Synthese
DOI 10.1007/s11229-012-0107-x

Do computer simulations support the Argument from Disagreement?

Aron Vallinder · Erik J. Olsson

Received: 28 December 2011 / Accepted: 29 March 2012

Abstract According to the Argument from Disagreement (AD) widespread and persistent disagreement on ethical issues indicates that our moral opinions are not influenced by moral facts, either because there are no such facts or because there are such facts but they fail to influence our moral opinions. In an innovative paper, Gustafsson and Peterson (Synthese, published online 16 October, 2010) study the argument by means of computer simulation of opinion dynamics, relying on the well-known model of Hegselmann and Krause (J Artif Soc Soc Simul 5(3):1–33, 2002; J Artif Soc Soc Simul 9(3):1–28, 2006). Their simulations indicate that if our moral opinions were influenced at least slightly by moral facts, we would quickly have reached consensus, even if our moral opinions were also affected by additional factors such as false authorities, external political shifts and random processes. Gustafsson and Peterson conclude that since no such consensus has been reached in real life, the simulation gives us increased reason to take seriously the AD. Our main claim in this paper is that these results are not as robust as Gustafsson and Peterson seem to think they are. If we run similar simulations in the alternative Laputa simulation environment developed by Angere and Olsson (Angere, Synthese, forthcoming and Olsson, Episteme 8(2):127–143, 2011) considerably less support for the AD is forthcoming.

Keywords Argument from Disagreement · Computer simulation · Formal epistemology · Bayesianism · Probability · Trust

A. Vallinder · E. J. Olsson (✉)
Department of Philosophy, Lund University, Kungshuset, 222 22 Lund, Sweden
e-mail: erik_j.olsson@fil.lu.se

Published online: 04 May 2012

1 Introduction

Many moral philosophers have defended a form of moral realism according to which there are objective moral facts that we can come to know (e.g. Boyd 1988; Brink 1989; Shafer-Landau 2003). The Argument from Disagreement (AD) seeks to show that there are no objective and epistemically efficacious moral facts, i.e., moral truths that affect our moral opinions. If objective and epistemically efficacious moral facts were to exist, then there would be no more interpersonal and intercultural disagreement in ethics than in, say, chemistry or mathematics. However, there is in fact much more disagreement in ethics than in other areas, and therefore advocates of the AD conclude that either no such facts exist or else, if they exist, they do not influence our moral opinions (e.g. Mackie 1977).

In their 2010 paper, Gustafsson and Peterson (G&P) identify two challenges for proponents of AD. First, a proponent has to offer support for the claim that variations in moral opinions are better explained by the moral non-realist hypothesis than by its realist rival. Why is it more reasonable to think that people would be less inclined to disagree if moral facts were to exist, especially given that moral opinions are likely to be affected by other factors as well, such as moral authorities? Second, a proponent of AD has to explain why there are no moral facts given that we actually agree on many moral issues. G&P mention as an example the wrongness of "torturing innocent children for fun" (p. 2).[1] The explanans seems to be not that we disagree on all moral issues but rather that we disagree on some moral issues, such as abortion and capital punishment, while we agree on many others.

G&P claim, in their paper, that these challenges can be met if we agree to adopt the methodology of computer simulations. According to G&P, their computer simulations offer a detailed explanation of why the existence of moral facts would lead to consensus, and why we do actually agree on some moral issues but not on others. They conclude that the net effect is to strengthen our reasons for taking AD seriously. Their simulations are based on the work of Hegselmann and Krause (2002, 2006). In this paper, we will challenge G&P's claim that their results are "very robust" (p. 3) by showing that another picture altogether emerges if, instead of the H&K model, we employ the alternative simulation environment Laputa (Angere forthcoming, Olsson 2011).

2 Gustafsson and Peterson's argument

G&P rely for their simulations on a Hegselmann and Krause's (2006) paper which assesses the chances for the truth to be found and broadly accepted under various conditions. H&K study this both mathematically and by means of computer simulations using an iterative procedure. Fortunately, we need not go into much of the mathematical details. It suffices, for our philosophical purposes, to get a reasonably firm grasp of the main ideas and the results that follow. One assumption is that each individual starts out with a certain opinion that is taken to be arbitrarily chosen. For

[1] All page references to G&P's article refer to the prepublished electronic version of their paper.

definiteness, we may think of it as a real number between zero and one representing some parameter that is being assessed. H&K thought of this parameter in non-moral terms, e.g. the weight of some physical object. The opinions are then assumed to be made public for all to see upon which each individual updates his or her opinion based on the opinions of the others. At this point it is assumed that a given individual takes into account only the other opinions that are, from her point of view, not too exotic.[2] In other words, a given individual bases her new opinion on (1) her own old opinion; and (2) those opinions that are within a certain distance ϵ from her own old opinion, where the number ϵ is referred to as the "confidence level." To be specific, the individual suspends judgment, as it were, between these views by taking on a new view corresponding to their average value.

An intriguing feature of H&K's model is what they call "truth attraction". H&K assume that there is a true opinion, T, in the space of possible opinions that may be capable of "attracting" individuals in the sense that they have a tendency to approach it, perhaps because they are using rational argumentation, reasonable thinking, sound experimental procedures, etc. Thus, H&K choose to abstract from the qualitative nature of the methods of inquiry, focusing instead on their quantitative reliability. There is in this model an objective as well as a social component determining the opinion of a given individual. Objectivity is, as we just saw, captured by the degree to which a given individual is attracted to the truth. The social component amounts to specifying how much weight the individual assigns to the opinions of her "peers." It is helpful to take a quick look at the main equation to which these considerations give rise:

$$x_i(t+1) = \alpha_i T + (1-\alpha_i) f_i(x(t)), \quad 1 \leq i \leq n,$$

where $x_i(t+1)$ is the new opinion of the ith individual, α_i the degree to which that individual is attracted to the truth, and $1-\alpha_i$ the degree to which her opinion is socially determined. Setting $\alpha_i > 0$ means that the ith individual is to some extent attracted to the truth. Setting $\alpha_i = 0$ means that there is no direct connect between her opinion and the truth but that her new opinion is rather the mere product of her own previous one and the opinions of her peers.

H&K proceed to conduct their computer simulations with amusing and occasionally unexpected results. Let us start with the case where none of the individuals is attracted to the truth, and where the confidence level of each individual is fairly low so that only a few other opinions are taken into account. What will happen when the main equation is used repeatedly to update the opinions of the individuals is that a number of clusters are eventually formed consisting of individuals sharing the same opinion, each cluster being out of reach of, and hence incapable of influencing, the others. The convergence of one of those clusters on the truth will of course be a purely random affair. H&K (2006, p. 6) describe the result of their simulations as "an eternal plurality of divergent views" (Fig. 1).

If, by contrast, the confidence level for the individuals is sufficiently raised, they will start averaging over the views of their peers. This will eventually lead to a situation

[2] This assumption may not be realistic. See Zollman (forthcoming) and Martini (2012).

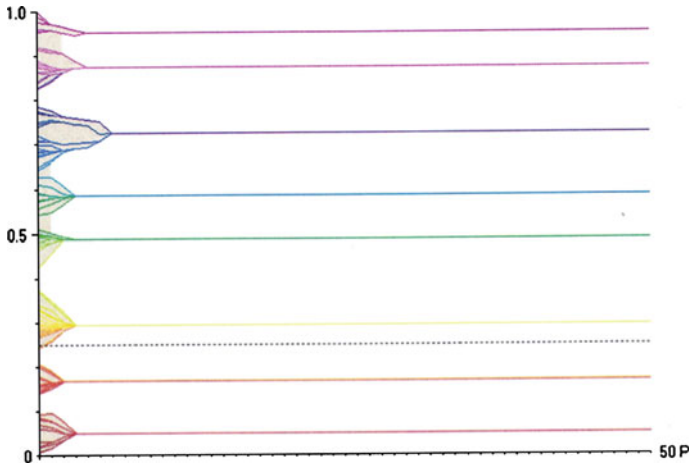

Fig. 1 Hundred agents, $T = 0.25$, random start distribution, confidence level = 0.05, $\alpha_i = 0$ for all agents i

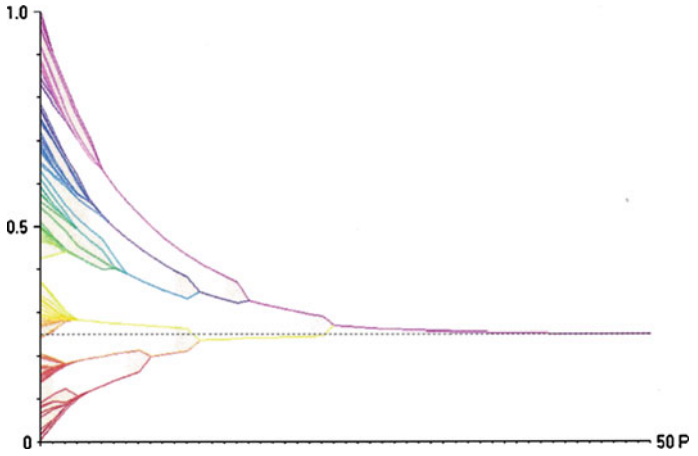

Fig. 2 Hundred agents, $T = 0.25$, random start distribution, confidence level = 0.05, $\alpha_i = 0.1$ for all agents i

where everyone converges on a particular opinion. This opinion will be false except in the exceptional case where it happens, by chance, to be true.

Suppose we assume, to take the other extreme, that all individuals are truth seekers in the sense of being attracted to the truth, if ever so slightly, and that circumstances are otherwise as in the scenario with a smaller confidence level. Then what will happen is that there will be the same initial tendency for clusters to be formed, but these clusters will now, as the updating process continues, gradually approach the truth (and hence also each other), though they may not get to there in finite time (Fig. 2).

As H&K show, this result is essentially preserved if only some of the inquirers have a positive degree of truth attraction, provided that the confidence interval is large enough. Thus H&K provide an example where only fifty percent of the individuals are

(slightly) attracted to the truth. The result after some initial clustering is that the social exchange process finally leads to a consensus that is at least fairly close to the truth. The inquirers who are not attracted to the truth will, because of their social nature, become indirectly connected to the truth through the information they receive from reliable peers.[3]

How is this relevant to AD? G&P (2010, p. 4) point out that we can think of the parameter being assessed in the H&K model as the "degree of moral praiseworthiness of the action under consideration". Conventionally they take 1 to represent maximum praiseworthiness and 0 maximum condemnation of the action (e.g. abortion). Suppose now that no inquirer is attracted to moral truth, either because there is no moral truth in the first place or because there is a moral truth but it does not affect us. Let us assume, in addition, that the confidence interval is small enough, i.e. that people take into account only the views of those who are of the same or almost the same opinion. In some infected moral issues such as abortion or capital punishment this may not be an unreasonable assumption. Then people should diverge widely in their opinion even after a considerable number of interactions, as in Fig. 1. Suppose, by contrast, that all inquirers are somewhat attracted to moral truth but that other things are equal. Then people should gradually converge on the true moral view, as in Fig. 2. Assume now that as a matter of fact people widely disagree in their moral opinions. It would follow that no inquirer is attracted to moral truth and that any moral realism that supposes otherwise is flawed.

G&P are of course aware of the fact that the H&K model is in many ways simplistic. In real life, there are various disturbing factors that a more complete model could take into account. It is a merit of their inquiry that some of those factors are explicitly attended to. To be specific, they provide an extension of the H&K model where they take into account the possibility of authorities, external shifts and random processes. An authority is conceived of as an influential person or organization, such as the pope or Greenpeace. Authorities obviously influence moral opinion in various ways, quite independently of the truth of the matter (if such there be). External shifts are sudden external shocks to the system, e.g. in the form of radical political shifts. Finally, moral opinion might change as the effect of a random process. G&P argue that the H&K model support AD even if all these additional mechanisms are taken into account, indeed even if they all are operative together.

We will not question the simulation results obtained by G&P. What we will question is their interpretation of those results. In particular, we will show that the results are not stable across different simulation models. We will show this by studying another simulation model, called Laputa, which is arguable the main competitor to the H&K model among models of communication that are, in G&P's words, "truth-sensitive".

[3] For an insightful recent discussion of the H&K model, see Douven and Kelp (2011). See also Olsson (2008).

3 The Laputa model

This section follows the structure of Angere (forthcoming). In Bayesian fashion, the epistemic state of an individual α at time t is given by a *credence function* $C^t_\alpha : \mathcal{L} \to [0, 1]$, where we can take $\mathcal{L}$ to be a classical propositional language. The expression $C^t_\alpha(p) = x$ should be read as "Agent α's credence in proposition p at time t is x." There are two ways for individuals to receive new information: inquiry and communication.

Inquiry is any method of altering a credence function that doesn't base itself on information from someone else in the network. We let $S^{t+}_{i\alpha}$ be the proposition 'α's inquiry gives a positive result at t', $S^{t-}_{i\alpha}$ be the proposition 'α's inquiry gives a negative result at t', and $S_{i\alpha}t =_{\text{df.}} S^{t+}_{i\alpha} \vee S^{t-}_{i\alpha}$ be the proposition 'α's inquiry gives some result at t'. The participants' properties as inquirers are represented by two probabilities: the chance $\mathrm{P}(S^t_{i\alpha})$ that, at any moment t, α receives a result from her inquiries, and the chance $\mathrm{P}(S^t_{i\alpha} \mid S^t_{i\alpha} \wedge p)$ that, when such a result is obtained, it is the right one. We will call $\mathrm{P}(S^t_{i\alpha})\alpha$'s *activity*, and $\mathrm{P}(S^t_{i\alpha} \mid S^t_{i\alpha} \wedge p)$ her *aptitude*. Both activity and aptitude will generally be constant over time.

When it comes to *communication*, we analogously define $S^{t+}_{\beta\alpha}$ as "β sends a positive message to α at t", $S^{t-}_{\beta\alpha}$ as "β sends a negative message to α at t", and $S^t_{\beta\alpha}$ as "β sends some message to α at t". The strength of a link $\beta\alpha$ is represented as a probability $\mathrm{P}(S_{\beta\alpha})$ that β sends some message to α.

The *threshold of assertion* is a value $T_{\beta\alpha} \in [0, 1]$ such that

- if $T_{\beta\alpha} > 0.5$, β sends a positive message to α only if $C_\beta(p) \geq T_{\beta\alpha}$, and a negative message only if $C_\beta(p) \leq 1 - T_{\beta\alpha}$;
- if $T_{\beta\alpha} < 0.5$, β sends a positive message to α only if $C_\beta(p) \leq T_{\beta\alpha}$, and a negative message only if $C_\beta(p) \geq 1 - T_{\beta\alpha}$; and
- if $T_{\beta\alpha} = 0.5$, β sends a positive or a negative message to α independently of what she believes, which is modeled by letting her pick what to say at random.

This means that a link that consists of systematic lying will have a threshold value below 0.5, and a link of truth-telling (as far as the link's source is aware) has a threshold value above 0.5.

Agent α's inquiry i and the other inquirers $\beta, \gamma, \ldots$who can talk to her, constitute her *sources*. We define the *reliability* of α's source σ as

$$R_{\sigma\alpha} =_{\text{df.}} \mathrm{P}(S^+_{\sigma\alpha} \mid S_{\sigma\alpha} \wedge p) = \mathrm{P}(S^-_{\sigma\alpha} \mid S_{\sigma\alpha} \wedge \neg p)$$

From this definition it follows directly that the reliability of α's inquiry is identical to her aptitude. Since the number of possible values for $R_{\sigma\alpha}$ is infinite, we need to represent α's credence as a density function instead of a probability distribution. Thus, for each inquirer α, each source σ, and each time t, we define a function $\tau^t_{\sigma\alpha}$: $[0, 1] \to [0, 1]$, called α's *trust function for σ at t*, such that

$$C^t_\alpha(a \leq R_{\sigma\alpha} \leq b) = \int_a^b \tau^t_{\sigma\alpha}(\rho)\, d\rho$$

for a, b in [0, 1]. $\tau_{\sigma\alpha}(\rho)$ then gives the credence density at ρ, and we can get the actual credence that α has in propositions about the reliability of her sources by integrating the function. We will also have use for the expression $1 - \tau^t_{\sigma\alpha}$, representing α's credence density for propositions about σ *not* being reliable, which we will refer to as $\bar{\tau}^t_{\sigma\alpha}$. Now, an inquirer's credences about chances should influence her credences about the outcomes of these chances. The way to do this is generally known as the principal principle (Lewis 1980). It states that if α knows that the chance that an event e will happen is ρ, then her credence in e should be exactly ρ. This means that in our case, the following (PP) must hold:

$$C^t_\alpha(S^{t+}_{\sigma\alpha} \mid S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge p) = \rho$$
$$C^t_\alpha(S^{t-}_{\sigma\alpha} \mid S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge \neg p) = \rho$$

for all t, i.e. α's credence in σ giving a positive report, given that the source gives any report at all, that σ's reliability is ρ, and that p is actually the case, should be ρ.

We will also use an independence assumption, here referred to as *communication independence* (CI):

$$C^t_\alpha(p \wedge S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho) = C^t_\alpha(p)C^t_\alpha(S^t_{\sigma\alpha})R^t_{\sigma\alpha}(p)$$

(CI) implies that whether σ says anything is independent both of whether p is actually true, and of σ's reliability.

Given (PP) and (CI) we can define the following expression for α's credence in σ's reliability (see Angere forthcoming for the derivation):

$$C^t_\alpha(S^{t+}_{\sigma\alpha} \mid p) = C^t_\alpha(S^t_{\sigma\alpha}) \int_0^1 \rho\tau^t_{\sigma\alpha}(\rho)\, d\rho$$

The integral in this expression is the expected value $\langle\tau^t_{\sigma\alpha}\rangle$ of the trust function $\tau^t_{\sigma\alpha}$, so that the above can be written as

$$C^t_\alpha(S^{t+}_{\sigma\alpha} \mid p) = C^t_\alpha(S^t_{\sigma\alpha})\langle\tau^t_{\sigma\alpha}\rangle$$

Analogously, we get

$$C^t_\alpha(S^{t+}_{\sigma\alpha} \mid \neg p) = C^t_\alpha(S^t_{\sigma\alpha})\langle\bar{\tau}^t_{\sigma\alpha}\rangle$$

Using Bayes' theorem and the theorem of total probability, we can now derive the expressions $C^t_\alpha(p \mid S^{t+}_{\sigma\alpha})$ and $C^t_\alpha(p \mid S^{t-}_{\sigma\alpha})$, the credence an inquirer should place in p at t given that she receives a positive or a negative message, respectively, form a single source σ:

Table 1 Qualitative rules for updating credence

Source trusted?	Is message surprising?		
	No	Neither	Yes
Yes	+	+	−
Neither	0	0	0
No	−	−	+

$$C_\alpha^t(p \mid S_{\sigma\alpha}^{t+}) = \frac{C_\alpha^t(p)\langle\tau_{\sigma\alpha}^t\rangle}{C_\alpha^t(p)\langle\tau_{\sigma\alpha}^t\rangle + C_\alpha^t(\neg p)\langle\bar{\tau}_{\sigma\alpha}^t\rangle}$$

$$C_\alpha^t(p \mid S_{\sigma\alpha}^{t-}) = \frac{C_\alpha^t(p)\langle\bar{\tau}_{\sigma\alpha}^t\rangle}{C_\alpha^t(p)\langle\bar{\tau}_{\sigma\alpha}^t\rangle + C_\alpha^t(\neg p)\langle\tau_{\sigma\alpha}^t\rangle}$$

By the Bayesian requirement of conditionalization, we must have $C_\alpha^{t+1} = C_\alpha^t(p|S_{\sigma\alpha}^{t+1})$, whenever σ is the only source giving information to α at t. Thus these formulae completely determine how α should update her credence in such a case. We say that an individual *trusts* a given source if the inquirer's credence in the reliability of the source is greater than 0.5; *distrusts* the source if it is less than 0.5; and *neither trusts nor distrusts* the source if it is exactly 0.5. We say that a message that p (not-p) was *surprising* to an individual if, prior to receiving the message, her credence in p (not-p) was less than 0.5; *unsurprising* if it was greater than 0.5; and *neither surprising nor unsurprising* if it is exactly 0.5. The qualitative update rules are given in Table 1. A '+' means that the current belief is reinforced (i.e. $C_\alpha^{t+1}(p) > C_\alpha^t(p)$ if $C_\alpha^t(p) > 0.5$, and $C_\alpha^{t+1}(p) < C_\alpha^t(p)$ if $C_\alpha^t(p) < 0.5$.), a '−' that the strength of the belief is weakened, and '0' that her credence is unchanged. See Olsson and Vallinder (forthcoming) for derivations.

When α receives messages from several sources at once the calculations become a bit more complex. Let $\sum_\alpha^t$ be the set of sources from which α receives information at t, and let $m_{\sigma\alpha}^t$ be the message that σ gives to α at t. Conditionalization requires that

$$C_\alpha^{t+1}(p) = C_\alpha^t\left(p \,\middle|\, \bigwedge S_{\sigma\alpha}^t m_{\sigma\alpha}^t\right)$$

where the conjunction runs over all sources σ in $\sum_\alpha^t$. We can simplify this by assuming that inquirers treat their sources as independent. This means that we adopt the following axiom of *source independence* (SI):

$$C_\alpha^t\left(\bigwedge S_{\sigma\alpha}^t m_{\sigma\alpha}^t \,\middle|\, p\right) = \prod C_\alpha^t\left(S_{\sigma\alpha}^t m_{\sigma\alpha}^t \,\middle|\, p\right)$$

According to the original interpretation of Laputa (Angere forthcoming; Olsson 2011), an individual communicating a positive or a negative report, was interpreted as her disclosing her opinion that p or not-p, respectively. However, as Olsson (forthcoming) notes, this makes the assumption of SI implausible. After only a few steps of the simulation, each individual's credence will typically be highly dependent on the messages she has received from others. As a remedy, Olsson proposes that we interpret $S_{\beta\alpha}^{t+}$ as 'β gives an argument *for* p to α at t', and $S_{\beta\alpha}^{t-}$ as 'β gives an argument

Table 2 Qualitative rules for updating trust

Source trusted?	Is message expected?		
	Yes	Neither	No
Yes	+	0	−
Neither	+	0	−
No	+	0	−

against p to α at t. If we further assume that all arguments given are novel and sound, (SI) will be a plausible assumption. This way, Laputa gains support from persuasive arguments theory (PAT). According to this theory, people move further in the direction they are currently leaning towards upon hearing novel persuasive arguments for the position, but not upon hearing arguments they are already familiar with (Isenberg 1986, pp. 1142–1144).

Given (SI), we are now in a position to derive an expression for how α should update her credence when she receives messages from several sources at once:

$$C_\alpha^t\left(p \,\middle|\, \bigwedge S_{\sigma\alpha}^t m_{\sigma\alpha}^t\right) = \frac{C_\alpha(p)\prod C_\alpha^t(S_\sigma^t m_{\sigma\alpha}^t \mid p)}{C_\alpha(p)\prod C_\alpha^t(S_\sigma^t m_{\sigma\alpha}^t \mid p) + C_\alpha(\neg p)\prod C_\alpha^t(S_\sigma^t m_{\sigma\alpha}^t \mid \neg p)}$$

But α also needs to keep track of how much to trust her sources. Given that α's trust function for the source σ is $\tau_{\sigma\alpha}^t$ at t, and that she receives a positive or negative message from σ then, her new trust function $\tau_{\sigma\alpha}^{t+1}$ is given either by

$$\tau_{\sigma\alpha}^{t+1} = \tau_{\sigma\alpha}^t(\rho)\frac{\rho C_\alpha^t(p) + (1-\rho)C_\alpha^t(\neg p)}{\langle\tau_{\sigma\alpha}^t\rangle C_\alpha^t(p) + \langle\bar{\tau}_{\sigma\alpha}^t\rangle C_\alpha^t(\neg p)} \tag{1}$$

or by

$$\tau_{\sigma\alpha}^{t+1} = \tau_{\sigma\alpha}^t(\rho)\frac{\rho C_\alpha^t(\neg p) + (1-\rho)C_\alpha^t(p)}{\langle\tau_{\sigma\alpha}^t\rangle C_\alpha^t(\neg p) + \langle\bar{\tau}_{\sigma\alpha}^t\rangle C_\alpha^t(p)} \tag{2}$$

See Angere (forthcoming) for derivations. The qualitative rules for updating trust are given in Table 2. Olsson and Vallinder (forthcoming) offer derivations.

A social network is represented as a directed graph, with nodes representing inquirers and links representing communication channels.

For each inquirer, a number of parameters can be set. The *initial degree of belief* is the inquirer's initial credence in p. *Inquiry accuracy* is her reliability in inquiry, $R_{i\alpha}$. The *inquiry chance* is the probability that the inquirer will conduct an inquiry in a given step of the simulation. The *inquiry trust* is the inquirer's degree of "self-trust", i.e. $\tau_{i\alpha}$. There are also a number of parameters for every link. The *listen trust* is the recipient's trust in the sender, i.e. $\tau_{\beta\alpha}$. The *threshold of assertion* is the value of $T_{\beta\alpha}$ for each inquirer α. Whether a message will be submitted depends on the *communication chance*.

When running Laputa, the program runs through a series of steps, each step representing a chance for an inquirer to conduct an inquiry, to communicate (send, listen)

with the other inquirers to which she is connected, or to do both. After each such step, Laputa computes new values for $C_\alpha(p)$ and $\tau_{\sigma\alpha}$ for every inquirer α, and every source σ.

Laputa also allows the user to set network features at a statistical level, by specifying how the parameters should be distributed over a network. In addition to the features we have already discussed, the *link chance* is the probability that there will be a link between any two inquirers in the network. The program can then randomly create large networks with those features, let them evolve and collect statistics. This is done in the "batch window".

4 Results in Laputa

We can study moral realism in Laputa by assuming that p is a moral proposition upon the truth or correctness of which there is considerable disagreement, e.g. "Abortion is morally permissible". A simulation in Laputa means a series of steps at each of which the inquirers may perform their own inquiry (i.e. consult their moral faculty) and/or communicate their moral view to their network peers. We now assume that the various parameters of the model have been set to values for which a claim can be made that they represent the normal case of communication. We can represent the absence of causally efficacious moral facts by assuming that the inquirer's moral faculty is completely unreliable with no connection whatsoever to moral truth if such there be (Fig. 3).

Some statistics for this case is seen in Fig. 4. Figure 4 may suggest that assuming inquirers to be unconnected to moral truth always gives rise to a split society, one part believing p and the other part believing not-p. However, this does not follow. The upper part of Fig. 4 only shows the statistical distribution of degrees of beliefs after Laputa has considered 100,000 societies. On the average, there will about as many inquirers in the p camp as in the not-p camp. This does not mean, however, that every society will be divided. On the contrary, it is compatible with our assumptions that some societies converge on the truth. During the simulation process, the final state of several societies is shown. Most of these will become divided if no inquirer is connected to the truth. This means that if inquirers are unconnected to moral truth the likely, but not inevitable, result is a society that is split on the issue in question. In other words, the "average" society will be one that exhibits a 50–50 split on the issue. As seen in the lower diagram of Fig. 4, there is on average no progression towards the truth: the average (degree of) belief over time stays the same.

We can represent the existence of causally efficacious moral facts by assuming that all inquirers are reliably connected to the truth. We assume, conventionally, that p is in fact true. We now run the same simulations as before, except that we set the inquirer accuracy parameter to .75. The result is shown in Fig. 5. Unlike the results in Fig. 4, in this case there is indeed progression towards the truth on average, even if that progression is modest, as seen in the lower diagram of Fig. 5.

As the statistics show, there is in Laputa no guarantee that reliable inquirers converge on the truth. A fair portion of the inquirers in fact end up believing what is false, even if they are reasonably reliable. Still societies consisting of reliable inquirers are

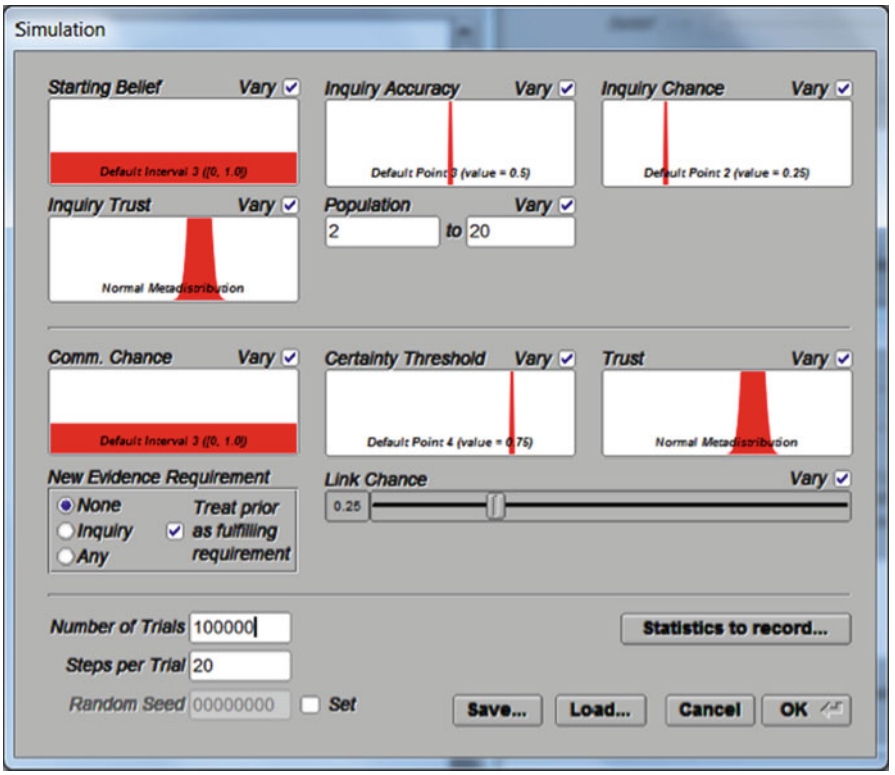

Fig. 3 Settings for the case of inquirers unconnected to moral truth (if such there be)

more likely to end up believing the truth than societies consisting of entirely unreliable inquirers, and the former societies are also more likely to converge on the truth than the latter. The "average" reliable society will be one where the majority believes the truth but where a substantial minority believes the falsity.

What follows from this? Suppose there are only two potential explanations available: (a) that people are not influenced by moral truth or (b) that they are (positively) influenced by moral truth. Suppose we have a 50–50 split for p contra not-p. Then the most plausible explanation in Laputa is that the inquirers are not influenced by moral truth (if such there be). Suppose, by contrast, that there is, say, a 70–30 split for p contra not-p. That would suggest that people are in some degree influenced by moral truth and that p is true. It all comes down to how deep the divide is. Supposedly, most deep moral conflicts are not properly described as 50–50 splits but rather as something like 70–30 splits. For instance, most nations have rejected capital punishment suggesting that most people judge capital punishment to be morally wrong, although there is a persistent and substantial minority that believes otherwise. If this is true, then most moral conflicts are better explained by reference to a moral truth capable of influencing our opinions rather than in terms of the lack of such influence altogether.

This would indeed follow were (a) the only explanation of a 50–50 split. Due to the sophistication of Laputa, there are many other situations, some more plausible

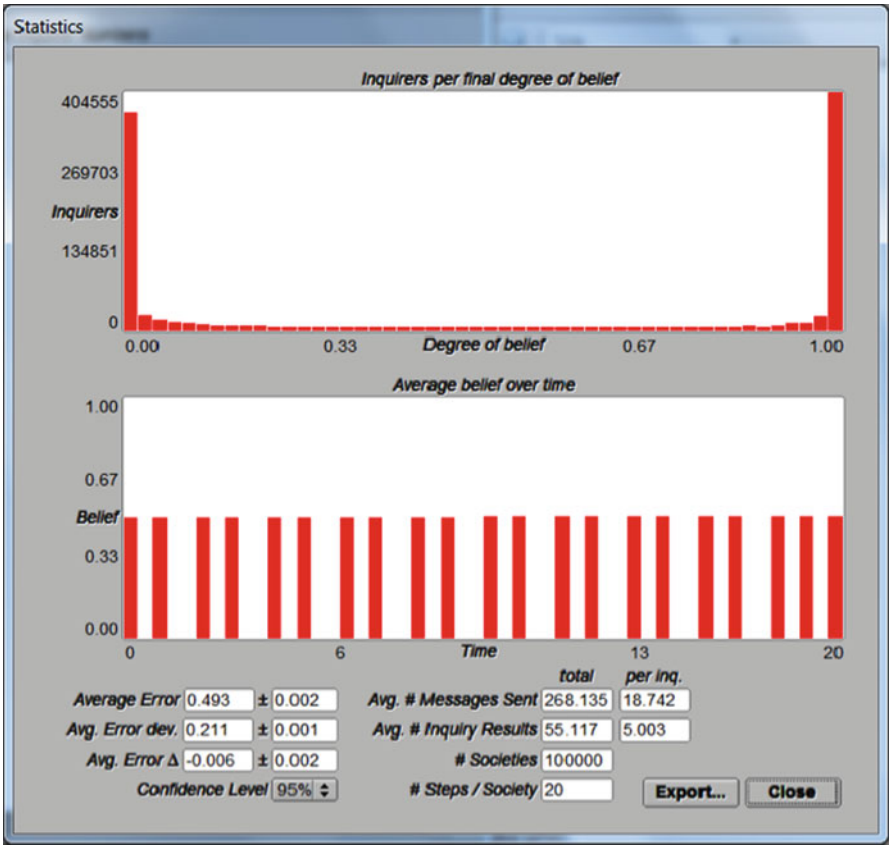

Fig. 4 Statistics for inquirers unconnected to moral truth (if such there be)

than others, that also give rise statistically to what is essentially a 50–50 split. One such situation is when inquirers are in fact reliable, but social trust is set below 0.5. Apart from this, everything is the same as in Fig. 3. However, due to the fact that they seriously distrust each other at the outset, indeed take each other to be systematic liars, they end up deeply divided. The result is the same as that presented in Fig. 4.

This phenomenon demonstrates the complexity of Laputa which arises because, unlike the H&K model, the Laputa model explicitly models and dynamically updates trust. The upshot is that if all we know is that people are deeply divided on some moral issue, there are—in Laputa—at least two possible explanations of this fact. One is the old one: that the opinions are not influence by moral truth. Another explanation, as we just saw, is that people are in fact influenced by moral truth. However, because they don't trust each other at the beginning, the dynamics of trust gradually lead people to adopt contrary opinions.

We will now try to explain how this happens step by step. We will close this section by studying an example of how a society consisting of serious (truth-telling) inquirers

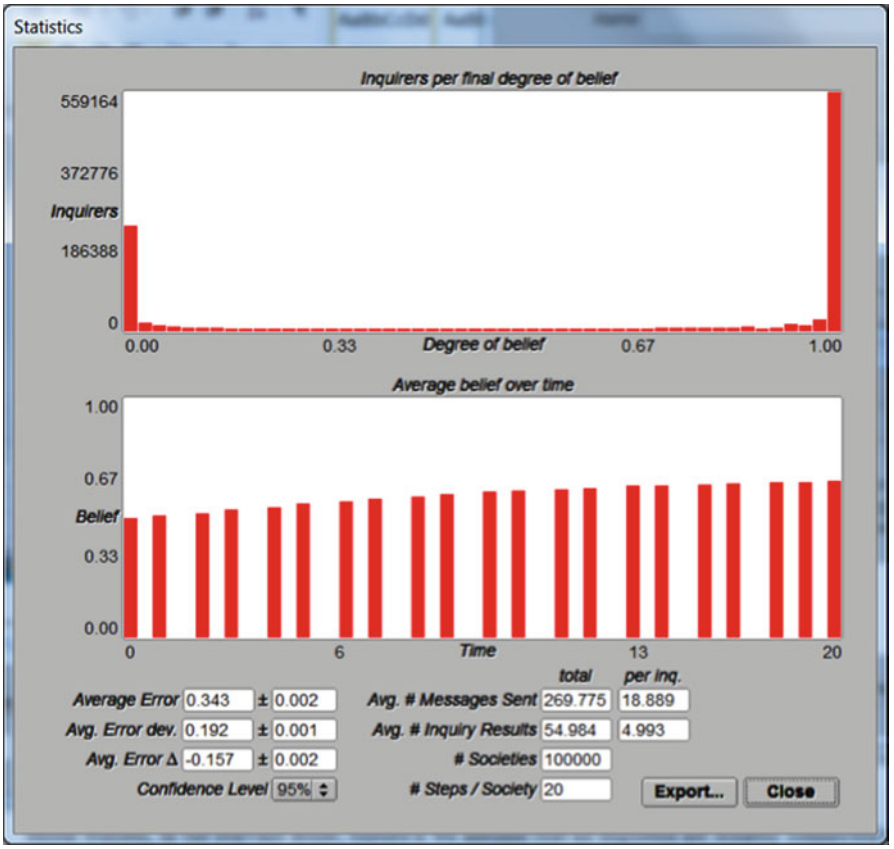

Fig. 5 Statistics for inquirers connected to moral truth

initially inclined to believe the same thing can still end up divided on the issue.[4] We will study a simple society consisting of only two inquirers: Inquirer 1 (Inq 1) and Inquirer 2 (Inq 2). We set the listen chance for Inquirer 1 to 0.94 and for Inquirer 2 to 0.88, and the threshold for both inquirers to 0.58. We choose a normally distributed trust function for both inquirers with expected value 0.38. Figure 6 shows how the inquirers degree of belief (DB) in p, and the expected value of their trust functions, change with time.

We see that after some fluctuations the general trend is that one inquirer will start believing p while the other will start believing not-p.

Laputa allows us to inspect the relevant parameters in a step-wise fashion to see what causes this result, while keeping in mind the qualitative rules for credence and trust updating that we mentioned in Sect. 3. This reveals that the following transpires:

1. Both inquirers initially give arguments for p because their DB in p is above the threshold of assertion.

[4] See also Olsson (forthcoming).

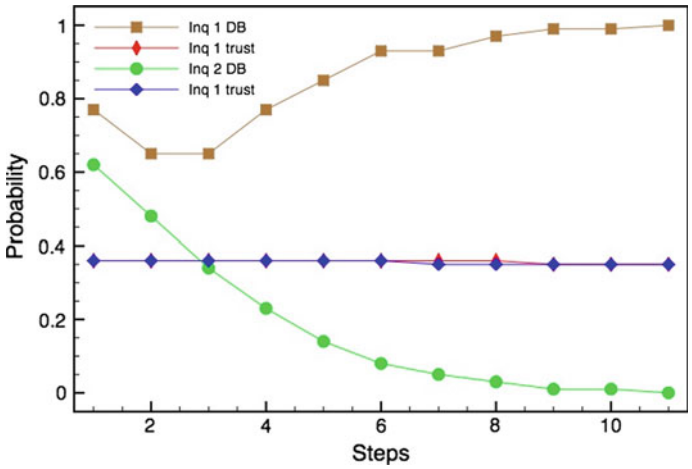

Fig. 6 DB and trust changing over time in a network with two inquirers

2. Since they distrust each other, they will take each other's arguments as evidence for not-p and lower their DB in p.
3. Inquirer 1 still has a DB in p which is above the threshold, and so she gives an argument for p.
4. Given her distrust in Inquirer 1, Inquirer 2 becomes rather confident that not-p, so she gives an argument for not-p.
5. Given her distrust in Inquirer 2, this is taken by Inquirer 1 to be evidence for the opposite, namely p.
6. By the same token, Inquirer 1 will continue to argue for p while Inquirer 2 will continue to argue for not-p, and they will become ever more confident in the conclusions of their arguments.
7. Eventually Inquirer 1 will become certain of p and Inquirer 2 certain of not-p.
8. Meanwhile they will continuously downgrade their degree of trust still further because, as they see it, they repeatedly receive surprising messages from a distrusted source.

We note that while divergence occurs with respect to credence in p, polarization occurs with respect to trust: the inquirers initially distrusted each other and this initial tendency is reinforced as the effect of deliberation.

There is in fact a third possible explanation of a 50–50 split in Laputa which, in a sense, represent the inverse of the explanation just given. In the explanation just given inquirers are de facto reliable but initially distrust each other. In the alternative explanation, the situation is reversed: inquirers are de facto systematically deceived in moral matters and yet they trust each other. That inquirers are systematically deceived means that people take a private moral "signal" to the effect that p (i.e. that abortion is permissible) as evidence for not-p (that abortion is impermissible). It can be shown that this situation also gives rise to a 50–50 split under otherwise the same conditions as before. However, this explanation does not strike the authors as very plausible in itself. What is common for the two alternative explanations of a 50–50 split is that

they involve a society lacking in *social calibration*: people's attitudes towards other people initially fail to adequately reflect the actual trustworthiness or reliability of the latter.

Finally, there are two other ways of obtaining a 50–50 split in Laputa. One is a variation on taking everyone to be a systematic liar. In this case, rather than taking *all* members of the network to be systematic liars, agents will treat roughly half the network in this fashion, and place high trust in the other half. All members are reliable in their inquiries, but the end result will again be a 50–50 split, via the mechanism presented step by step above. This might be more empirically plausible than treating everyone as a liar. On many moral issues, the position a person takes will be correlated with what social group she belongs to.[5] It is not implausible to assume that people place higher trust in members of their own group, and thereby we arrive at this particular simulation setting. However, it is unlikely that people communicate just as much with people from the other group as they do with those in their own. If we accommodate this, the result will no longer be a 50–50 split, but rather a 70–30 or 80–20 split, with the exact numbers depending on the ratio of "inside" vs. "outside" communication. This suggests another possible explanation for a 50–50 split: have two subnetworks that barely communicate with each other. In one subnetwork, agents' inquiries are positively correlated with truth, and in the other they are negatively correlated with truth. Members of the truth-tracking subnetwork will end up believing the truth, and members of the other subnetwork will end up believing the falsity.[6]

5 Is Laputa the better model?

We have seen that the two main "truth-sensitive" models of communication, Laputa and the H&K model, give different verdicts on the AD. A natural next step is to ask which is the better model. Before we do this we will ask under what conditions inquirers converge on the truth in the Laputa model. Interactions in Laputa will typically be highly complex, making analytical methods ill-suited. However, for simpler cases we can prove some convergence results. At a bare minimum, we should expect that an agent who only receive positive messages from a trusted source will converge on the truth. We should also expect that, under the same conditions, the expected value of the agent's trust function, $\langle \tau^t_{\sigma\alpha} \rangle$, will converge to 1 in the limit. And this is indeed so.

Theorem 1 *If* $\langle \tau^t_{\sigma\alpha} \rangle > 0.5$ *for every* t, *and* $C^0_\alpha(p) > 0$, *then* $\lim_{t\to\infty} C^t_\alpha(p) = 1$ *for an individual* α *receiving positive messages from* σ, *and receiving no messages from other sources.*

Proof The conditions place us in the upper leftmost corner of Table 1, where we find a + sign, meaning that the agent's current belief is reinforced. This means that $C^t_\alpha(p), C^{t+1}_\alpha(p), C^{t+2}_\alpha(p), \ldots$ is a strictly monotonically increasing sequence. Since the least upper bound for $C^t_\alpha(p)$ is 1, it follows that the sequence will converge on this value, which is also the true value. □

[5] See e.g. Jelen and Wilcox (2003) for a case study on abortion.

[6] However, this type of disagreement may not be of the persistent kind that is taken to be of special interest in this context.

Theorem 2 *If* $C^t_\alpha(p) > 0.5$ *for every t, and* $\langle \tau^0_{\sigma\alpha} \rangle > 0.5$*, then* $\lim_{t\to\infty} \langle \tau^t_{\sigma\alpha} \rangle = 1$ *for an individual* α *receiving positive messages from* σ*, and receiving no messages from other sources.*

Proof Here, the conditions place us in the upper leftmost corner of Table 2, where we find a + sign, meaning that the expected value of the trust function, $\langle \tau^t_{\sigma\alpha} \rangle$, is increased. So $\langle \tau^t_{\sigma\alpha} \rangle$, $\langle \tau^{t+1}_{\sigma\alpha} \rangle$, $\langle \tau^{t+2}_{\sigma\alpha} \rangle$, ... form a strictly monotonically increasing sequence. Again, the least upper bound is 1, so $\langle \tau^t_{\sigma\alpha} \rangle$ will converge on this value. □

There are a number of differences between Laputa and the H&K model. First, in Laputa there are only two potential answers to the inquirers' question: p or not-p. In the H&K model, by contrast, there is a continuum of potential answers corresponding to the interval between 0 and 1. Which of these is more suitable for representing the present question of whether abortion is permissible or not? If one adopts the H&K model, it follows that moral opinions can be represented by real numbers, and that one could hold that abortion is permissible to some degree x in the unit interval. But typically, we would rather say that abortion is permissible or impermissible under such or such circumstances. For this reason, it is not clear whether the H&K model is a good representation of the question. On the other hand, if the statement "Abortion is morally permissible" is fleshed out in a non-ambiguous way (e.g. as "In some circumstances, abortion is morally permissible"), the binary framework of Laputa can be applied naturally. Second, in Laputa beliefs are updated using standard Bayesian conditionalization, whereas H&K employ a more novel, some would say idiosyncratic, approach where opinions are updated by weighing together truth attraction and social influence in a linear fashion. This difference could count in favor of Bayesian conditionalization, at least until H&K's linear updating has received wider scrutiny. Third, Laputa explicitly represents and dynamically updates trust while it is somewhat unclear what role, if any, trust plays in the H&K model. Superficially, considerations of trust does not seem to enter into the latter. However, this interpretation of the model raises the question why people are supposed to take into account only opinions that are close to their own.[7] On an alternative interpretation, individuals in the H&K model trust only those other individuals that share more or less the same opinion. This would explain why they take into account only opinions close to their own. If so, trust does enter into the model, albeit in an implicit way which is open to serious criticism because it makes trust contingent solely on the current view of the individual in question. On a more plausible picture, which in our view is faithfully represented in Laputa, the extent to which we should trust a given individual does not depend only on her current opinion but on our previous degree of trust which can be seen as a compressed representation of that individual's track record. How trust should best be represented and updated is a complex issue that we cannot do full justice to here. It is even unclear whether trust is or should be updated in every process of group deliberation. It seems unlikely that we are constantly busy updating or downgrading our trust in close friends when discussing with them, while the same procedure may be perfectly legitimate in the case of anonymous internet conversations with strangers. Nevertheless, to the extent

[7] See Martini (2012) on this point.

that trust should be updated, we tend to think, for the reasons just given, that Laputa is the more plausible model.

6 Conclusion

We saw that in Laputa a 70–30 split can be obtained even if all inquirers are reliable and initially trust one another. If this is the kind of split we observe on a moral issue, it is not problematic for moral realism from the perspective of Laputa. However, a 50–50 split cannot be obtained with the same settings, except in the trivial case in which nobody is conducting any inquiry, i.e. consulting their moral intuitions. Instead, we saw that there are four possibilities: (a) the inquirers have a reliability of 0.5 (i.e. moral realism in the relevant sense is false), (b) they are reliable but initially take each other to be systematic liars, or (c) they are systematically deceived but initially trust each other, (d) they are reliable but initially trust half their network and distrust the other half. Even if we disregard (c), that still leaves open two explanations that are compatible with moral realism.

We have argued that the support conferred upon the AD by Gustafsson and Peterson's simulation approach is not robust across simulation platforms. In another respect, however, we are in full agreement with these authors. They write (G&P 2010, p. 18):

> However, the most important conclusion is perhaps that the methodology we use appears to be fruitful for moral philosophers wishing to discuss the meta-ethical significance of moral disagreement. By giving up some of the seemingly plausible assumptions we use in our model, it may very well be possible to construct some version of moral realism that is immune to the AD—if so, this would then tell us something important about the structure of a plausible form or moral realism. We thus conclude that computer simulation provide us with a new tool for assessing meta-ethical debates about moral disagreement.

The simulation model we have used is arguably a purely Bayesian model which, for that very reason, has the potential to be viewed as a normative rationality model of group inquiry and communication. Thus, the model can be used for studying moral convergence from an idealized perspective similar to that taken by many moral realists (for a discussion, see Doris and Stich 2006). Our results provide the moral realist with additional potential explanations of moral disagreement given the assumed truth of their standpoint, although our goal has not been to defend moral realism but rather to illustrate the complexity and sophistication of current truth-sensitive models of group inquiry and communication and the corresponding difficulty in interpreting the simulation results emerging from those models.

References

Angere, S. (forthcoming). Knowledge in a social network. *Synthese*.

Boyd, R. N. (1988). How to be a moral realist. In G. Sayre-McCord (Ed.), *Essays on moral realism* (pp. 181–228). Ithaca: Cornell University Press.

Brink, D. O. (1989). *Moral realism and the foundations of ethics*. Cambridge: Cambridge University Press.

Doris, J., & Stich, S. (2006). Moral psychology: Empirical approaches. In *The Stanford encyclopedia of philosophy*.

Douven, I., & Kelp, C. (2011). Truth approximation, social epistemology, and opinion dynamics. *Erkenntnis, 75*(2), 271–283.

Gustafsson, J. E., & Peterson, M. (2010). A computer simulation of the argument from disagreement. *Synthese*, published online 16 October.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation, 5*(3), 1–33.

Hegselmann, R., & Krause, U. (2006). Truth and cognitive division of labour: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation, 9*(3), 1–28.

Isenberg, D. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology, 50*(6), 1141–1151.

Jelen, T., & Wilcox, C. (2003). Causes and consequences of public attitudes toward abortion: A review and research agenda. *Political Research Quarterly, 56*(4), 489–500.

Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. Harmondsworth: Penguin.

Martini, C. (2012). Consensus formation in networked groups. In S. Hartmann & S. Okasha (Eds.), *EPSA philosophy of science: Amsterdam 2009* (pp. 199–215). Dordrecht: Springer.

Olsson, E. J. (2008). Knowledge, truth, and bullshit: Reflections on Frankfurt. *Midwest Studies in Philosophy, 32*, 94–110.

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

Olsson, E. J. (forthcoming). Is polarization rational? A Bayesian simulation model of group deliberation.

Olsson, E. J., & Vallinder, A. (forthcoming). Norms of assertion and communication in social networks. *Synthese*.

Shafer-Landau, R. (2003). *Moral realism: A defence*. Oxford: Oxford University Press.

Zollman, K. (forthcoming). Social network structure and the achievement of consensus. *Politics, Philosophy and Economics*.