Synthese
DOI 10.1007/s11229-013-0375-0

Trust and the value of overconfidence: a Bayesian perspective on social network communication

Aron Vallinder · Erik J. Olsson

Received: 3 October 2011 / Accepted: 27 February 2012

Abstract The paper presents and defends a Bayesian theory of trust in social networks. In the first part of the paper, we provide justifications for the basic assumptions behind the model, and we give reasons for thinking that the model has plausible consequences for certain kinds of communication. In the second part of the paper we investigate the phenomenon of overconfidence. Many psychological studies have found that people think they are more reliable than they actually are. Using a simulation environment that has been developed in order to make our model computationally tractable we show that in our model inquirers are indeed sometimes better off from an epistemic perspective overestimating the reliability of their own inquiries. We also show, by contrast, that people are rarely better off overestimating the reliability of others. On the basis of these observations we formulate a novel hypothesis about the value of overconfidence.

Keywords Trust · Overconfidence · Bayesianism · Social network · Communication · Probability · Reliability

1 Introduction

Bayesians are committed to the view that an epistemic agent's belief state at any given time can be represented as a probability distribution over propositions in some language. Bayesians also believe that a rational agent should react to incoming evidence by means of conditionalization. Thus the new degree of belief an agent assigns to a proposition should equal the old conditional probability of the proposition on the evidence. However, when we receive information from other sources we also tend to adjust our trust in those sources. If the information was expected, this should count,

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: 11 December 2013

if ever so slightly, in favor of trusting the source. If the information was surprising, that might lead us to reduce our trust in the source. How this can be modeled within the framework of Bayesianism is very much an open question.

The first aim of this paper is to present and defend a particular way of modeling and updating trust called *Laputa* that was developed by Staffan Angere in collaboration with one of the authors (Olsson) (see Angere, to appear and Olsson 2011). This is done in Sects. 2 and 3. The rest of the paper is devoted to the phenomenon of overconfidence and especially the sense in which overconfidence might be rational from an epistemological standpoint.

2 A Bayesian model of communication in social networks

By a social network we will mean a set of inquirers with links between them representing communication channels (e.g. email connections, Facebook friendship relations etc.). If there is a link from inquirer A to inquirer B, that means that A can send a message to B. All inquirers focus on answering the question whether p is true, where p is any proposition that can be true or false. The messages they can send are either "p" or "$\neg p$". Each inquirer also has available a private information source, which we refer to as "inquiry". This can be just about any source external to the social network, e.g. a scientific instrument, an informer, a computer database, and so on. Internal sources (network peers) and external sources are treated in a completely analogous fashion in the model. What we say about "sources" in the following therefore applies equally to sources of both kinds.

Each inquirer assigns to p, at time t, a certain credence, $C_t(p)$ (subjective probability). Each inquirer also assigns to each information source a certain degree of trust at t. We can now pose our main problems:

- The Credence Problem: How to update an inquirer's credence in p given new information?
- The Trust Problem: How to update an inquirer's trust in a given source given new information from that source?

Being good Bayesians, we want to solve these two problems by means of *conditionalization on the new evidence*. For the credence problem this means that

$$C_{t+1}(p) = C_t(p \mid \text{source } S \text{ says that } p)$$

or

$$C_{t+1}(p) = C_t(p \mid \text{source } S \text{ says that } \neg p),$$

depending on what the source S says. But how do we compute the right hand side of these equations? Clearly our new credence in p after having listened to S will depend on how much trust we placed in S. So already the credence problem requires that we also model epistemic trust—but how?

The proposal is that we think of trust as being also a form of credence, namely *credence in the reliability of the source*. This idea is not new but goes back to the

Scandinavian School of Evidentiary Value (Edman 1973; Ekelöf 1983; Halldén 1973; Hansson 1983). More recently it has been drawn upon extensively in the literature on epistemic coherence (Olsson 2002a,b, 2005; Angere 2008; Schubert 2010). This earlier work focuses essentially on two possibilities concerning reliability: being fully reliable (truth telling) and being fully unreliable (randomizing). But what about half-way reliable inquirers and what about systematic liars? Let us by a source S's degree of reliability with respect to p mean the (objective) probability that S says that p given (i) that S says anything at all and (ii) that p is true. Let us by a source S's degree of unreliability with respect to p mean the (objective) probability that S says that $\neg p$ given (i) that S says anything at all and (ii) that p is true. Laputa takes into account every possible form of reliability and every possible form of unreliability: for every degree of reliability, the inquirer's *trust function* assigns a credence to the proposition expressing that the source is reliable to that degree and, moreover, for every possible degree of unreliability, the inquirer's *distrust function* assigns a credence to the proposition expressing that the source is unreliable to that degree. The distrust function can be defined in terms of the trust function given that a source S's degree of unreliability $= 1 -$ (S's degree of reliability). For example, an agent's trust function may assign a particular credence to the proposition that the source is 75 % reliable. This makes it clear that trust values are second order probabilities: subjective probabilities about objective probabilities.

We will first address the credence problem for one source and then extend this solution to a solution to the credence problem for n sources. We need a few assumptions in order to be able to proceed.

> (Source Symmetry) S's reliability with respect to p equals S's reliability with respect to $\neg p$.

This assumption rules out cases in which an agent is more likely to detect that p, if p is true, than that $\neg p$, if p is false. While strictly speaking not necessary, Source Symmetry simplifies the model considerably.[1]

We will also need some way of connecting subjective credences with objective chances. This is achieved as follows:

> (Principal Principle) On the assumptions that the source S is (objectively) reliable to degree r, that S will report anything at all and that p is true, an inquirer A should assign credence r to the proposition that S will report that p.

The original Principal Principle goes back to Lewis (1980) and states essentially that an agent's credence in a given proposition, on the assumption that the objective chance of that proposition equals c, should be c. What we here refer to by the same name, following Angere (to appear), is but a special case of that general, almost tautologically-sounding principle.

Finally, we also make the following innocent assumption:

> (Communication Independence) Whether a source S says something is independent of whether p is true as well as of S's degree of reliability.

[1] For a discussion of this assumption and how it can be relaxed, see Olsson (2011).

As we shall see, these assumptions allow us to compute $C_{t+1}(p) = C_t(p|\ S$ says that $p)$, where the latter depends on (i) $C_t(p)$ and (ii) the inquirer's trust function for S (or rather its expected value).

This takes care of the credence problem for the special case of one source. What about the case of n sources?

$$C_{t+1}\,(p) = C_t\,(p \mid \text{source } S_1 \text{says } p, \text{source } S_2 \text{ says } \neg p, \ldots)\,.$$

In order to tackle this case we need to add a further assumption:

> (Source Independence) Each inquirer assumes that the other inquirers are reporting independently.

Source Independence can be expressed in a standard way as a form of conditional independence: the credence assigned to the proposition that source S_1 will report that p is independent of the credence assigned to the proposition that source S_2 will report that p, and so on, conditional on the truth/falsity of p. Given Source Independence, thus interpreted, the general credence problem has a purely mathematical solution (Angere, to appear; Olsson 2013). Independence, in this sense, is often postulated in the literature on epistemic coherence and in artificial intelligence, and it is one of the cornerstones of the theory of Bayesian networks (see e.g. Spohn 1980; Pearl 1988).[2]

It should be noted that Source Independence is plausible as a psychological assumption, i.e., as a default rule of information processing: lacking any reason to think otherwise, we commonly assume that the information we receive from various sources was independently reported, i.e. that the sources have not agreed to give the message beforehand. Source Independence is less plausible from a normative standpoint. As inquirers communicate over time they become increasingly dependent, making the Source Independence assumption increasingly unrealistic, although it can still be accepted as a useful idealization. A possible alternative solution is to reinterpret the model: we may choose to interpret a message to the effect that p ($\neg p$) is true as a claim to the effect that there is a new independent reasons in favor of p ($\neg p$). This path is taken in Olsson (2013).[3,4] In this section and the next, we will be concerned solely with the one-source credence and trust problems.

[2] Despite this similarity, Bayesian networks should be carefully distinguished from networks in our sense.

[3] Incidentally that move also solves a problem of repetition. Suppose one inquirer S in the network is repeatedly reporting the same message, say, p. This will make that inquirer's peers repeatedly update with the information "S said that p". If the messages exchanged between inquirers are simply thought of as claims to the effect that p is true or false, this is not very plausible. If, however, we instead interpret a message that p ($\neg p$) as a message to the effect that there is a novel or independent reason for p ($\neg p$), this reaction to repetition is as it should be.

[4] As pointed out by an anonymous referee, source independence is not a necessary condition for confirmation. Consider a case in which several inquirers believe that p (e.g., "global warming is real") on account of deferring to one and the same expert. The testimonial judgments to the effect that p that these deferring inquirer may make are not independent of one another in the conditional sense. Still, it seems that the fact that a large number of inquirers (dependently) report that p should increase one's credence in the proposition that p. This kind of scenario is studied at length in Olsson (2002a,b) and in Olsson (2005, Sect. 3.2.3), where it is characterized as involving "dependent reliability". The question whether such cases can be modeled in Laputa is a complex one which depends on various other issues, such as how we choose to

Let us now turn to the trust problem: the problem of how to update an inquirer's trust function in the light of new evidence. Interestingly, no additional assumptions are needed to solve the trust problem (and we don't need Source Independence). As we will see, where the source says that p we can now compute

$$T_{t+1}(S \text{ is reliable to degree } r) = T_t(S \text{ is reliable to degree } r \mid S \text{ says that } p),$$

where the right hand side is a function of (i) r, (ii) $C_t(p)$, and (iii) the inquirer's trust function for S at t (or rather the expected value of the trust function).

This concludes our bird-eye exposition of the model. We will now show how to represent these ideas within a Bayesian probabilistic framework, focusing on the one-source case. The epistemic state of a person α at time t is assumed to be given by a *credence function* $C^t_\alpha : L \to [0, 1]$. L can be taken to be a classical propositional language, and C^t_α is assumed to fulfill the standard axioms of a probability measure. We assume, conventionally, that p happens to be true, since this will simplify calculations further on.

Not all participants' approaches to inquiry are the same, and they tend to vary both in their degree of activity and their effectiveness. Let $S^t_{i\alpha} p$ be the proposition "α's inquiry gives the result that p at time t", $S^t_{i\alpha} \neg p$ be the proposition "α's inquiry gives the result that $\neg p$ at t", and $S^t_{i\alpha} p \vee S^t_{i\alpha} \neg p$ the proposition that α's inquiry gives *some* result at t. We represent the participants' properties *qua* inquirers by two probabilities: the chance $P(S^t_{i\alpha})$ that, at any moment t, α receives a result from her inquiries, and the chance $P(S^t_{i\alpha} p \mid S^t_{i\alpha} \wedge p)$ that, when such a result is obtained, it is the right one. $P(S^t_{i\alpha})$ will be referred to as α's *activity*, and $P(S^t_{i\alpha} p \mid S^t_{i\alpha} \wedge p)$ as her *aptitude*. As a simplification, we will assume α's activity and aptitude to be constant over time, so we will generally write them without the time index t.

Analogously to the inquiry notation we define

$S^t_{\beta\alpha} p =_{df} \beta$ says that p to α at t
$S^t_{\beta\alpha} \neg p =_{df} \beta$ says that $\neg p$ to α at t
$S^t_{\beta\alpha} =_{df} \beta$ says that p or that $\neg p$ to α at t.

The strength of a link $\beta\alpha$ is then representable as a probability $P(S_{\beta\alpha})$, being the chance that β communicates that p or that $\neg p$ to α, at any given moment t.

Given that β communicates with α, what does she say? And what makes her say it? These questions are answered by a property of the link $\beta\alpha$ that we will call its *threshold of assertion* or just threshold for short: a value $T_{\beta\alpha}$ between 0 and 1, such that

If $T_{\beta\alpha} > 0.5$, β tells α that p only if $C_\beta(p) \geq T_{\beta\alpha}$, and that $\neg p$ only if $C_\beta(p) \leq 1 - T_{\beta\alpha}$;
If $T_{\beta\alpha} < 0.5$, β tells α that p only if $C_\beta(p) \leq T_{\beta\alpha}$, and that $\neg p$ only if $C_\beta(p) \geq 1 - T_{\beta\alpha}$; and

Footnote 4 continued
interpret communication in the system. We would prefer to save that discussion for a later occasion as it does not bear directly on the points we wish to make in the present article.

If $T_{\beta\alpha} = 0.5$, β can tell α that p or that $\neg p$ independently of what she believes, which is modeled by letting her pick what to say randomly.

In other words, an inquirer will say that p ($\neg p$) only if her credence in p ($\neg p$) has reached the level at which her threshold of assertion has been set.

We now define α's source σ's *reliability* as

$$R_{\sigma\alpha} =_{df} \; P\left(S_{\sigma\alpha} p \mid S_{\sigma\alpha} \wedge p\right) = P\left(S_{\sigma\alpha} \neg p \mid S_{\sigma\alpha} \wedge \neg p\right).$$

This is where the assumption of Source Symmetry comes in: the definition presupposes that the probability that any source gives the answer p, if p is the case, is equal to the probability that it gives the answer $\neg p$, if $\neg p$ is the case.

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

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

for a, b in [0,1]. $t_{\sigma\alpha}(\rho)$ then gives the credence density at ρ, and we can obtain the actual credence that α has in propositions about the reliability of her sources by integrating this function. We will also have use for the expression $1-\tau^t_{\sigma\alpha}$ (which represents α's credence density for propositions about σ *not* being reliable) which we will refer to as $\bar{\tau}^t_{\sigma\alpha}$.

Now, as we saw earlier, in connection with the Principal Principle, an inquirer's credences about chances should influence her credences about the outcomes of these chances. We can now formally represent the relevant special case of that principle as follows:

$$C^t_\alpha\left(S^t_{\sigma\alpha} p \mid S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge p\right) = \rho$$
$$C^t_\alpha\left(S^t_{\sigma\alpha} \neg p \mid S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge \neg p\right) = \rho$$

for all t, i.e. α's credence in σ giving the report p should be ρ on the assumptions (i) that the source gives any report at all, (ii) that σ's reliability is ρ, and (iii) that p actually is the case.

Finally, Communication Independence can be expressed in the following fashion (CI):

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

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

$$\text{(T1)} \quad C^t_\alpha\left(S^t_{\sigma\alpha} \mid p\right) = C^t_\alpha\left(S^t_{\sigma\alpha}\right) \int_0^1 \rho \tau^t_{\sigma\alpha}(\rho)\, d\rho.$$

The integral in this expression is the expected value $\langle \tau^t_{\sigma\alpha} \rangle$ of the trust function $\tau^t_{\sigma\alpha}$, whence

$$\text{(T2)} \quad C^t_\alpha \left(S^t_{\sigma\alpha} \mid p \right) = C^t_\alpha \left(S^t_{\sigma\alpha} \right) \langle \tau^t_{\sigma\alpha} \rangle .$$

Similarly,

$$\text{(T3)} \quad C^t_\alpha (S^t_{\sigma\alpha} \mid \neg p) = C^t_\alpha \left(S^t_{\sigma\alpha} \right) \langle \tau^t_{\sigma\alpha} \rangle .$$

We are now in a position to calculate $C^t_\alpha (p \mid S^t_{\sigma\alpha} p)$ and $C^t_\alpha (p \mid S^t_{\sigma\alpha} \neg p)$, the credence an agent should place in p at t given that the source σ says that p or $\neg p$, respectively:

$$\text{(C1)} \quad C^t_\alpha \left(p \mid S^t_{\sigma\alpha} p \right) = \frac{C^t_\alpha (p) \langle \tau^t_{\sigma\alpha} \rangle}{C^t_\alpha (p) \langle \tau^t_{\sigma\alpha} \rangle + C^t_\alpha (\neg p) \langle \tau^t_{\sigma\alpha} \rangle} .$$

$$\text{(C2)} \quad C^t_\alpha \left(p \mid S^t_{\sigma\alpha} \neg p \right) = \frac{C^t_\alpha (p) \langle \tau^t_{\sigma\alpha} \rangle}{C^t_\alpha (p) \langle \tau^t_{\sigma\alpha} \rangle + C^t_\alpha (\neg p) \langle \tau^t_{\sigma\alpha} \rangle} .$$

where $\langle \tau^t_{\sigma\alpha} \rangle$ is the expected value of the trust function $t^t_{\sigma\alpha}$. By the Bayesian requirement of conditionalization, we must have

$$\text{(C3)} \quad C^{t+1}_\alpha = C^t_\alpha \left(p \mid S^t_{\sigma\alpha} p \right) ,$$

whenever σ is the only source giving information to α at t. This means that our formula completely determines how α should update her credence in such a case. For the many-sources case we need, as we indicated earlier, the additional assumption of Source Independence. We refer to Angere (to appear) for details.[5] As for trust, it is updated according to

$$\tau^{t+1}_{\delta\alpha} (\rho) = \tau^t_{\delta\alpha} (\rho) \frac{\rho C^t_\alpha (p) + (1 - \rho) \left(C^t_\alpha (\neg p) \right)}{\langle \tau^t_{\sigma\alpha} \rangle C^t_\alpha (p) + \langle \tau^t_{\sigma\alpha} \rangle \left(C^t_\alpha (\neg p) \right)}$$

or

$$\tau^{t+1}_{\delta\alpha} (\rho) = \tau^t_{\delta\alpha} (\rho) \frac{\rho C^t_\alpha (\neg p) + (1 - \rho)(C^t_\alpha (p))}{\langle \tau^t_{\sigma\alpha} \rangle C^t_\alpha (\neg p) + \langle \tau^t_{\sigma\alpha} \rangle (C^t_\alpha (p))}$$

depending on whether the message received was p or $\neg p$.

3 Derived rules of trust

Do our solutions to the credence and trust problems satisfy reasonable qualitative rules for how these things should be updated? When determining how the credence in p changes, we are looking for the conditions under which

[5] See Zollman (2007) for an alternative Bayesian model of communication in social networks which does not, however, allow trust to be represented and updated.

(a) $C_\alpha^{t+1}(p) = C_\alpha^t(p)$
(b) $C_\alpha^{t+1}(p) > C_\alpha^t(p)$
(c) $C_\alpha^{t+1}(p) < C_\alpha^t(p)$.

Beginning with (a), (C1) gives us that

$$\frac{C_\alpha^t(p)\langle\tau_{\sigma\alpha}^t\rangle}{C_\alpha^t(p)\langle\tau_{\sigma\alpha}^t\rangle + (1 - C_\alpha^t(p))\langle 1 - \tau_{\sigma\alpha}^t\rangle} = C_\alpha^t(p).$$

After simplification, this is equivalent to

$$2C_\alpha^t(p)\,\langle\tau_{\sigma\alpha}^t\rangle - C_\alpha^t(p) - 2\langle\tau_{\sigma\alpha}^t\rangle + 1 = 0$$

for any $C_\alpha^t(p) \neq 0$. If we also have $C_\alpha^t(p) \neq 1$, we can further simplify the expression as

$$\langle\tau_{\sigma\alpha}^t\rangle = \frac{C_\alpha^t(p) - 1}{2C_\alpha^t(p) - 2} = 0.5.$$

As for (b) and (c), we then get $\langle\tau_{\sigma\alpha}^t\rangle > 0.5$ and $\langle\tau_{\sigma\alpha}^t\rangle < 0.5$, respectively. From (a) we can now see that if the credence is to remain completely unchanged, $\langle\tau_{\sigma\alpha}^t\rangle$ must be exactly 0.5. If $\langle\tau_{\sigma\alpha}^t\rangle$ on the other hand is greater than 0.5, it follows that we must have $C_\alpha^{t+1}(p) > C_\alpha^t(p)$, i.e. the credence is increased. Similarly, if we have $\langle\tau_{\sigma\alpha}^t\rangle < 0.5$, credence is decreased. The derivations are completely analogous in the case when an agent receives a message that $\neg p$.

Let us say that a source is *trusted* if our credence in the reliability of the source is greater than 0.5; *distrusted* if our credence in the reliability of the course is less than 0.5; and *neither trusted nor distrusted* otherwise. We say that a message is *expected* if our credence in it is greater than 0.5; *unexpected* if our credence in it is less than 0.5; and *neither expected nor unexpected* otherwise. The "+" sign means in the following that the message reinforces the inquirer's current belief (i.e. her confidence increases if above 0.5 and decreases if below 0.5). The "–" sign means that the message weakens the inquirer's current belief (i.e. her confidence decreases if above 0.5 and increases if below 0.5). 0 means that the inquirer's credence is left unchanged. We can now summarize the results of our calculations in a table (Table 1).

Suppose, for example, that inquirer A assigns p a credence of 0.75, and that A trusts the source S. Inquirer A now receives the message p from S. This is an expected message coming from a trusted source. Thus we have a situation corresponding to the upper left hand corner of Table 1. The "+" sign there indicates that A's credence in

Table 1 Summary of the derived rules for updating credences in the one-source case

	Message expected	Neither nor	Message unexpected
Source trusted	+	+	−
Neither nor	0	0	0
Source distrusted	−	−	+

p will increase. Or suppose that inquirer A assigns p a credence of 0.25, and that A distrusts the source S. A now receives the message p from S. Thus A is receiving an unexpected message from a distrusted source. This case corresponds to the lower right hand corner of Table 1. This will then make A's degree of belief stronger, i.e. A will believe more strongly that $\neg p$ is the case.

We can also study the effect of prior expectation on posterior trust. Here we are looking for the conditions under which

(i) $\tau_{\delta\alpha}^{t+1}(\rho) = \tau_{\delta\alpha}^{t}(\rho)$
(ii) $\tau_{\delta\alpha}^{t+1}(\rho) > \tau_{\delta\alpha}^{t}(\rho)$
(iii) $\tau_{\delta\alpha}^{t+1}(\rho) < \tau_{\delta\alpha}^{t}(\rho)$.

According to the rule for updating trust, we have that

$$\tau_{\delta\alpha}^{t+1}(\rho) = \tau_{\delta\alpha}^{t}(\rho)\,\frac{\rho C_{\alpha}^{t}(p) + (1-\rho)\left(1 - C_{\alpha}^{t}(p)\right)}{\langle\tau_{\sigma\alpha}^{t}\rangle C_{\alpha}^{t}(p) + \left(1 - \langle\tau_{\sigma\alpha}^{t}\rangle\right)\left(1 - C_{\alpha}^{t}(p)\right)}.$$

Given (i), this is equivalent to

$$\frac{\rho C_{\alpha}^{t}(p) + (1-\rho)\left(1 - C_{\alpha}^{t}(p)\right)}{\langle\tau_{\sigma\alpha}^{t}\rangle C_{\alpha}^{t}(p) + \left(1 - \langle\tau_{\sigma\alpha}^{t}\rangle\right)\left(1 - C_{\alpha}^{t}(p)\right)} = 1.$$

Further simplification gives us that

$$C_{\alpha}^{t}(p) = \frac{\rho - \langle\tau_{\sigma\alpha}^{t}\rangle}{2\rho - 2\langle\tau_{\sigma\alpha}^{t}\rangle} = 0.5.$$

For (ii) and (iii), we have $C_{\alpha}^{t}(p) > 0.5$ and $C_{\alpha}^{t}(p) < 0.5$, respectively. If the trust function is to remain unchanged, we must have $C_{\alpha}^{t}(p) = 0.5$, i.e. the message is neither expected nor unexpected. If there is to be an increase in the trust function, we must have $C_{\alpha}^{t}(p) > 0.5$, i.e. the message is expected. Finally, if there is to be a decrease in the trust function, we must have $C_{\alpha}^{t}(p) < 0.5$, i.e. the message is unexpected. Parallel derivations for the case when the message $\neg p$ is sent give us the same expressions. These results are summarized in Table 2, where "+" stands for an increase in trust, "−" for a decrease and 0 for no change.

By combining the information in Tables 1 and 2, we get a good sense of what effect a particular report will have on an inquirer's credence in p and trust in the source. Suppose, for example, that inquirer A assigns p a credence of 0.75 and trusts the source

Table 2 Summary of the derived rules for updating trust values in the one-source case

	Message expected	Neither nor	Message unexpected
Source trusted	+	0	−
Neither nor	+	0	−
Source distrusted	+	0	−

S. Inquirer A now receives the message p from S. This corresponds to the upper left hand corner of Tables 1 and 2, being a case of expected information stemming from a trusted source. The "+" sign in Table 1 indicates that A's credence in p will be raised. The "+" sign in Table 2 indicates that A's trust in S will also become higher.

In the example, A reacted to incoming confirming evidence by raising not only her credence in the confirmed proposition but also by increasing her trust in the source. The question is: is this an objectionable form of "conformation bias"? According to Nickerson (1998), confirmation bias is "the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (p. 175). Nickerson goes on to list five types of such problematic biases:

(a) Restriction of attention to a favored hypothesis
(b) Preferential treatment of evidence supporting existing beliefs
(c) Looking only or primarily for positive cases
(d) Overweighting positive confirmatory instances
(e) Seeing only what one is looking for.

For better or worse, there is no clear sense in which inquirers in Laputa can restrict attention, preferentially treat evidence or look for something rather than for something else. The fact that the inquirers lack the corresponding cognitive resources and abilities has the fortunate effect of making them immune to confirmation biases of kinds (a), (b), (c), and (e). It remains to consider whether they succumb to biases of kind (d): overweighting positive confirmatory instances. The first question we need to ask is "overweighting positive confirmatory instances in relation to what?" Nickerson's answer is: in relation to opposing evidence (p. 181). Objectionable forms of biases of kind (d) are in other words characterized by asymmetric treatment of positive and negative evidence. But as Tables 1 and 2 show, inquirers in Laputa treat evidence symmetrically: expected messages are taken to confirm both the current view regarding p and the reliability of the source, and unexpected messages are taken to disconfirm both the current view regarding p and the reliability of the source. And if the message was neither expected nor unexpected, its effect on the current view is taken to be confirmatory if the source is trusted and disconfirmatory otherwise. We conclude that there is no reason to think that inquirers in Laputa succumb to any kind of confirmation bias.[6]

In our model, inquirers continuously update their trust in their sources—their own inquiry as well as what they learn from others—depending on what those sources say and the prior degree of trust placed in them. Inquirers are constantly monitoring what their sources are saying and updating their trust accordingly, as if they were always "on alert". While there are situations in daily life in which we need to monitor our sources in this way, we often simply take the reliability of people we are communicating with for granted, especially if we know them well. It is only when there are clear signs of trouble that we take a more skeptical, distrusting stance. There are two factors that may explain why we tend to trust the people we engage with on a daily basis, and why, when there

[6] We could of course imagine an extended model in which communication links are dynamically created in the process of collective inquiry. In such a model, inquirers could be biased to establish links to other inquirers whom they think will confirm their current view, in which case the issue of confirmation bias could indeed be legitimately raised.

is a problem, there are still limits to how much we will allow single reports to influence our trust. One factor is that there might be social norms requiring us to trust in certain circumstances (Faulkner 2010). Presumably, it is part of being a good family member that one places some unconditional trust in the other members. The other factor is that if we have relied on someone for a longer period of time, finding the reports of that person regularly confirmed, he or she has in our eyes built up a considerable track record as an informant. Given the considerable weight of the evidence favoring trust, a few misfortunes may not significantly alter the trust we place in him or her.

The Laputa model, as it stands, does not represent norms; nor does the update mechanism take into account the weight of evidence in favor of trust or distrust, weight of evidence being a concept that is notoriously difficult to represent in a Bayesian framework. This does not prevent the model for being a reasonably realistic one for more skeptical forms of belief and trust updating. One application we have in mind is communication between strangers in an online context for which it is less plausible to think that norms of trust are operating, and in which—at least initially—participants may not yet have established convincing track records of truth telling in the eyes of the persons they are communicating with. Whether the model is realistic as a model of certain forms of online communication, and other scenarios in which norms or track records are largely absent, is ultimately an empirical issue which cannot be completely settled from the position of the armchair. What we can do, as philosophers, is to inquire further into the consequences of this way of representing the dynamics of belief and trust. Our next task will be to study the effect and possible merits of overconfidence from this perspective.

4 The value of overconfidence

It is often observed that human beings are overconfident: we overestimate the reliability of our own judgments.[7] Since our model allows us to model both the actual reliability of sources and the credence inquirer's place in the reliability of those sources, we can model overconfidence by setting the latter credence to a value that exceeds the actual reliability. What we are mainly interested in is whether there is an epistemic value in being overconfident as opposed to being perfectly calibrated. In the following, we will address the issue of overconfidence both with regard to an inquirer's capacity as an inquirer and with regard to other inquirers' capacities as informants. Our study will be conducted in a multi-agent setting.

One problem here is obviously what to mean by "epistemic value". It is a virtue of our model that this notion can be made exact. Following Goldman (1999) we will take the average increase or decrease in credence in the truth, called *veritistic value* (or V-value for short), as the proper measure of epistemic value. Thus, a social practice—a concept that is here understood in the broadest possible sense—such as "being overconfident" will in the fullness of time affect the credence inquirers place in p, which is henceforward assumed to be the true answer to the question whether p. If that effect is positive, so that the average credence in p is raised as the result

[7] See Harvey (1997) for a review of the psychological literature on overconfidence.

of the practice being followed, then the practice is said to have (positive) veritistic value. For example, a community of inquirer may, before inquiry and communication takes place, assign p a credence of 0.6 on the average. Now we let them inquire and communicate for a certain period of time, while being overconfident, after which we inspect their credences in p once more. Suppose we find that the average credence is now 0.7. This would mean that we have a gain of 0.1 in veritistic value.

Obviously, any such process of inquiry and communication takes place in the context of certain background conditions, e.g. that the inquirers are reliable to a particular degree, that they started out with particular prior credences and so on. So the result we get if we follow the above procedure is only the veristic value of a particular *application* of the practice of being overconfident. The trick, if we wish to become independent of particular applications, is to consider a lot of possible applications and then take the average of all the veristic values that they give rise to.

In principle we could do all this by hand, but it would require hiring a team of mathematicians to do all the calculations. Fortunately, a simulation environment has been developed (by Staffan Angere) which allows us to do the computations mechanically. The simulation program is described in Olsson (2011) and we will not repeat the details here.[8] The crucial fact is that the program allows us to study the effect of a social practice, such as overconfidence, for a large number of applications. The program collects veritistic data from the various applications and calculates the overall veristic value automatically based on that data.

We begin by investigating the role of inquiry trust in a network of barely reliable inquirers. "Barely reliable" here means that the reliability = 0.6. We assume that the prior credences are sampled from a uniform distribution over the unit interval, and that the same is true for the activity level of each inquirer. What we are interested in is how the veristic value varies with the expected value of the trust function. We start out by assuming that the inquirers do not communicate but are only engaged in inquiry. The result is seen in Fig. 1, where we have also included the corresponding curves for higher reliability values (0.7 and 0.8, respectively).

As we can see in Fig. 1, the rise in veritistic value is sharpest when the expected value of the trust function is just above 0.5 and the veritistic value continues to increase even as the expected value of the trust function exceeds the inquirer's actual reliability. This shows that, at least in some cases, inquirers in Laputa will be better off veritistically overestimating their own reliability. This effect is more pronounced for lower reliability values.[9]

When we allow agents to communicate (with communication trust = 0.6) and vary the inquiry trust, we again get a V-value that rises steadily with the expected inquiry trust at least for reasonable values for the threshold of assertion.[10]

It is trickier to measure the impact of communication trust without taking inquiry trust into account, since there will not be a change in V-value in a network with

[8] The program Laputa can be downloaded from http://sourceforge.net/projects/epistemenet/.

[9] The following parameter values were used in Laputa. Starting belief, inquiry chance and communication chance were all set to a flat distribution over the unit interval. Population was set to 20, certainty threshold to 0.99, steps to 100 and link change to 0.

[10] For more on the veritistic effect of varying the threshold of assertion, see Olsson and Vallinder (2013).

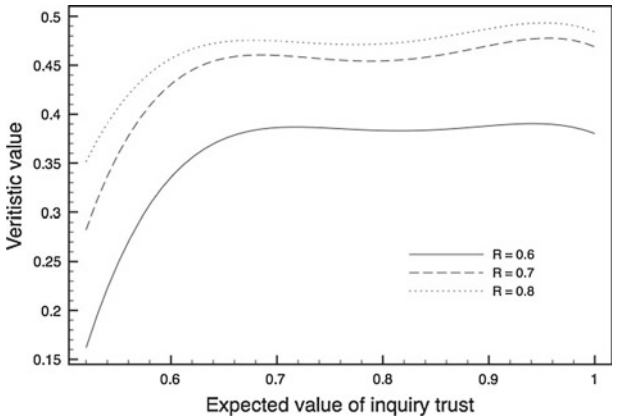

Fig. 1 The veristic value as a function of inquiry trust for agents who do not communicate, as obtained for three different values of inquiry reliability

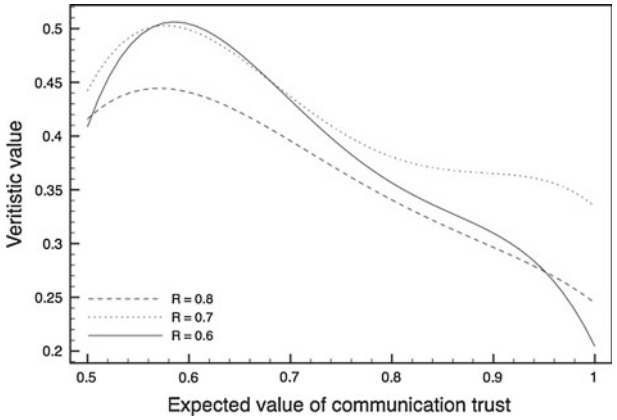

Fig. 2 The veristic value as a function of communication trust for three different values of inquiry reliability

randomly distributed initial credences if there is no inquiry going on (cf. Olsson 2011). We have obtained results for three different values of inquiry reliability, as seen in Fig. 2.

Here we see that a V-optimal value for the communication trust is reached somewhere in the range 0.5–0.6, before the veritistic value begins to drop off again. If we repeat the same experiment with a lower threshold value, the V-optimal communication trust slides further towards 0.5, which in effect means that agents are relying more on inquiry than on communication. For even higher threshold values (e.g. 0.999), the V-optimal communication trust will also be slightly below 0.6.[11]

Why are inquirers better off veritistically if they overestimate their reliability? Since an inquirer is only barely reliable, it may well happen that she is unlucky and

[11] The same parameter values were used as for the preceding experiment, except that inquiry chance was set to 0.6 and link chance to 0.25.

receives a series of "bad results" from her inquiries, i.e. messages to the effect that $\neg p$ is the case (assuming, as always, that p is true). If this happens, this will be interpreted by the inquirer at first as a series of unexpected messages, assuming her present credence in p exceeds 0.5, forcing her to downgrade her trust in her inquiring capacity in accordance with Table 2 (upper right hand corner). At the same time, the inquirer's credence in p will also be reduced, as Table 1 shows. This may lead to the inquirer start thinking she is actually unreliable as an inquirer and that $\neg p$ is probably true. If now the "good results"—a lot of p messages—start coming in, as they normally should considering the inquirer's objective reliability, the inquirer will interpret them as bad results, and as confirming her unreliability. In a suggestive phrase, the inquirer has entered a "spiral of distrust" which she will have a hard time extracting herself from. This is where overconfidence comes in. An overconfident inquirer is less likely to be intimidated by an accidental series of bad results and therefore less likely to enter a potentially disastrous spiral of the kind we just called attention to. Of course, the matter is worse still if the inquirer's initial credence in p is below 0.5, as this makes a series of perceived bad results even more likely, making a distrust spiral more probable and overconfidence correspondingly more valuable. An inquirer could very well start with a credence below 0.5 if her prior credence is sampled from a uniform distribution over the unit interval, as was the case in our experiments.

If this explanation is correct, we should expect that lowering the probability of an accidental series of bad results by increasing inquirers' reliability will have the effect that overconfidence is no longer as valuable as before. To test this, we compared three different values for inquiry reliability (0.6, 0.7 and 0.8). For each of these values, we increased inquiry trust from 0.6 to 1, by steps of 0.1. Since we are interested in the value of overconfidence, we then compared the first increase in inquiry trust that led to overconfidence for the three different reliability values (i.e. the increase in inquiry trust from 0.6 to 0.7 for $R = 0.6$, the increase from 0.7 to 0.8 for $R = 0.7$, etc). For $R = 0.6$, the V-value increased by 70 %; for $R = 0.7$ it increased by 19 %, and for $R = 0.8$ it increased by 9 %. As the increase in V-value diminishes when reliability goes up, this gives prima facie support to our explanation. The support is not conclusive since there is a diminishing marginal return in V-value when inquiry trust approaches 1 (as seen in Fig. 1), which could also influence our test results.

Why does the inquirer not experience a similar advantage from being overconfident in the reports of other inquirers? One possible explanation for this is that while the results of inquiry only depend on the inquirer's reliability, the results of communication depend not only on the communicator's reliability, but also on her initial credence. Since initial credences were evenly distributed between 0 and 1 in our setup, this means that communication is noisier than inquiry, making overconfidence relatively risky from a veritistic standpoint. This explanation stands in line with one we presented for a similar phenomenon in the context of assertion thresholds (Olsson and Vallinder 2013). We can test this explanation by instead having initial credences evenly distributed between 0.5 and 1. In this case, if our explanation is correct, agents shouldn't be as penalized for placing very high trust in others, because communication will be less noisy. Simulation results show that, for all tested expected values of the

communication trust function above 0.52, virtually all agents in the network converge on the truth. This result holds even as the expected value reaches 1.

These results might be relevant to the debate over peer disagreement. In our framework, we could interpret epistemic peers as agents that are (i) equally reliable in their inquiries, and (ii) equally good at weighing the results of their inquiries: that is, they have identical inquiry trust functions. According to the "equal weight view", you should give the same weight to the opinion of a peer as you give your own (Christensen 2007; Elga 2005). Another option would be to assign greater weight to your own opinion. This is the "steadfast view" (Kelly 2005). One natural way of representing the equal weight view in our framework is as saying that your communication trust function for peers should be identical to your inquiry trust function. From this perspective, our results lend support to the steadfast view, i.e. to the thought that you should give more weight to your own inquiry. However, it should be noted that in many of the cases considered in the literature on peer disagreement, there is only one instance of communication, and no further inquiry is taking place. Our simulations pertain to a related but distinct class of cases. Moreover, one might question our conclusion on the basis that real epistemic peers are unlikely to have their initial credences evenly distributed between 0 and 1. As we saw, when initial credences are closer to the truth and distributed more narrowly, being overconfident in the reliability of other inquirers has no negative epistemic effects.

5 Conclusion

We presented and defended a Bayesian model of trust in social networks. We started out by providing justifications for the basic assumption behind the model. They were seen to include standard Bayesian assumptions as well as a few substantial additional principles:

- Trust as credence in the source's reliability
- The Principal Principle
- Source Independence.

We also assumed Source Symmetry and Communication Independence but they can be classified as simplifying assumptions of a seemingly innocent kind. We found that all the substantial assumptions have a firm independent standing in the philosophical literature. This particular way of viewing trust as a form of credence derives from the Scandinavian School of Evidentiary Value. The Principal Principle, although hotly debated, is still a principle which many philosophers find attractive as providing a link between subjective credence and objective chance. Finally, Source Independence is an assumption that one finds in many applications of probability theory, and as we saw it plays a central role in the celebrated theory of Bayesian networks.

We went on to derive a number of qualitative updating principles for credence in p as well as for trust. Some of those principles reminded us of the issue of confirmation bias in cognitive psychology. On closer scrutiny, we found that the model does not embody or legitimize any objectionable form of such bias. We also noted that the way trust is monitored and updated in the model corresponds to a potentially deceptive

situation in which norms of trust or track records have not been established, e.g. online communication between strangers.

In Sect. 4 we studied the effect of overconfidence using a simulation environment that has been developed in order to make our model computationally tractable. We showed that in our model inquirers are sometimes better off from an epistemic perspective overestimating the reliability of their own inquiries. Our explanation of this phenomenon, for which we offered some (inconclusive) independent evidence, was that overconfidence protects the inquirer from a kind of self-defeating doubt that may arise from observing a string of bad results. We put forward this as a possibly novel partial explanation of why people are overconfident. McKay and Dennet (2009) suggest that so-called "positive illusions" are adaptive from an evolutionary point of view, and Johnson and Fowler (2011) present a model which shows that overconfident populations are stable in a wider range of environment than unbiased ones. As far as we know, however, ours is the first explanation that takes overconfidence to be beneficial from a purely *epistemic* point of view. We also showed that people are rarely better off overestimating the reliability of others, an effect that we attributed to the noise inherent in reports from others resulting from randomly distributed prior credences.

Acknowledgments We would like to thank two anonymous referees for their input which led to many significant improvements and clarifications.

References

Angere, S. (2008). Coherence as a heuristic. *Mind*, *117*, 1–26.
Angere, S. (to appear). Knowledge in a social network. *Synthese*.
Christensen, D. (2007). Epistemology of disagreement: The good news. *Philosophical Review*, *116*(2), 187–217.
Edman, M. (1973). Adding independent pieces of evidence. In B. Hansson (Ed.), *Modality, morality and other problems of sense and nonsense* (pp. 180–188). Lund: Gleerup.
Ekelöf, P.-O. (1983). My thoughts on evidentiary value. In P. Gärdefors, B. Hansson, & N.-E. Sahlin (Eds.), *Evidentiary value: Philosophical, judicial and psychological aspects of a theory* (pp. 9–26). Lund: Library of Theoria.
Elga, A. (2005). Reflection and disagreement. *Noûs*, *41*(3), 478–502.
Faulkner, P. (2010). Norms of trust. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Social epistemology* (pp. 129–147). Oxford: Oxford University Press.
Goldman, A. I. (1999). *Knowledge in a social world*. Oxford: Clarendon Press.
Halldén, S. (1973). Indiciemekanismer. *Tidskrift for Rettsvitenskap*, *86*, 55–64.
Hansson, B. (1983). Epistemology and evidence. In P. Gärdefors, B. Hansson, & N.-E. Sahlin (Eds.), *Evidentiary value: Philosophical, judicial and psychological aspects of a theory* (pp. 75–97). Lund: Library of Theoria.
Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Science*, *1*(2), 78–82.
Johnson, D., & Fowler, J. (2011). The evolution of overconfidence. *Nature*, *477*, 317–320.
Kelly, T. (2005). The epistemic significance of disagreement. *Oxford Studies in Epistemology*, *1*, 167–196.
Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2). Berkeley, CA: University of California Press.
McKay, R., & Dennet, D. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, *32*, 493–561.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.
Olsson, E. J. (2002a). Corroborating testimony, probability and surprise. *British Journal for the Philosophy of Science*, *53*, 273–288.

Olsson, E. J. (2002b). Corroborating testimony and ignorance: A reply to Bovens, Fitelson, Hartmann and Snyder. *British Journal for the Philosophy of Science*, *53*, 565–572.

Olsson, E. J. (2005). *Against coherence: Truth, probability and justification*. Oxford: Oxford University Press.

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 F. Zenker (Ed.), *Bayesian argumentation, Synthese library* (pp. 113–134). New York: Springer.

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

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Palo Alto, CA: Morgan-Kaufmann.

Schubert, S. (2010). Coherence and reliability: The case of overlapping testimonies. *Erkenntnis*, *74*, 263–275.

Spohn, W. (1980). Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic*, *9*, 73–99.

Zollman, K. J. (2007). The communication structure of epistemic communities. *Philosophy of Science*, *74*(5), 574–587.