubiquitous polarization we see in social networks in Laputa? Researchers working in the Bayesian tradition have shown how their models are compatible with polarization among agents. Pallavicini *et al.* consider two such accounts at some length, one due to Jern, Chang and Kemp (2014) and the other to Kelly (2008). Both studies conclude that Bayesian updating is compatible with polarization in cases in which agents have different background beliefs. Jern *et al*. make this point as follows (p. 209):

> [S]uppose that a high cholesterol test result is most probable when a patient has Disease 1 and low blood sugar, or when a patient has Disease 2 and high blood sugar. Then two doctors with different prior beliefs about the patient's blood sugar level may draw opposite conclusions about the most probable disease upon seeing the same cholesterol test result D.

This explanation seems to transfer directly to Laputa. In Laputa, agents may have different background beliefs not only regarding the proposition p but also regarding the trustworthiness of a source. Suppose, to consider a case similar to that investigated by Jern *et al*. (2014), that A has a high credence in p and trust the source and B has a low credence in p (high credence in not-p) and distrusts the same source, i.e. considers it to be potentially biased (reporting false propositions). Now the source says that p. For A, this is an expected message coming from a trusted source. By the Laputa updating table (upper left-most cell in Table 1), A will believe p even more strongly than before. For B, it is an unexpected message coming from a distrusted source. By the Laputa table (lower right-most cell in Table 1), B will believe not-p even more strongly than before. Thus, agents polarize and what is responsible for this are precisely differences in background beliefs and the effects those differences have given the underlying Bayesian machinery. Those effects essentially mean, in the case of credence updating, that evidence coming from sources believed to be

trustworthy is taken at face value, whereas evidence coming from sources believed to be biased is taken as "evidence to the contrary".

Yet, Pallavicini *et al.* disagree with this explanation, writing (p. 35): "The group polarization we see in our simulations does not depend on any particular prior assumptions made by subjects in the group, as our polarization results are robust for more than 200 different groups." Their conclusion is that "we can not amend the explanation from Jern *et al.* to argue that our polarization results are rational" (ibid.). One could object that it is not necessary for Laputa agents to polarize in the way just demonstrated that "particular prior assumptions" are made by subjects in the group, whatever this might mean more precisely, so long as agents have qualitatively different beliefs concerning p and the trustworthiness of the source.

Pallavicini *et al.* (p. 36) consider a similar explanation due to Kelly (2008):

> [I]t may be possible to give an interpretation of what goes on in the simulations which is compatible with Kelly's account. The information in the Bayesian network consists of agents communicating with each other and the agents doing inquiries on their own. The received messages and results of inquiry are what would be the narrow evidence on Kelly's view. All of the agents in the simulation update their beliefs in the same way, based on a formula that incorporates the agent's prior belief, all of the information that an agent receives at a given time (the narrow evidence) and how much the agent trusts the sources that are giving the information … This updating based on the collection of the prior belief, the narrow evidence and the trust in the sources could be understood as the broader evidence.

Furthermore (p. 36):

> Since in none of our simulations the agents have the same narrow evidence and since the agents do not share their broad evidence when communicating …, it makes sense on Kelly's view that the agents update their degrees of beliefs in different directions.

Yet, once more Pallavicini *et al*. find this sort of explanation problematic (p. 36):

> However, this interpretation seems insufficient to explain why agents polarize in our simulation. The interpretation assumes a very detailed process for how the agents treat and generate evidence, which is not captured by the mechanics of the model. In the model the agents just receive some information and update their degrees of belief and degrees of trust accordingly. This setup means that the model is compatible with various different interpretations for how to understand this behavior.

However, as demonstrated above in connection with the discussion of Jern *et al.*, it follows directly from the Laputa updating table that agents will polarize if they have qualitatively different prior beliefs concerning p and qualitatively different trust assignments. Following Jern *et al.* (2014) and Kelly (2008), these differences in background beliefs together with the underlying Bayesian machinery fully explain how polarization can arise in models like Laputa.

Note that the only thing that is needed to explain how polarization can occur in Laputa in the above "Jern-style" example are the rules for updating credences in the Laputa table; the rules for updating trust play no role in the explanation. This is, of course, completely in accord with the previously mentioned finding by Pallavicini *et al.* that polarization takes place in simulations even when trust-updating is turned off.

Having, clearly incorrectly, rejected this kind of explanation of polarization in Laputa, what do Pallavicini *et al*. propose in its stead? Their own analysis is that Laputa lacks a certain feature that they think is required of a full model of epistemic rationality, namely a mechanism that takes "higher-order evidence" into account. In their view, this is what explains polarization in Laputa. The idea is that two agents who see each other as epistemic equals, in terms of diligence, carefulness and the like, but disagree regarding a proposition p following communication should not only update by adjusting their credence in p and adjusting their trust in the source, as is the case in the Laputa model as it stands; they should also downgrade their trust in their own abilities to inquire properly.

In Laputa, this would mean that disagreement should lead not only to lower trust in one's peer, but also to lower trust in one's own inquiry. Pallavicini *et al.* are silent on how, exactly, to implement this proposal for a revised updating rule in Laputa.

Taking this proposal seriously would have beneficial effects in simulations, Pallavicini *et al.* think (p. 37):

> Now consider the implications of this for the simulation. Can the Bayesian agents in our simulations represent and process higher-order evidence in the way suggested by the above cases? The answer is 'yes' when it comes to information from inquiry, but 'no' when it comes to information from communication. Since the vast majority of information in the simulation comes from communication, this partial Bayesian agent blindness towards higher-order evidence might be quite significant for explaining why they polarize to the surprising extent that they do, and why this polarization is much stronger than what we would expect to see among ordinary epistemically well-functioning human beings.

We note that the ambition has been lowered from explaining polarization *per se* to explaining the surprising extent of polarization. At any rate, Pallavicini *et al.* clarify their view as follows (p. 39):

> Here we have a hypothesis about why ideally rational Bayesian agents in the simulation behave so surprisingly. Agents are responsive to first-order evidence in communication for or against p, but the fail to treat the fact of disagreement as higher-order evidence and fail to adjust their first-order beliefs in their own abilities [i.e. their inquiry trust] accordingly. If they did, we might speculate, they would tend not to be as confident in their ever more extreme views as they are. Moreover, if we assume that fully rational epistemic agents should be responsive to higher-order evidence, then these Bayesian agents are not fully rational. It is not that they are irrational, rather a Bayesian agent only constitutes a partial model of full rationality.

My first point is that the cause of polarization in Laputa, or its extent, is not that the model lacks a mechanism for handling "higher-order evidence" along the lines suggested by Pallavicini *et al*. One way to see this is to observe that polarization, as we noted in connection with Olsson's 2013 study, occurs even if we turn off inquiry altogether. We recall that Olsson's examples concerned a "closed-room debate" in which inquiry chance was set to 0, and yet, as he observed, polarization occurred, indeed to a very considerable extent. Why is this a relevant observation here? The reason is that an updating rule of the kind proposed by Pallavicini *et al.* according to which communication in the network should affect not only credences in p and social trust among agents, but also the receiving agent's inquiry trust, can obviously have any effect only if agents actually engage in inquiry. If they don't, inquiry trust may be updated as much as you like during communication; it won't have any effect on what transpires in the network. In particular, it won't have any effect on whether or not, or the extent to which, agents update their credence in p and, as an effect thereof, polarize.

It is still possible that a revised rule like the one suggested by Pallavicini *et al.* could affect polarization if inquiry is turned on (that is, if inquiry chance is set to a non-zero value). Even so, because, as Pallavinici *et al.* note, "the vast majority of information in the simulation comes from communication" (p. 37), it is unlikely that a rule that differs from the current updating rule for communication only in the effect it has on inquiry should have a major effect on simulation results. In the absence of a formally precise specification of the rule, we cannot know for sure, however.

To clarify, even though I think the study by Pallavicini and her co-authors fails to identify the root cause of polarization in Laputa, their investigations into the impact of various factors on the extent of polarization, and many other insightful observations, add greatly to our understanding of the model.

4. Polarization and epistemic value

The net effect of the simulations carried out by Pallavicini *et al.* is, in fact, a stronger argument for the rationality of polarization. Not only is polarization compatible with Bayesian updating; polarization is, as their study amply demonstrates, omnipresent in social networks governed by such

updating. Now, besides Bayesian updating there is a further standard of epistemic rationality that is relevant in connection with polarization, namely, whether or not deliberation can increase epistemic value. More precisely, can deliberation that lead to higher epistemic value for the group leave that group seriously divided on the issue at hand and, if so, under what conditions does this transpire? This issue is not studied in the article by Pallavicini and her co-workers. Of course, "epistemic value" may mean different things. I will explain the account I favor in a moment.

In shedding light on this matter, let us return to the simple case of two agents, let us call them John and Mary, with opposing prior beliefs who both strengthen their beliefs after receiving the same information from the same source (Figure 2). Let us assume that the source says that p is true. Both John and Mary receive this message and no other information. Mary is initially inclined to believe that p is true, whereas John is initially inclined to believe that not-p is true. Both John and Mary are unsure about the reliability of the source: it is probably no better than tossing a coin but it may also be somewhat reliable or somewhat biased.

Formally:

- The source's degree of belief in p = 1, its certainty threshold is below that value (0.72), and communication chance is set to 1

- Mary's degree of belief in p = 0.75

- John's degree of belief in p = 0.25

- Inquiry chance = 0 for both John and Mary

- Both John and Mary have a trust function corresponding to a normal distribution with mean (expected trust) 0.5 and standard deviation 0.1

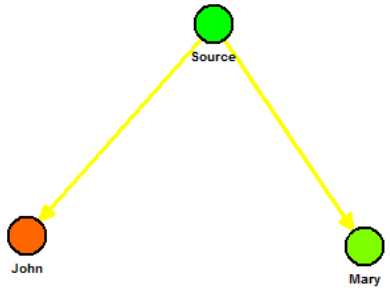

Figure 2: Network of John and Mary listening to the same source.

First, a few observations about polarization. If we run this network in Laputa single network mode, nothing happens at time 0. Polarization kicks in only at time 2, after the source has repeated that p. If, as in this case, both agents have an expected trust of 0.5 in the source, believing it to be no better than chance, two rounds are required to obtain belief polarization

Let us now adjust the situation, so that John is initially inclined to think that the source is probably biased, and Mary has an inclination in the opposite direct. Specifically, John and Mary both have a trust function corresponding to a normal distribution with standard deviation 0.1, but John's trust function has the mean 0.46 and Mary's trust function has the mean 0.54. Then we obtain belief polarization after just one round. This confirms what we already knew: where agents have qualitatively different views about the trustworthiness of a source, the same evidence can prompt them to update their degrees of belief in opposite directions, and this happens even when no inquiry is taking place.[6]

[6] It might be argued that, since the agents assign different trust values to the same source, one of them must be wrong. Being wrong about the trustworthiness of a source is, one might think, a sign of irrationality. For instance, if John assigns a lot of trust to the reports of a creationist, he is irrational. If so, the above simulation does not establish that *rational* agents can update their beliefs in opposite direction. Against this one could object that rationality is a matter of internal coherence as opposed to one's connection to the world. On this picture, even an ideally rational agent may be utterly mistaken in his beliefs. Similarly, according to classic Bayesianism, rationality does not dictate what initial beliefs an agent should have (including beliefs about the reliability of sources) so long as they are coherent in the sense of satisfying the Kolmogorov axioms. The present work is situated in this influential tradition.

To return to the present issue, it remains to be investigated whether polarization is rational in the further sense of being the outcome of a practice that has positive epistemic value. We will assume that one proposition, p, is true and, following the authoritative account in Goldman (1999), focus on epistemic value, or E-value for short, in the sense of "veritistic value". The main intuition is that the closer an inquirer's degree of belief in a true proposition p is to 1, the better it is. Thus having a credence of 0.9 in p is better than having a credence of 0.8 in p, for a true proposition p. The veritistic value of an inquirer's degree of belief in the true proposition p can be identified with that degree of belief. For example, if an inquirer assigns credence 0.6 to the true proposition p, then the veritistic value of that assignment is simply 0.6.

The E-value of a network state can be defined as the average degree of belief in the true proposition p among the agents in the network. Thus, a network in which every agent has degree of belief 1 in the truth has E-value 1 (maximum), and a network in which every agent has degree of belief 0 in the truth has E-value 0 (minimum). We are, above all, interested in whether a given epistemic practice of deliberation, as defined by the initial constraints on the network parameters, raises or lowers epistemic value. For this purpose, we define E-value ∇ to be the final E-value (after the simulation) minus the initial E-value (before the simulation) of the network. Thus, a positive E-value ∇ means that agents in the network, on average, have come closer to (full belief in) the truth as the effect of engaging in inquiry or communication. A negative E-value ∇ means that agents, on average, have move farther away from (full belief in) the truth.

Let us consider John and Mary again. We will study three cases, differing in how, exactly, John and Mary assign initial credences to p.

Case 1: John and Mary are equally far from 0.5 in their belief. To be specific, we assume that John assigns credence 0.25 to p and Mary assigns credence 0.75 to p. Simulation reveals that in this case John and Mary will polarize (Figure 3). After 20 time steps John almost fully believes not-p, whereas Mary almost fully believes p. Meanwhile, E-value does not change at all. In particular, there is no

increase in E-value resulting from John and Mary engaging with the source. It is easy to see why: since John and Mary are initially equally far from 0.5 in their initial credences, and John distrusts the source as much as Mary trusts it, their credences will move apart symmetrically from 0.5, meaning that the E-value (average credence in p, the true proposition) will remain 0.5.

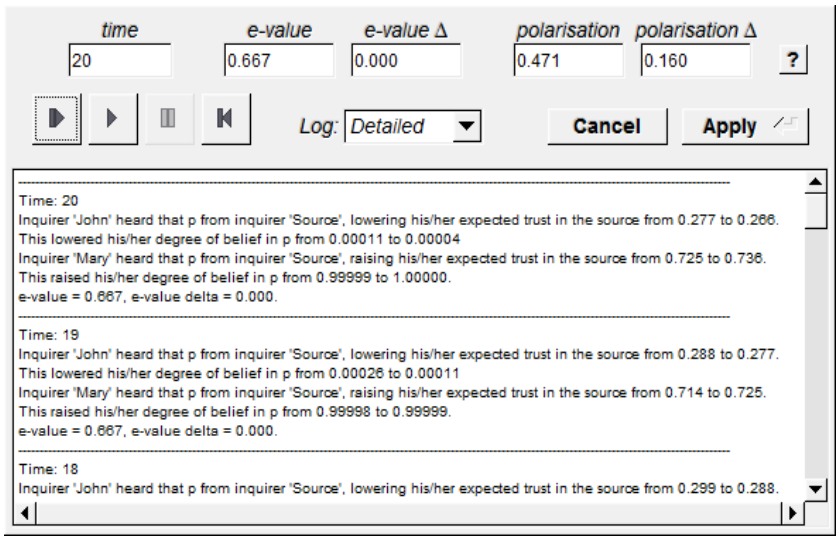

Figure 3: Output of Laputa in Case 1.

Case 2: Mary is farther away from 0.5 than John is. To be specific, we assume that John, as before, assigns credence 0.25 to p but that Mary now assigns credence 0.90 to p. This assignment leads to polarization as well (Figure 4), but this time there is a change in E-value, but the change is negative (-0.05, to be precise). This means that John and Mary, as a collective, have moved farther away from (full belief in) the truth. Again, the explanation is straightforward. Given the circumstances, the average credence in p is initially 0.575. After 20 simulations rounds, however, since John and Mary have reached opposite poles, the average credence in p is 0.5. So there has been a decrease in E-value.

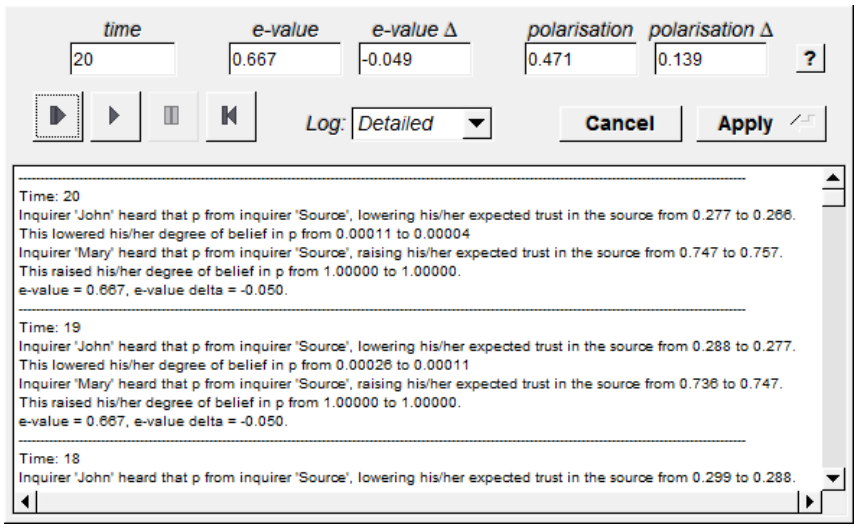

Figure 4: Output of Laputa in Case 2.

Case 3: Finally, we consider a case in which John is farther way from 0.5 than Mary is. As in the first case, Mary assigns credence 0.75 to p but now John assigns a mere 0.1 credence to p. In this case, we do not only get polarization, but also – lo and behold – a positive E-value (Figure 5)! The reason why this is so should be clear by now: the average credence in p was initially slightly below 0.5, or 0.425 to be precise. After 20 rounds of activity, John and Mary have reached opposite poles, and so the average credence is 0.5, i.e. slightly higher than it was initially.

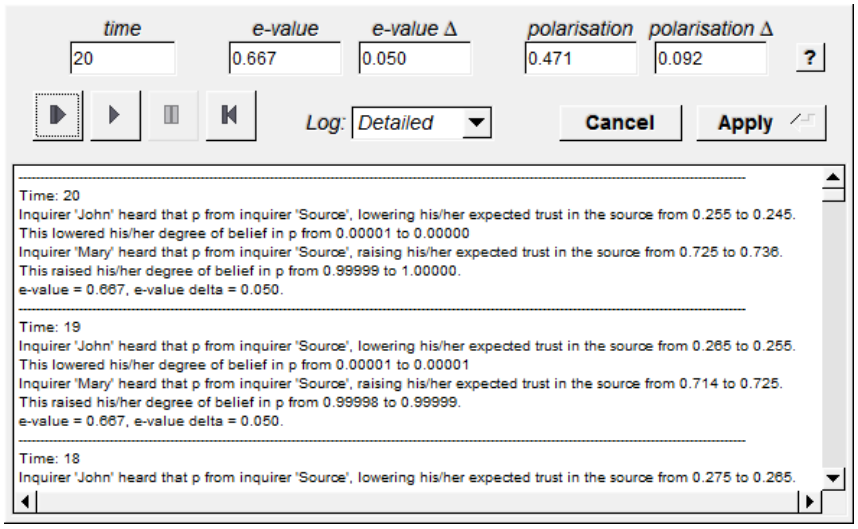

Figure 5: Output of Laputa in Case 3.

What I have done so far is giving an “existence proof”: there exist scenarios in Laputa in which agents that collectively benefit from increased veritistic value polarize. This shows that collective rationality in the sense of increased average veritistic value in the group is compatible with polarization. A stronger claim would be that polarization is rational in the sense that it arises from a practice of social inquiry that increases average veritistic value in the long run. In fact, it is quite common in Laputa that agents that benefit, collectively, from increased veritistic value also become increasingly polarized. A typical case is depicted in Figure 6.

Figure 6: Typical result of simulation in batch mode with increase in epistemic value as well as polarization. E-value has increased by 0.1190 and polarization has increased by 0.2570.

The upshot is that polarization is rational not only in the sense that it emerges from Bayesian updating in the cases under consideration; polarization is also sometimes rational in the sense that a situation in which agents polarize can be one in which the collective benefits from positive epistemic value, a situation that commonly occurs in Laputa simulations, although assessing the precise extent to which this happens would require a more detailed study.

5. Conclusion

Pallavicini, Hallsson and Kappel (2018) is an important study into the robustness of polarization in the Bayesian Laputa model. The authors extend the results reported in Olsson (2013) quite considerably, showing that groups of Bayesian agents polarize under a surprisingly broad range of circumstances. However, rather than concluding that polarization may be rational, they argue that the results are, in the end, explained by an alleged failure of Laputa to capture rationality in its full sense. What, in their view, is not accounted for in Laputa is the role of higher-order evidence in cases in which network peers disagree. This lack is what ultimately explains the fact that agents polarize, they think. As a remedy, they propose a revised updating mechanisms, though without providing a detailed rule, let alone demonstrating analytically or by computer simulation that their proposal would prevent groups from polarizing.

I showed that incorporating higher-order evidence in the manner proposed by Pallavicini *et al.* in fact fails to block polarization in Laputa. Instead, a closer examination revealed that what drives polarization is simply Laputa rules for updating credences on the basis of the expectedness of a message in conjunction with a recognition that the source might be biased, i.e. systematically giving false information. A criticism of the rationality of polarization in Laputa style frameworks would need to address the rationality of expectation-based updating or the rationality of countenancing the possibility that a source might be biased – or, possibly, both.[7] Finally, I demonstrated, by means of simulations, that polarization is rational in a further sense: epistemic practices that lead to increases in epistemic value at the group level can be, and in practice often are, associated with increased polarization, if epistemic value is construed, following Alvin Goldman (1999), as (average) veritistic value. Pace Pallavicini *et al.*, the upshot of all this is a strengthened case for the rationality of polarization.

[7] See Hahn, Merdes and von Sydow (2018) for an extensive and insightful discussion of expectation-based credence updating and its role in reasoning.

References

Angere, S., and Olsson, E. J. (2017). Publish late, publish rarely! Network density and group performance in scientific communication. In T. Boyer, C. Mayo-Wilson, & M. Weisberg (Eds.), *Scientific collaboration and collective knowledge*, Oxford University Press: 34-62.

Angere, S., Olsson, E. J., and Genot, E. (2015). Inquiry and deliberation in judicial systems: The problem of jury size. In C. Baskent (Ed.), *Perspectives on interrogative models of inquiry: Developments in inquiry and questions*, Springer: 35-56.

Collins, P. J., Hahn, U., von Gerber, Y. & Olsson, E. J. (2018). The Bi-directional Relationship Between Source Characteristics and Message Content, *Frontiers in Psychology*, 9. URL: https://doi.org/10.3389/fpsyg.2018.00018

Cook, J., and Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks, *Topics in Cognitive Science*, 8(1): 160–179.

Hahn, U., Hansen, J. U., and Olsson, E. J. (2018). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent, *Synthese*: 1-31. URL: https://link.springer.com/article/10.1007/s11229-018-01936-6

Easwaran, K., Fenton-glynn, L., Hitchcock, C., and Velasco, J. D. (2016). Updating on the credences of others: Disagreement, agreement and synergy, *Philosophers' Imprint*, 16: 1–39.

Goldman, A. I. (1999). Knowledge in a social world. New York: Oxford University Press.

Hahn, U., Merdes, C., and von Sydow, M. (2018). How Good Is Your Evidence and How Would You Know? *Topics in Cognitive Science*, 10: 660–678

Jern, A., Chang, K.-m. K., and Kemp, C. (2014). Belief polarization is not always irrational, *Psychological Review,* 121(2): 206-224.

Kelly, T. (2008). Disagreement, dogmatism, and belief polarization, *Journal of Philosophy*, 105(10): 611-633.

Olsson, E. J. (2011). A simulation approach to veritistic social epistemology, *Episteme*, 8(2): 127-143.

Olsson, E. J. (2013). A Bayesian simulation model of group deliberation and polarization. In Zenker, F. (Ed.), *Bayesian argumentation: The practical side of probability*, Synthese Library, New York: Springer: 113-134.

Olsson, E. J. (2018). A diachronic perspective on peer disagreement in veritistic social epistemology. *Synthese*: 1-19. URL: https://link.springer.com/article/10.1007%2Fs11229-018-01935-7

Olsson, E. J., and Vallinder, A. (2013). Norms of assertion and communication in social networks, *Synthese*, 190: 1437-1454.

Pallavicini, J. Hallsson, B., and Kappel, K. (2018). Polarization in groups of Bayesian agents, *Synthese*: 1-55. URL: https://link.springer.com/article/10.1007/s11229-018-01978-w

Vallinder, A., and Olsson, E. J. (2013a). Do computer simulations support the argument from disagreement? *Synthese*, 190(8): 1437-1454.

Vallinder, A., and Olsson, E. J. (2013b). Trust and the value of overconfidence: a Bayesian perspective on social network communication, *Synthese*, 191(9): 1991–200.