Mis-and Disinformation in a Bounded Confidence Model

Igor Douven, Rainer Hegselmann

Mis- and Disinformation in a Bounded Confidence Model

Article *in* Artificial Intelligence · November 2020

DOI: 10.1016/j.artint.2020.103415

CITATION
1

READS
177

2 authors:

Igor Douven
French National Centre for Scientific Research
200 PUBLICATIONS **2,499** CITATIONS

SEE PROFILE

Rainer Hegselmann
Frankfurt School of Finance & Management
65 PUBLICATIONS **3,230** CITATIONS

SEE PROFILE

Mis- and Disinformation in a Bounded Confidence Model

Igor Douven
SND / CNRS / Sorbonne University
igor.douven@sorbonne-universite.fr

Rainer Hegselmann
Frankfurt School of Finance and Management
r.hegselmann@fs.de

Abstract

The bounded confidence model has been widely used to formally study groups of agents who are sharing opinions with those in their epistemic neighborhood. We revisit the model with an eye toward studying mis- and disinformation campaigns, which have been much in the news of late. To that end, we introduce typed agents into the model, specifically agents who can be irresponsible in different ways, most notably, by being deceitful, but also by being reluctant to try and obtain information from the world directly. We further add a mechanism of confidence dynamics to the model, which—among other things—allows agents to adapt the closeness threshold for counting others as being their epistemic neighbors. This will be used to study the effectiveness of possible defense mechanisms against mis- and disinformation efforts.

Keywords: agent-based modeling; belief change; bounded confidence; confidence dynamics; disinformation; misinformation; non-truthfulness.

1 Introduction

There is much recent work on how best to organize communities of interacting agents if such communities are to achieve some fixed goal. For instance, researchers have looked into whether putting in place certain types of communication structures enables us to improve the efficacy of our belief-forming practices, and if so, which types of communication structures are most helpful in that respect (Zollman, 2007; Kummerfeld & Zollman, 2016; Douven & Wenmackers, 2017; Rosenstock, Bruner, & O'Connor, 2017; Douven, 2019, 2021a; Hahn, Hansen, & Olsson, 2020). Most of this work assumes communities to consist of strictly benevolent agents, all of whom are willing to contribute to the common goal. The present paper relinquishes this assumption and considers the possibility that some agents interact with less than benign motives, even to the extent that they are overtly or covertly carrying out mis- or disinformation campaigns, where (to a first approximation) by a *mis*information campaign we mean one to encourage belief in one or more falsehoods, while by a *dis*information campaign we mean one to manufacture doubt about, or otherwise undermine trust in, one or more truths, so as to discourage others from believing those truths (see O'Connor & Weatherall, 2019). Our main focus will be on whether certain kinds of communication structures—certain sets of rules for determining who to talk and listen to, and for how to take their opinions into account—offer better protection against such campaigns than others.

We trust that the importance of our research topic needs little stressing. Mis- and disinformation campaigns, for financial or political gains, and sometimes for reasons we are still grappling to understand, are the order of the day. At least since the campaigns preceding the vote on Brexit and the 2016 presidential election in the United States, it is generally recognized how dangerous this trend is, jeopardizing the foundations of Western democracies and possibly even the future of our planet as a habitable place. And in its report of 2 February, 2020, the World Health Organization warned that the COVID-19 pandemic is accompanied by an "infodemic"—a stream of myths and rumors about the disease, in particular about prevention measures and cures—which makes it difficult for the public to identify trustworthy sources of information and thereby poses an immediate threat to public health.

This recognition of the harm that can be done by mis- and disinformation has already led to a great number of publications offering analyses of what is at the root of the evil and often also proposing countermeasures that could, or ought to, be taken. These publications have come from academic researchers, but also from think tanks, and governmental and nongovernmental bodies.[1]

Most of these publications have focused on socio-economic issues—such as economic inequality, differences in educational background, and the vanishing of the middle class—on the role the Internet and social media play in the dissemination of falsehoods, or on policies and legislation that might help combat mis- and disinformation campaigns, for instance, by addressing income inequality, ensuring equal access to high-quality education, and regulating the Internet and curbing the power of social media.

Much of this research has led to valuable new insights, and many of the suggested fixes are causes worth fighting for. At the same time, we believe previous research has left some questions about the finer mechanics of mis- and disinformation campaigns unanswered, for instance, whether certain communication structures may make us more vulnerable to such campaigns than others, which strategies the ill-intending might use most effectively, whether we ourselves might be able to protect our society against campaigns of the said types by changing our attitudes toward the opinions of others (e.g., those we most vehemently disagree with), or how to quantify the specifically *epistemic* damage done by mis- and disinformation.

This paper aims to address these and related questions with the help of agent-based computational models, which have become increasingly popular in the field of artificial intelligence (see, e.g., Shoham, Powers, & Grenager, 2007; Tamargo, Garcia, Falappa, & Simari, 2014; Nunes & Antunes, 2015; Gottifredi et al., 2018; Douven, 2019). More specifically, our methodological starting point is the so-called bounded confidence model (or BC model, for short) developed in Krause (2000, 2015) and Hegselmann and Krause (2002, 2005, 2006, 2009, 2015, 2019), which studies groups of epistemically interacting agents.[2] In changing their opinions, the agents in this model are sensitive, to a certain

[1] Contributions by academic researchers include, most notably, Proctor and Schiebinger (2008), Mocanu et al. (2015), Del Vicario et al. (2016), Nichols (2017), Temin (2017), Mason (2018), Vosoughi, Roy, and Aral (2018), O'Connor and Weatherall (2017, 2019), and Weatherall, O'Connor, and Bruner (2020). Noteworthy institutional reports have been issued by the World Economic Forum (Howell, 2013), the Rand Corporation (Kavanagh & Rich, 2018), and, in 2019, the Digital, Culture, Media and Sport Committee of the British House of Commons. Work by journalists is also to be acknowledged; see, for instance, Pomerantsev (2019) and various reports by the *New York Times* (e.g., https://www.nytimes.com/2018/11/12/opinion/russia-meddling-disinformation-fake-news-elections.html and https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html).

[2] For related models, see Deffuant et al. (2000), Dittmer (2001), Weisbuch et al. (2002), Jacobmeier (2004), Semeshenko, Gordon, and Nadal (2008), Lu, Korniss, and Szymanski (2009), Tsang and Larson (2014), Gao et al. (2017), and Chen and Lou (2019).

degree, to the opinions of other agents in the group. Additionally, in Hegselmann and Krause (2006, 2009), the agents are sensitive to information coming directly from the world they inhabit. This captures the idea that, while there is an unmistakable social aspect to learning, we also learn by inspecting the world directly for ourselves. The idea underlying the BC model makes sense not only from a normative standpoint (see Sect. 2.2 for more on this); if not introspectively clear, there is undeniable evidence that we do form our opinions on the basis of both the opinions of others and the results of our probing the world directly (Mason, Conrey, & Smith, 2007; Lorenz et al., 2011; Mason & Watts, 2012).

In its extant form, however, the BC model assumes all agents to act responsibly in that they do not hide their opinions from others in the group, let alone lie about their real opinions to mislead those others, and that they are open to the information that comes in from the world. In view of the above remarks on mis- and disinformation campaigns, one can only conclude that those assumptions are far enough removed from reality for the BC model to be at risk of being inapplicable to some of the socio-epistemic, and in particular politico-epistemic, phenomena that, arguably, are currently of the greatest interest to us.

This paper aims to take some first steps toward mending that situation by proposing two extensions to the model. Specifically, we extend the model by introducing (i) new types of agents and (ii) the mechanism of confidence dynamics. As for (i), communities in the extended model can consist not only of agents that are epistemically responsible in the sense explained above, but also of epistemically *irresponsible* agents, where the irresponsibility of the latter type of agents can vary: they can dogmatically stick to an opinion, but they can also be just not interested enough in the truth to make any truth-finding efforts themselves. As for (ii), the original BC model fixes from the start the degree to which agents rely on others, where this degree is also mostly (though not always) taken to be the same for all agents in a community. But in reality people will not all be trusting others to the same degree, nor will those degrees be fixed once and for all; to the contrary, they are likely to be influenced by the extent to which those they interact with trust others, and will adjust their level of trust depending on what they experience in their encounters with others.

No one should expect a formal analysis of communication structures, on its own, to suggest an easy fix of the problems that mis- and disinformation campaigns are causing. What we are aiming at instead is to achieve a deeper understanding of *why* these problems have proven so recalcitrant, and to get at least some sense of the direction or directions in which progress may lie.

We start, in Section 2.1, by summarizing the BC model in its original form and by highlighting some important limitations. In Section 3, we present the first extension of the model, featuring different types of agents. Section 4 adds the second extension, introducing the concept of confidence dynamics. In both sections, we also show how the new machinery can be used to address questions concerning the degrees to which evildoers are able to exploit different settings of the parameters of the model. These parameters correspond to various ways in which agents can be liberal in counting others as their peers—the agents they deem worthy of letting themselves be influenced by—and to the weight they give to their peers' opinions, but also to the measure in which a community is infiltrated by forces aiming to undercut the truth-finding process of others as well as to the proportion of members unwilling to attend to worldly evidence.

2 Theoretical background

In this section, we review previous work on the BC model and some of its notable variants, which collectively serve as a starting point for the present research. We also say more about the questions that motivate our endeavor.

2.1 Bounded confidence updating

The broad availability of fast and powerful computers has made agent-based computational modeling a popular tool for studying complex social phenomena that are difficult or even impossible to investigate analytically. A relatively recent branch of this program focuses on socio-epistemic phenomena, specifically aspects of knowledge and belief acquisition to which the interaction among agents is central. For example, it is nowadays regarded as a truism that most successes of modern science could not have been achieved by researchers working in complete isolation of one another (Kitcher, 1992; Gribbin, 2002).

A widely used agent-based computational model is the one first presented in Krause (2000) and Hegselmann and Krause (2002) in which agents change their opinions by "averaging" (in some way) over the opinions of those epistemically close (in some sense) to them. In Hegselmann and Krause (2006, 2009), the agents also receive direct evidence from the world. Thereby the model covers in a highly idealized way the fundamental structure of our epistemic situation: learning from others and, at the same time, learning from the world. Many publications have used the model for investigating descriptive questions, most notably, questions concerning the conditions that lead a community of initially disagreeing agents to reach a consensus and those that lead to polarization (e.g., Lorenz, 2003, 2008). A major focus of studies lay, and (due to many open questions) still lies, on the time that it takes to reach a stable final pattern (Chazelle, 2011; Kurz & Rambau, 2011; Kurz, 2015; Hegarty & Wedin, 2016). Other work has recruited the model to shed light on a number of normative issues of interest mostly to philosophers, for instance, concerning the practice of assertion (Olsson, 2008), the resolution of disagreement amongst peers (Douven, 2010), and efficient truth approximation (Douven & Kelp, 2011).

Let $\tau \in [0, 1]$ be the value of some parameter (whose exact nature we leave unspecified) that the agents in a given community are trying to determine. In the BC model, these agents update their estimates of τ repeatedly, at discrete points in time, where an agent updates on the basis of (i) information about τ received from the world and (ii) the estimates of τ of those agents who are within her bounded confidence interval, or BCI, meaning that their estimate of τ is within some distance ε of the agent's own estimate. Formally, agent x_i's opinion concerning τ after the $(u+1)$-st update is defined to be

$$x_i(u+1) \quad = \quad \frac{1-\alpha}{\left|X_i(u)\right|} \sum_{j \in X_i(u)} x_j(u) \quad + \quad \alpha\tau, \tag{BC}$$

with $x_j(u)$ being the opinion of agent x_j after update u,

$$X_i(u) \quad := \quad \left\{ j \colon \left| x_i(u) - x_j(u) \right| \leqslant \varepsilon \right\}$$

the set of agents within agent x_i's BCI after update u, and $\alpha \in [0, 1]$ a parameter determining the weight the agent gives to the "evidential" part of the updating relative to the "social" part.

The formalism is easiest understood through an illustration. Figure 1 shows, for different settings of the parameters, how 50 agents who start out by randomly picking an initial estimate of the value

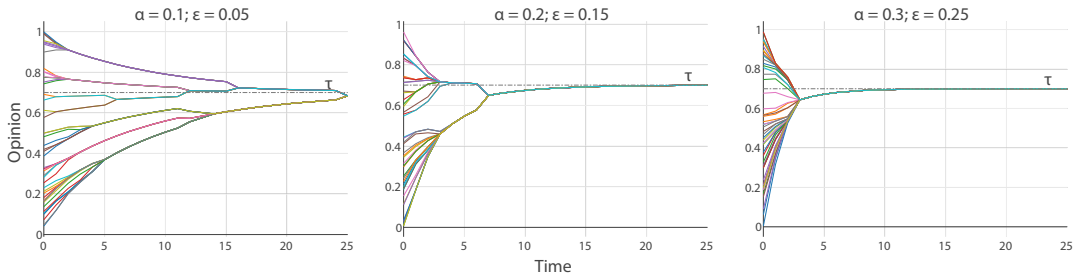

Figure 1: Repeated BC updating of communities of 50 agents, for $\tau = 0.7$ with different values for α and ε.

of τ converge to τ (= 0.7) almost completely, despite the fact that each update is based to a much greater extent (even in the $\alpha = 0.3$ case) on the estimates of those within an agent's BCI than on the information coming in from the world. The figure further shows that while the agents converge on the value of τ relatively fast in all three situations, the parameter settings have a notable impact on exactly *how* fast the convergence occurs.

Many working on agent-based modeling have considered the BC model a good starting point for their own research because it generally takes little effort to extend or otherwise tweak the model to one's own needs (Hegselmann, 2004). Douven (2010), Douven and Riegler (2010), and De Langhe (2013) present straightforward extensions of the model which are all meant to investigate situations in which agents receive "noisy" evidence. Crosscombe and Lawry (2016) are interested in the issue of vagueness and present a further extension of the BC model which is populated by agents whose beliefs can consist of intervals rather than point estimates. Somewhat more complicated extensions are to be found in Lorenz (2003, 2008), Jacobmeier (2004), and Pluchino, Latora, and Rapisarda (2006). In these extensions, agents hold beliefs about multiple issues at the same time rather than about a single parameter. While these extensions are still restricted to agents having *numerical* beliefs on *unconnected* issues, Riegler and Douven (2009) propose an extension of the BC model populated by agents who can hold many beliefs on issues that are not necessarily numerical and that can be logically interconnected (see, in the same vein, Wenmackers, Vanpoucke, & Douven, 2012, 2014).

Douven and Wenmackers (2017) present an extension that deviates even further from the original BC model (see also Douven, 2019). Their extension features agents whose belief states at a given time are characterized by probability functions on a set of self-consistent, mutually exclusive, and jointly exhaustive hypotheses. They let the agents in the model interact by pooling the probabilities of those within their BCI, where this notion is redefined in probabilistic terms but entirely in the spirit of the original model. Douven and Wenmackers use this version of the BC model to unpack the updating on worldly evidence, which in the original BC model is a black box (see also Douven, 2021b). In particular, they unpack it in two different ways—one a version of Bayes' rule, the other a formalization of so-called Inference to the Best Explanation—and compare their behaviors along a number of epistemically important dimensions (such as their accuracy; see below). The extensions to be presented in the following could be combined with Douven and Wenmackers' model, but this is a topic for future research; in the present paper, we stick to treating the updating on worldly evidence as a black box, as is done in (BC).

2.2 Varieties of deceit

It is natural to suppose that liars aim to have us believe whatever it is they falsely assert. That is not necessarily the case, however. Their purposes might be served as well if their false assertions make us retract beliefs we had previously adopted, or keep us from accepting something our evidence would otherwise have inclined us to believe. Depending on what their purpose is—converting us to a view they falsely profess or diverting us from a view we tend to endorse—they may want to follow different strategies of deceit.

Consider a politician firmly convinced that their base is going to support them whatever their view on issue X is (X might for instance be climate change, or the trade deficit with China, or the threat of Iran, or immigration); most of their potential voters do not care much about X, which however is of great importance to a sizeable number of voters who are seriously considering voting for an opponent. A politician in this kind of situation—if cynical enough—may reason that a *subtle* lie on X (e.g., "Sea levels are rising more slowly than scientists report") will not succeed in creating enough doubt about climate science to have a noticeable effect on the turn-out for their opponent. By contrast, a *blatant* lie (e.g., "Climate scientists have been bribed by the liberal elites to publish results supporting green policies") may work, in that it will result in just enough doubt about the opponent's view on X (e.g., "We should take drastic measures to reduce carbon emissions, even if that comes at the expense of economic growth") to curb some of the public's erstwhile enthusiasm for the candidate, which in turn may be just enough to make them stay home on election day. This possibility is far from academic, as we have all been able to read in reports about the Brexit campaign as well as the Trump campaign for the 2016 presidential election (see also O'Connor & Weatherall, 2019).

Accordingly, we distinguish between two types of campaign, to wit:

Misinformation campaign: an effort to deceive a target public about a given proposition X, with conversion as a goal, that is, aiming to make that public believe a falsehood contrary to X (i.e., some proposition Y such that Y entails the negation of X);

Disinformation campaign: an effort to deceive a target public about a given proposition X, with diversion as an explicit goal, that is, aiming to lure away that public from believing X (not necessarily by making it believe some other proposition inconsistent with X).

Assuming an at least minimally rational public, a successful misinformation campaign will automatically mean a successful disinformation campaign. However, for that same reason, a disinformation campaign will typically have a greater chance of success than a misinformation campaign. Whether a disinformation campaign is enough will depend on the situation.[3]

For example, from the perspective of tobacco producers, all that matters may be that people *do not believe* that smoking is detrimental to their health (say, by raising doubts about certain scientific studies), not so much that they *believe* that smoking is *not* detrimental to their health. The former may be all that is needed to keep tobacco sales at a profitable level; trying to convince the public that smoking is actually safe may then come at an additional cost with no corresponding return. Similarly, to win an election, it may be enough to suppress voter enthusiasm among potential voters for your opponent. These people may never vote for *you*, but just to dampen voter turnout for your opponent may be enough for you to win. And to diminish this kind of enthusiasm, it may be enough to divert

[3] Hegselmann et al. (2015) discuss in detail problems of an optimal campaign design for a BC dynamics where however no true value is involved.

the public from the truth, not necessarily to make them believe whatever lies you are spreading. On the other hand, most Brexiteers seem to have made a serious effort to convince their countrymen that Leave was the right choice to make in the referendum (see the report from the committee of the British House of Commons referenced in note 1).

As mentioned in the introduction, the question whether there is anything we can do to protect ourselves against these kinds of campaigns has been much in the limelight both in academic and non-academic publications. Against the background of the mathematical model of communication we are considering here, we may ask whether it could help to be selective in counting others as our peers (i.e., to set ε to a small value) but then give a lot of weight in updating to their opinions (so set α to a relatively small value). Or would it be more effective to be rather inclusive as regards peerhood (set ε to a relatively large value) but then not attach too much weight to our peers' opinion (set α to some high value)? More generally, are there combinations of α and ε values that minimize the likelihood for a misinformation campaign to succeed? If so, do the same combinations offer maximum protection against disinformation campaigns?

There is at least one straightforward answer to these questions, to wit, set either ε equal to 0 or α equal to 1, or both. If you only go by the evidence you get directly from the world, none of the other community members will have any influence on how you form your opinions. A fortiori, none of them will be able to either convert you to their view or divert you from the truth that the worldly evidence is steering you toward.

Note, however, that this would amount to giving up on social learning entirely, and that—we submit—is not a realistic option. Arguably, we would be essentially helpless were it not for what we have learned from others. But even if not, it is certainly correct that, as Schurz (2019, p. 193) points out, individual learning tends to be much more costly than social learning: "Many unsuccessful trial-and-error steps are involved in individual learning that can be avoided by just being informed about the results of these steps."[4] Allowing ourselves to be influenced by the views of others is not just something we happen to do, it is something we *ought* to do—*in addition* to investigating the world ourselves, rather than as a replacement for that. Hence, what we should be looking for is a possibility to be at the same time somewhat sensitive to the opinions of others and still relatively safe from mis- and disinformation campaigns.

Precisely because the BC model recognizes that social learning is not an all-or-nothing matter, and that we may want to be more or less liberal in counting others as peers and may want to give more or less weight to our peers' opinions, it appears eminently suitable to investigate the above kind of questions. Unfortunately, however, and as indicated already, in its present form the model makes no provision for representing untruthful agents, the kind of agents running mis- or disinformation campaigns. The assumption that all agents are truthful may, for many purposes, be a harmless and perhaps even useful idealization—but it makes the model unusable as a tool for answering questions about mis- and disinformation campaigns.

As we mentioned, however, one reason why the BC model has gained popularity is its great flexibility. In the following we aim to show that, because of this flexibility, the above questions do not motivate a radical rethinking of the computational modeling of epistemically interacting agents. To the contrary, it is relatively easy to "concretize" (Nowak, 1980; Kuipers, 2001) or "de-idealize" (Mäki, 1992) the BC model with an eye toward modeling interactions among agents not all of whom are truthful or willing to make a serious effort of informing themselves.

[4] On the importance of social learning, see also Goldman (1999).

In the next section, we take a first step toward this concretization, by extending the BC model to one that covers communities with non-truthful agents. In Section 4, we address another limitation of the BC model, to wit, that it treats the level of "open-mindedness," in the sense of willingness to take others' opinions into account, as being fixed, instead of being itself open to change. The first extension allows us to model the kind of mis- and disinformation campaigns that motivated the present project, the second allows us to model possible countermeasures against such campaigns.

3 Making room for non-truthfulness

In the extension to the BC model to be introduced in this section, agents are *typed* according to how epistemically responsible they are. We look at the effects of irresponsible agents on the truth-seeking endeavors of the responsible agents, where the latter are the agents that already populated the original BC model.

3.1 Typed agents

The agents that populated the original BC model will henceforth be called "truth-seekers." In a first step, we introduce only one new type of agents, to be called "campaigners." Agents of this new type do not update in the normal way. In fact, in a clear sense they do not update at all, but rather stick to a fixed opinion $\varrho \in [0, 1]$ about the value of τ. In all situations to be considered, it will hold that $\varrho \neq \tau$, though important questions will concern how far removed ϱ is from τ.[5]

Formally, then, the extension to be studied first is characterized by the following bounded-confidence-with-campaigners (BCC) updating operation:

$$x_i(u+1) \;=\; \begin{cases} \frac{1-\alpha}{|X_i(u)|} \sum_{j \in X_i(u)} x_j(u) \;+\; \alpha\tau & \text{if } x_i \text{ is a truth-seeker} \\ \varrho & \text{if } x_i \text{ is a campaigner,} \end{cases} \tag{BCC}$$

where $x_j(u)$ and $X_i(u)$ are as defined previously.

As a further extension, we also want to introduce a type of agent unwilling to gather, or even listen to, worldly evidence but only updating by averaging the opinions of those within her BCI. We refer to such agents as "free riders." These agents may not have an agenda, hidden or otherwise, to deflect the truth-seekers from the truth. But their unwillingness to make a serious effort to inform themselves, other than by listening to their epistemic neighbors, may nonetheless make them complicit in mis- or disinformation efforts. Whether that is really so is something we hope to determine.

The formal specification of this bounded-confidence-with-campaigners-and-free-riders (BCCF) model is as follows:

$$x_i(u+1) \;=\; \begin{cases} \frac{1-\alpha}{|X_i(u)|} \sum_{j \in X_i(u)} x_j(u) \;+\; \alpha\tau & \text{if } x_i \text{ is a truth-seeker} \\ \frac{1}{|X_i(u)|} \sum_{j \in X_i(u)} x_j(u) & \text{if } x_i \text{ is a free rider} \\ \varrho & \text{if } x_i \text{ is a campaigner,} \end{cases} \tag{BCCF}$$

[5] What we call here "campaigners" are called "radicals" in Hegselmann (2014, 2020) and Hegselmann and Krause (2015). Now, in the context of this paper on mis- and disinformation, the essence is that there is a constant false signal which is transmitted n times and that these transmissions are "heard" by all agents that have the false signal within their BCI. In the present paper, we interpret that situation in terms of the workings of a group of campaigners. Another interpretation one might want to consider is that of a leader with a certain degree of charisma (indicated by n) or a team of campaigners together with a charismatic leader (their combined total strength or intensity given by n). See also Hegselmann and Krause (2015, Sects. 1.3 and 6). See Section 5 for more on the interpretation of the notion of campaigner.

again assuming the earlier definitions of $x_j(u)$ and $X_i(u)$.

Needless to say, our focus will be only on the truth-seekers and, when present, the free riders; as for the campaigners, there is nothing to know about them that we do not know already from their definition: they simply stick to their fixed opinion under all circumstances, no matter how many updates we consider.

The first questions now to be looked at are (i) how "successful" campaigning is, specifically how successful campaigners are in converting truth-seekers (or at least in diverting them from the truth), and (ii) how much damage to a society the presence of campaigners can do, depending on the circumstances.

3.2 Conversion

To develop an understanding of the impact campaigners can have, we start with the simplest model, setting $\alpha = 0$. Thus, whatever the value of τ may be, τ does not have any influence on the dynamics. Here and elsewhere, the communities we look at will always include 50 truth-seekers, unless indicated otherwise.[6] For the first experiment, we consider a variable number of campaigners, that is, agents who do not attend to any other opinion (or to reality) but stick to some fixed opinion. The questions to be asked concern the ability of the campaigners to convince others, and the extent to which this depends on the value of ε and on the number of campaigners present in a community. We run simulations in which the truth-seekers begin with a random opinion, meaning that, for each agent individually, the initial opinion is drawn randomly from U(0, 1).[7] All simulations will run till a point of very low rate of change is reached, which we formally define to be the update u such that $\left|x_i(u) - x_i(u+1)\right| \leqslant 10^{-5}$ for all i, and which below we often refer to as fixed point (even if technically it need not be a fixed point).

Figure 2 shows, for each combination of number of campaigners (going from 1 to 50, in steps of 1) and value of ε (going from 0.01 to 0.5, in steps of 0.01), the average number of others that have come to adopt the campaigners' opinion in the fixed point (meaning, more specifically, that their opinion was within 10^{-3} from ϱ). It is clearly seen that how many others get converted, on average, depends on the number of campaigners present in the community and on how liberal the others are in counting others as their peers, as well as on what exactly the campaigners' opinion is.

We might already seem to face a puzzling phenomenon here. If we look along the vertical line $\varepsilon = 0.3$ in the left panel of Figure 2, we see that conversion is *more* successful with *fewer* campaigners present than with *more* of them present, where one might have expected to find the opposite. On

[6] Although, in the present case, where α is set to 0, it might be more appropriate to use shudder quotes and say that the community comprises 50 "truth-seekers" (rather than truth-seekers), given that these agents are not actually assigning any weight to τ.

[7] Under certain circumstances, the BC dynamics can cause numerical problems because of the constitutive queries $\left|x_i(u) - x_j(u)\right| \leqslant \varepsilon$ that have to be answered to get the index set $X_i(u)$. Since the floating point arithmetic only approximates real numbers, the given answers can easily go wrong if an agent's opinion is exactly on the border of another agent's BCI. Hegselmann and Krause (2015, pp. 493 ff) demonstrate how severe the resulting numerical errors can be: the dynamics often gets numerically corrupted already in the first updating step. However, in their paper the numerical dangers are the consequence of a specific equidistant starting distribution together with a certain set of ε values that can easily lead to situations where one agent's opinion is exactly on the border of another agent's BCI. As a protective measure, Hegselmann and Krause (2015) use an equidistant starting distribution in which such situations do not occur. (Another protective measure would be to simply add a tiny value, for instance 10^{-12}, to ε.) In this paper, we do not use equidistant starting distributions; our starting distributions are always *random*. For a random start distribution, the probability that one agent's opinion is exactly on the border of another agent's BCI is almost surely zero.

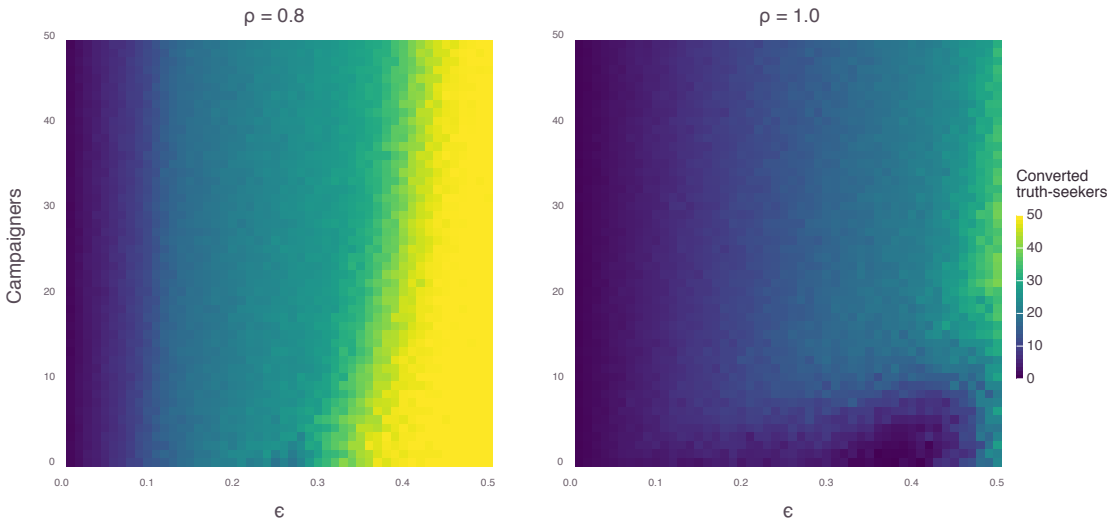

Figure 2: For campaigners holding an opinion of 0.8 (left) and 1 (right), showing the average number of truth-seekers that have been converted in the fixed point (i.e., whose opinion is within 10^{-3} from ϱ in the fixed point), for number of campaigners increasing from 1 to 50, and ε increasing from 0 to 0.5, in increments of 0.01. Truth-seekers start with opinions drawn from U(0, 1). See the text for details.

closer inspection, however, the phenomenon is easily understandable. It is due to the fact that the more campaigners there are, the greater the pull their opinion ϱ exerts on the truth-seekers in their vicinity, and the greater the chance there will occur an early split among the truth-seekers. With fewer campaigners, truth-seekers in their vicinity are also pulled in their direction, but not as strongly, whence it is easier for other truth-seekers to catch on. And by catching on, these others can also come under the influence of the campaigners and thereby eventually end up believing ϱ. By contrast, if an early split occurs, some truth-seekers may forever remain out of the reach of the campaigners. The single runs shown in Figure 3, one featuring eight campaigners (left panel), the other fifty (right panel), illustrate this phenomenon. (Note that convergence takes much longer for the case with fewer campaigners than for that with more.)

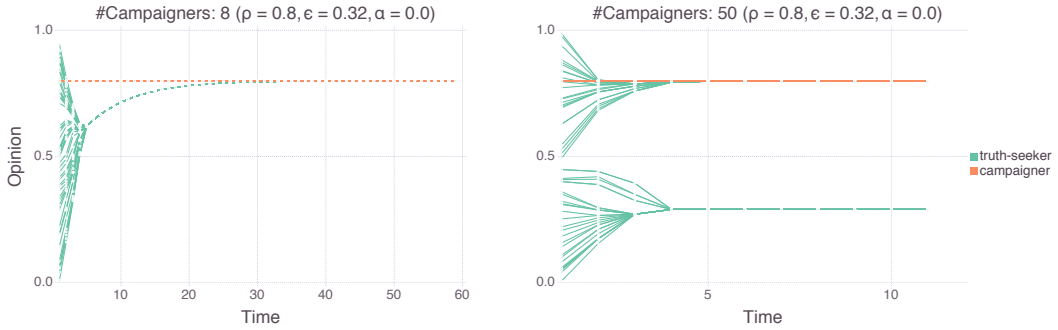

Figure 3: Illustration of why smaller numbers of campaigners may be able to convert more truth-seekers. See the text for further explanation.

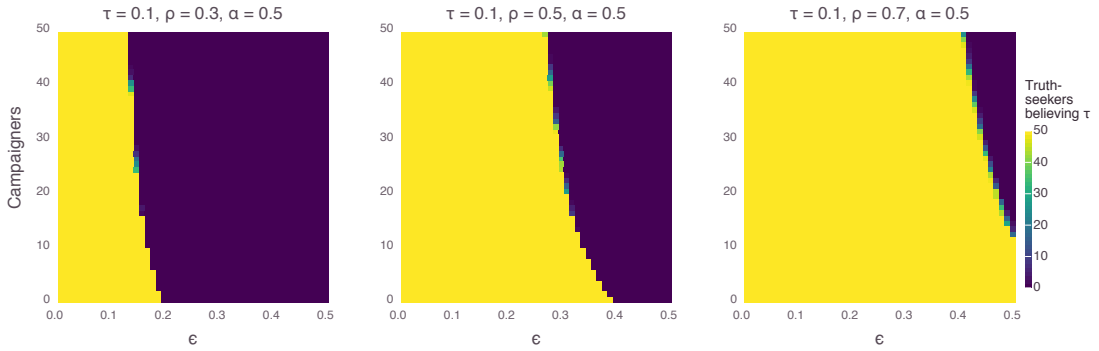

Figure 4: Average number of truth-seekers believing the truth in the fixed point (i.e., whose opinion is within 10^{-3} from τ in the fixed point), where the average is over 100 simulations per combination of number of campaigners (from 1 to 50) and ε (from 0 to 0.5, in steps of 0.01).

Now let us have a look at the more interesting kind of case in which α is *not* equal to 0 and so the truth-seekers also base their updates on evidence coming from the world. We want to ask two questions: (i) Supposing given values for τ and α, and letting ε vary as in the above simulations, what is the ability of the campaigners to convince truth-seekers of their fixed opinion (i.e., what is the success rate of campaigning)? The foregoing result indicates that this may depend on what that opinion is and also on how many campaigners there are in the community, so we consider different values of ϱ and let number of campaigners vary as before. (ii) Given the same suppositions, what is the ability of the campaigners to lure away the truth-seekers from τ? Note that the first question directly bears on the possibility of leading a successful *mis*information campaign—a campaign to make the public believe some falsehood—while the second is relevant to *dis*information campaigns, that is, campaigns meant to steer the public away from the truth (but not caring about what the public comes to believe, as long as it is not the truth).

Figure 4 indicates an answer to the second question. It shows, averaged over 100 simulations for each combination of number of campaigners and ε value, the number of truth-seekers whose opinion equals τ in the fixed point (i.e., whose opinion was within 10^{-3} from τ).[8] The results are for three values of ϱ only, but the trend is manifest. The rate at which truth-seekers have been effectively disinformed in the fixed point depends on both ε and the number of campaigners, but most importantly it depends on how far from the truth the campaigners' opinion is. We see that, from the perspective of the campaigners, supposing their goal is disinformation, it pays to be subtle. This will be a recurring theme.

We are not showing any figures related to the first question, simply because there is not much of interest to be shown: for the aforementioned parameter settings, *none* of the truth-seekers get converted to the campaigners' opinion. It is in fact only when the campaigners get *very* subtle—for instance, by holding an opinion of 0.11 (where τ is still 0.1)—and when we lower α quite a bit, like to .05, that conversion starts to occur, but, depending on exactly how close to 0 we set α, this may still only occur if the number of campaigners present in the community is near its maximum. Under circumstances

[8] To forestall misunderstanding, note that Figures 2 and 4 use the same color scheme to indicate counts of truth-seekers, but in the former these are counts of truth-seekers believing ϱ in the fixed point while in the latter they are counts of truth-seekers believing τ in the fixed point.

that are just barely less extreme, truth-seekers, while massively lured away from the truth, are still enough in touch with the world to not become misinformed.

So far, we have only considered communities that consisted of truth-seekers and campaigners. But in addition we ran simulations with communities that also featured fixed numbers of free riders, agents who are not dogmatic but who update strictly by averaging the opinions of those within their BCI. More exactly, we reran three times the simulations whose results are visualized in Figure 4, once with 10 free riders added, once with 25, and once with 50. The question we were interested in was what effect the presence of those numbers of free riders had on the number of truth-seekers that ended up, in the fixed point, believing the truth.

Figure 5 brings out the effects for the nine kinds of situations (the three different values for ϱ considered in the earlier simulations times the three different numbers of free riders) by plotting the results, for any combination of number of campaigners present and ε value, of subtracting the average number of truth-seekers that hold the truth in the communities *with* free riders from the average number of truth-seekers that hold the truth in the communities *without* free riders (both averages taken over 100 simulations per combination of campaigners and ε value). Note that, in principle, this can yield results from −50 to 50. In fact, however, the results were never negative—the truth-seekers in communities *without* free riders did always at least as good as the truth-seekers in the communities *with* free riders—and were for various combinations of parameter values even positive, typically at or very close to the maximum of 50, meaning that for various combinations of parameter settings, the presence of free riders has a clear negative epistemic effect on the truth-seekers. It is remarkable that the results are largely insensitive to how *many* free riders there are: the risk brought about by 50 free riders seems only slightly greater than that brought about by just 10.

To further clarify the results: the yellow slivers in Figure 5 indicate combinations of parameter settings for which truth-seekers that tend to *arrive* at the truth in the fixed point when *no* free riders are present are *diverted* from the truth in the fixed point when free riders *are* present. In other words, in the yellow areas, the presence of free riders helps to bring about the success of disinformation attempts that would otherwise have faltered.

The patterns in the various plots may at first appear mysterious, but in fact they are easy to explain. As Figure 6 illustrates, depending on the value of ε, the campaigners may be able to hold all or some free riders hostage, not necessarily in the sense that those free riders side with the campaigners (i.e., adopt ϱ as their opinion), but at least in the sense that their opinions remain under the influence of the campaigners' fixed opinion. In turn, and again depending on ε, the free riders may retain an influence on the truth-seekers' opinions, at least enough so to keep the latter from believing the truth.

To go a bit more into the details, all four panels in Figure 6 show outcomes of runs with 50 truth-seekers, 50 free riders, and 30 campaigners, and with $\tau = 0.1$, $\varrho = 0.7$, and $\alpha = 0.5$. The runs differed in the value of ε that was assumed. In the upper-left panel, where $\varepsilon = 0.2$, the campaigners' influence reaches down to 0.5, but not lower. And we see that free riders starting with an opinion below 0.5 rapidly come to side with the truth-seekers, who all end up believing the truth, not held back by any of the free riders. That is very different in the upper-right panel. Because there $\varepsilon = 0.4$, the campaigners' influence stretches down all the way to 0.3, as a result of which the free riders remain torn between the campaigners' opinion and the opinion the truth-seekers quickly come to converge on. Unfortunately for the latter, because the free riders remain within their BCI, they are stuck with an opinion that is not quite the truth. And then, as the bottom row of Figure 6 shows, there are also some values of ε for which it is more or less a toss-up (depending on the random distribution of opinions at the

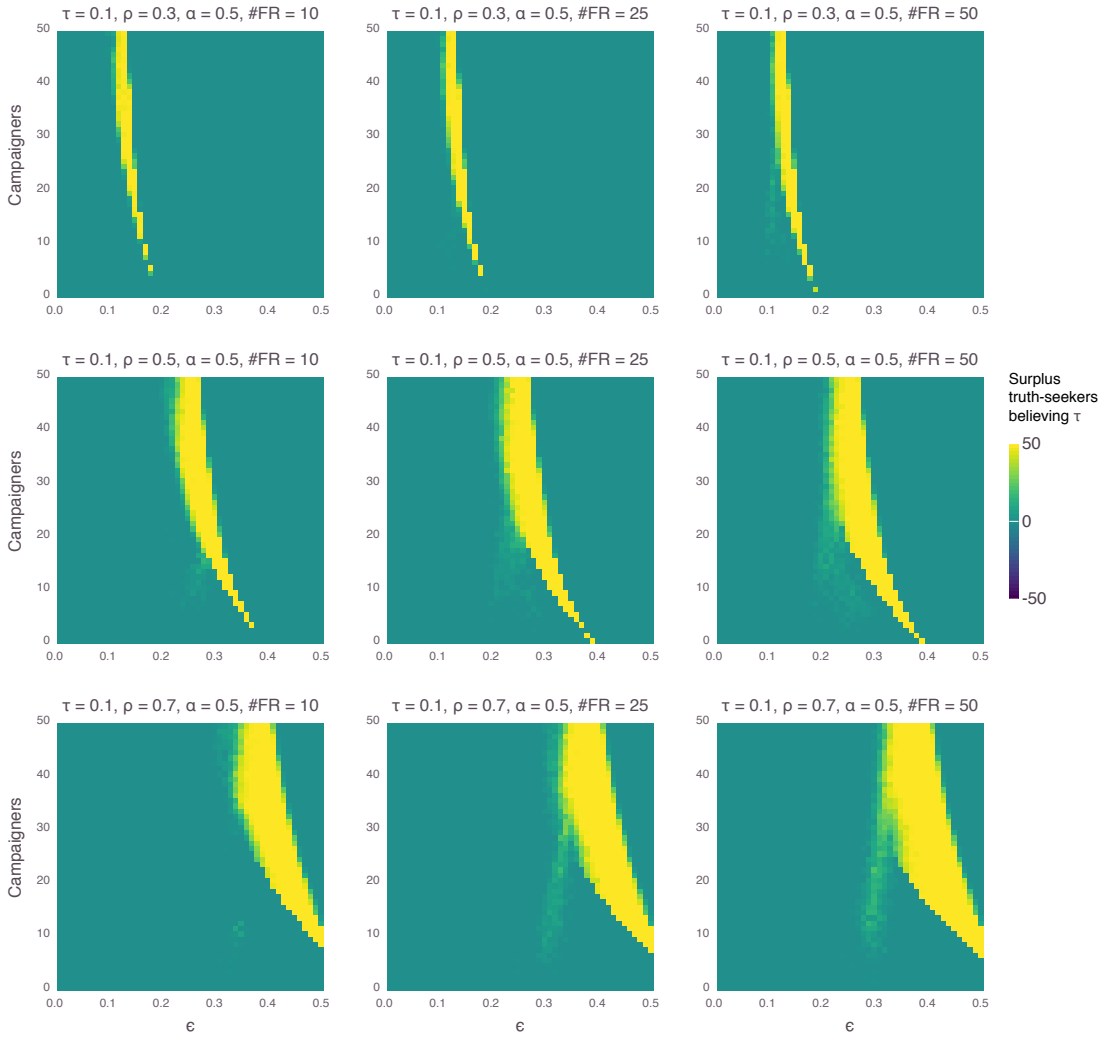

Figure 5: Average number of truth-seekers believing the truth in the fixed point (averaged over 100 simulations per combination of number of campaigners and ε value) with 10 (left column), 25 (middle column), and 50 (right column) free riders present subtracted from average number of truth-seekers believing the truth when *no* free riders are present (again averaged over 100 simulations per combination of number of campaigners and ε value), with rows corresponding to different settings for τ, ϱ, and ε. In particular, yellow areas indicate combinations of number of campaigners and ε value for which all truth-seekers end up believing the truth when no free riders are present and none of them end up believing the truth when the indicated numbers of free riders are present.

start) whether or not any free riders will be captured by the campaigners' opinion—and hence also whether the truth-seekers will remain under the influence of opinions that are partly influenced by the campaigners' opinion.[9]

[9] Not shown here, but easy to understand, is the fact that, as seen in Figure 5, the number of free riders present in a run has hardly any impact on the result. In general, in those situations in which free riders do keep truth-seekers from believing the truth, their number has some influence on how far from the truth the truth-seekers' opinions end up: they are closer to the truth with fewer free riders present. (But note that Figure 5 only registers *whether* the truth-seekers ended up believing

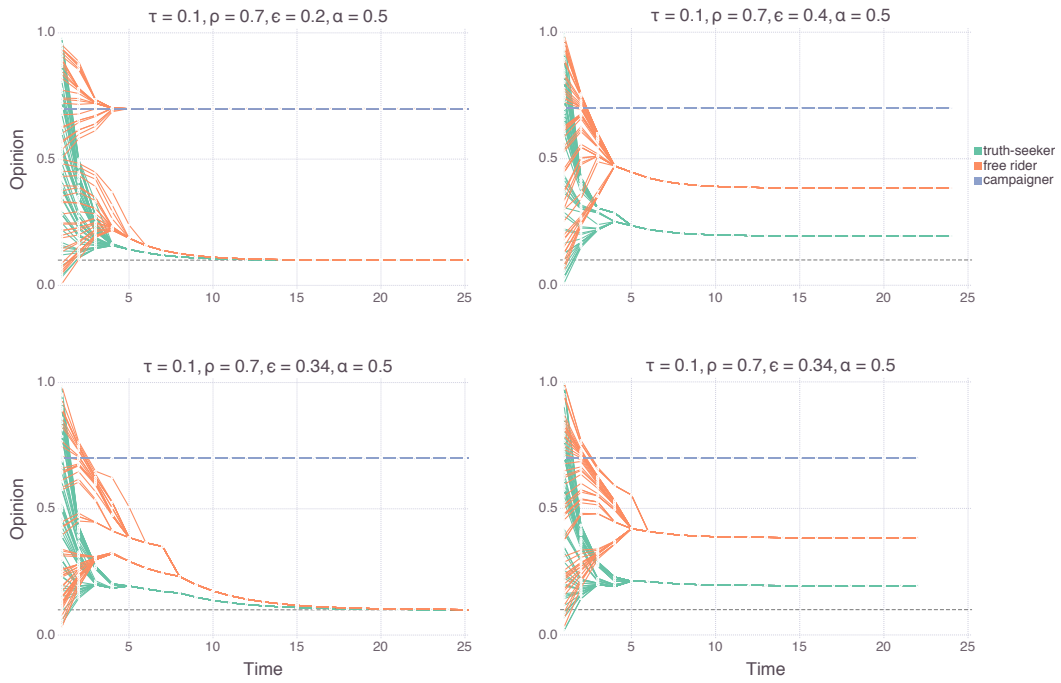

Figure 6: Four runs with mixed communities, the top row illustrating the finding that whether free riders keep truth-seekers from arriving at the truth depends on the value of ε. The bottom row shows two runs with identical parameter settings, illustrating the finding that, for some values of ε, truth-seekers sometimes will, and sometimes will not, reach their goal.

As for the more general lesson to be learned from Figure 5, we see that even though for most of the parameter settings—all those covered by green in the figure—the presence of free riders makes no difference as far as our current criterion goes, there are still quite a number of parameter settings where they do have a big impact. This means that, while they may have no interest in furthering the campaigners' cause, free riders at least take the risk of doing so. Of course, this will not hold if free riders can somehow rule out in advance that any of the situations in which they do harm obtains. But it is not part of any BC model that agents *know* what their peers' BCIs are (even though, in the models studied so far, all agents have the same BCI; this will no longer be true in an extension to be considered shortly). Hence, in view of the results shown in Figure 5, it seems reasonable to conclude that free riders are to some extent (possibly unwittingly) furthering the campaigners' cause.

Below, we will be more precise about what damage free riders can do. Before turning to that question, however, we want to briefly mention a variation of the above simulations that we will not explore in any depth here, but that interested readers may want to investigate using the code in the Supplementary Materials. So far, we kept one parameter constant, viz., number of truth-seekers (which was always 50). Instead, one could consider a setup in which the total numbers of agents—truth-seekers + free riders + campaigners—is kept constant, and in which we go through all combinations of numbers

the truth, not *how far* they were from the truth. We canvass the impact free riders have on how far from the truth the truth-seekers end up in Sect. 3.3.) In the kind of borderline situations depicted in the bottom row of Figure 6, number of free riders present has some impact on whether a run does or does not end with the truth-seekers believing the truth.

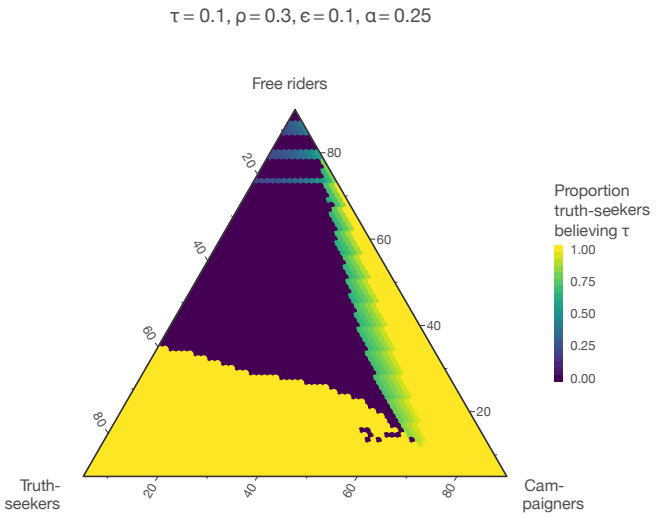

Figure 7: Ternary plot showing proportion of truth-seekers believing the truth in the fixed point, for a community of 90 agents, considering all possible combinations of numbers of agents from the three types.

of the three types of agents yielding that constant. The question about conversion (and also questions to be addressed in the remainder of this paper) could then be asked for each of those combinations.

Figure 7 gives the results for just one setting of τ, ϱ, ε, and α, and with total number of agents kept at 90. This figure is a so-called ternary plot. For any point represented in the plot, the proportion of agents of the type associated with a given vertex is the shortest distance from the point to the edge opposite the vertex, divided by the sum of the lengths of the shortest distances from the point to the various edges.[10] So, for instance, the geometric center of the plot corresponds to a population of agents in which all three types are equally represented. Each point in the plot shows the outcome of one simulation for the corresponding combination of agents.

The figure reveals a somewhat intricate pattern. At the same time it is clear, however, that the relative number of free riders in the population is decisive for whether the truth-seekers reach their goal. Where there are relatively few of the former, the truth-seekers arrive at the truth even when vastly outnumbered by campaigners. By contrast, where free riders constitute a sizable portion of the population, they keep the truth-seekers away from the truth *even if there are only very few campaigners in the population*. While we leave a further analysis of this result, and more generally the pursuit of this variant of our previous simulations, for future research, it is still worth noting how even these preliminary results underscore the above conclusion about the negative role of free riders.

3.3 Measuring societal costs

The truth-seekers have a clear goal, to wit, finding out the truth. If campaigners can make it harder or even impossible for truth-seekers to reach that goal, that could be said to constitute societal damage. We propose to measure the amount of damage being done by campaigners in terms of the *accuracy* of

[10] Needless to say, only points corresponding to triples of integer values are occupied.

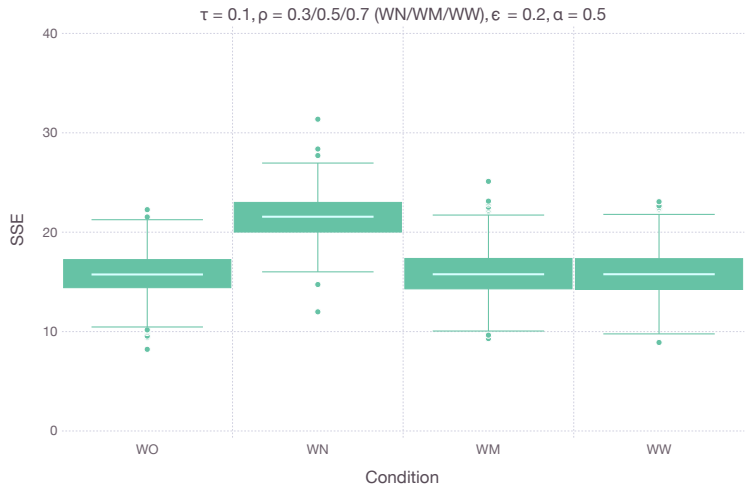

Figure 8: Box plot showing results from 1,000 simulations for each of four conditions, each simulation measuring the total sum of squared distances from the truth for all truth-seekers at each of the 50 time step. In the WO condition, there are only truth-seekers; in the other conditions there are also 25 campaigners, holding an opinion close to (WN condition), neither close to nor distant from (WM condition), and distant from (WW condition) the truth.

the opinions of the truth-seekers, more specifically by asking how much less accurate (if at all) they are when their society includes campaigners, as compared to a situation in which campaigners are entirely absent. To be still more specific, we express inaccuracy as the sum of squared errors over a given number of agents and a given number of time steps (i.e., updates). So, where the truth is given as τ and agent x_i's opinion at time t is $x_i(t)$, her inaccuracy at t equals $\big(\tau - x_i(t)\big)^2$. Given a community whose truth-seekers are $\{x_1, \ldots, x_n\}$, who update their opinions at m consecutive time steps t_j, the total inaccuracy of that community, taken over the said time steps, equals $\sum_{i=1}^{n} \sum_{j=1}^{m} \big(\tau - x_i(t_j)\big)^2$.

The following example illustrates the kind of effect the presence of campaigners can have on the accuracy of a community:

Example 3.1 We compare four different situations featuring communities of agents who interact according to the BC model extended to incorporate campaigners, and who receive input from the world. In all, there are 50 truth-seekers, who, starting with an opinion drawn randomly per agent, are going to update 50 times.[11] Also in all situations, $\tau = 0.1$, $\varepsilon = 0.2$, and $\alpha = 0.5$. In the first situation, the community consists strictly of truth-seekers. In the other three, there are also 25 campaigners, the difference between these three communities being that in one, those campaigners hold an opinion that is relatively close to τ, while in the second, the campaigners hold an opinion further removed from τ, and in the third, the campaigners hold an opinion far removed from τ; specifically, the campaigners in these communities stick to the opinions 0.3, 0.5, and 0.7, respectively.

[11] Earlier, we took convergence of opinions as a stopping point for our simulations. That is not a good idea if we want to make accuracy comparisons. Else, a community quickly converging to an opinion far from the truth might still compare favorably with a community converging slowly to the truth, simply because we would, in the latter case, be summing over a greater number of updates. Thus, to produce meaningful comparisons of the sort we are interested here, we fix the number of updates.

The two questions now are whether the presence of campaigners makes a difference to the accuracy of the truth-seekers, and whether it matters how far from the truth the campaigners' opinion is. To answer these questions, we ran 1,000 simulations for each of the four communities and measured in each simulation the total inaccuracy (the sum of squared errors, or SSE) of the truth-seekers. Figure 8 graphically represents the outcomes from the simulations; WO is the condition without campaigners, and WN, WM, and WW are the conditions with campaigners and with, respectively, a narrow, moderate, and wide gap between τ and ϱ. It appears that the truth-seekers in the second condition are in general less accurate than the ones in the other conditions, while there appear to be only minor differences between the results from the WO, WM, and WW conditions.

These visual impressions were confirmed by conducting a one-way ANOVA, with SSE as dependent variable and condition as independent variable. The ANOVA showed the effect of condition to be large and highly significant: $F(3, 3996) = 1661$, $p < .0001$, $\eta^2 = 0.55$. A Tukey post hoc test revealed that the mean of the WN condition (21.54, ±2.13) differed significantly from the mean of the WO condition (15.78, ±2.18) as well as from the mean of the WM (15.85, ±2.33) and WW (15.83, ±2.23) conditions, all at $p < .0001$. None of the other means differed significantly from one another (all ps $> .9$).

Thus, the answer to our first question is that the presence of campaigners can come at a significant societal cost, even if it is not guaranteed to do so. Remarkably, campaigners, at least in the case at hand, compromise the accuracy of their truth-seeking community members the most if they are "subtle." At first blush, one might expect greater extremism on the part of the campaigners to do more damage to the truth-seekers, in which case the fact that the extremists did not significantly affect the accuracy of the truth-seekers at all may come as a surprise. On more careful consideration, however, it makes a lot of sense that blatant lying, or at least spreading blatant falsehoods, is going to be ineffective, simply because the blatancy of the falsehoods makes those spreading them more easily identifiable as agents whose opinions are to be discounted. ♦

This is only an example, but the finding that campaigners are damaging only if they are subtle is quite robust, holding across a range of combinations of parameter settings. Naturally, when we set α to some very low value, so that the influence of worldly input becomes negligible relative to the social part of what goes into the updating, the presence of subtle campaigners will tend to *increase* accuracy, simply because they help steer the truth-seekers' opinions to a value at least close to the truth, where without the campaigners' influence, the truth-seekers' opinions would be all over the place, and so more likely to be further removed from the truth. That being subtle is, under a broad range of conditions, the better strategy from the campaigners' perspective is a finding we encountered before, and later on we shall encounter it again.

Above, we looked at the impact that free riders could have on the conversion rate. Let us also look if there is, or may be, any societal cost associated with their presence.

Example 3.2 We saw that, for the conversion rate, the number of free riders present did not seem to matter much. For most combinations of number of campaigners and ε values, *if* the presence of free riders had an effect on conversion, it did so whether there were as many free riders as truth-seekers or only a small minority of free riders. It does not follow that the number of free riders will not matter to societal cost, which we understand as the overall increase in inaccuracy in the truth-seekers' opinions that the free riders bring about.[12]

[12] Although in note 9 we already hinted that free riders can have an effect on inaccuracy.

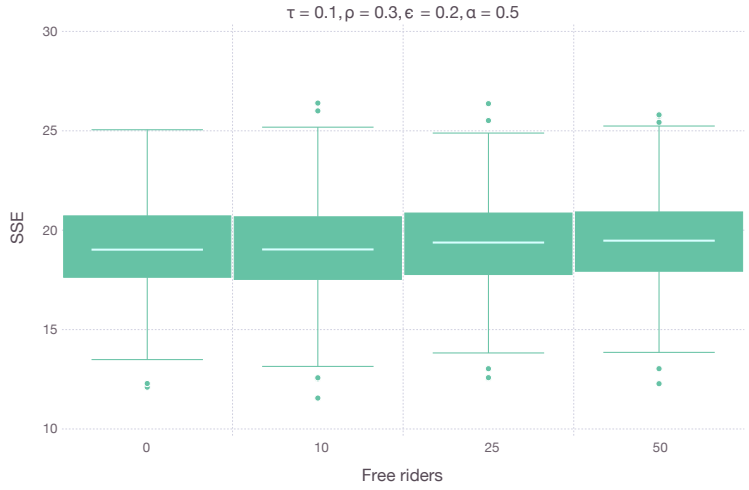

Figure 9: Same as for Figure 8, except that now the conditions correspond to different numbers of free riders being present.

Thus, consider four communities, all of which contain 50 truth-seekers and 15 campaigners, but the first of which contains only those agents, the second contains also 10 free riders, the third 25 free riders, and the fourth 50. Moreover, the following parameter settings hold in all of them: $\tau = 0.1$, $\varrho = 0.3$, $\varepsilon = 0.2$, and $\alpha = 0.5$. We are measuring again the total sum of squared errors over 50 updates, running 1,000 simulations per community.

A visual comparison of the results is given in Figure 9. We discern a slight upward trend in SSEs as the number of free riders increases, although it is not immediately obvious from the figure whether the impact is significant. An ANOVA reveals that it is: $F(3, 3996) = 5.63$, $p = .0008$. Conducting a Tukey post hoc test shows that the mean of the community with 50 free riders (19.45, ±2.24) is reliably higher than the mean of the community without free riders (19.13, ±2.23) and the mean of the community with 10 free riders (19.09, ±2.26); both ps $< .01$. There were no further statistically significant differences in means.

It hence appears that, even in a community with relatively few campaigners, free riders can make the truth-seekers significantly less accurate. On the other hand, as the figure suggests, and as is confirmed by looking at the effect size associated with the just-reported ANOVA—$\eta^2 = 0.004$, which conventionally counts as small—the impact of the free riders is, for all we know so far, very limited.

But this turns out to be highly sensitive to the setting of the parameters. Rerunning all of the above, but now for $\varepsilon = 0.3$ and $\alpha = 0.25$, yields a very different picture, as already emerges from Figure 10. Not surprisingly, this time an ANOVA showed the impact of the number of free riders in a community to be highly significant—$F(3, 3996) = 312.5$, $p < .0001$—with a Tukey post hoc test showing the means of all four conditions (38.45, ±3.54; 39.13, ±3.60; 40.49, ±3.85; 43.18, ±3.94) to differ significantly from each other; all ps $< .0001$. It is equally unsurprising that we find a large effect size: $\eta^2 = 0.19$. ♦

Free riders, even if there were no chance that they impacted the accuracy of the truth-seekers, are morally blameworthy for the very general reason that they reap the benefits of the work done by others while not making any contribution from which others could benefit in return. But Example 3.2 shows

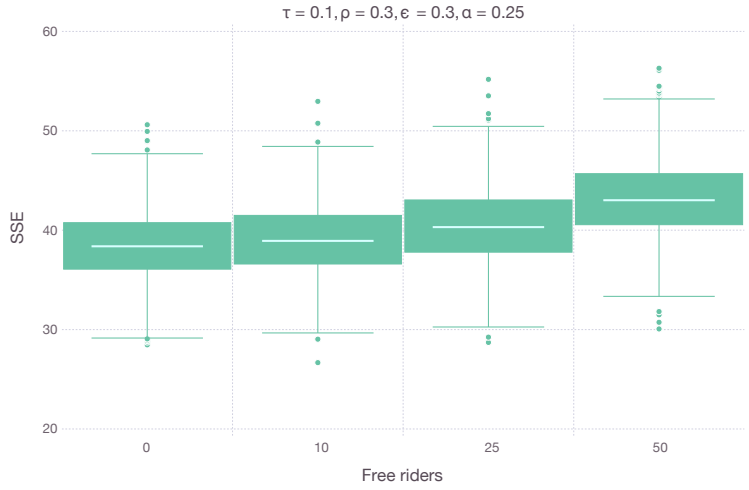

Figure 10: Same as Figure 9 except that values for ε and α are different.

that, in the area of opinion dynamics, there is an additional reason, namely, precisely the fact that free riders *can* significantly and greatly compromise the accuracy of the opinions of those willing to directly seek information about the world; though a comparison of the two above cases suggests that the free riders will have more of an impact the more the truth-seekers, in forming new opinions, *also* rely on others—which is in effect precisely what one would expect.

4 Adding confidence dynamics

In the original BC model as well as in all known extensions of it, the BCI as determined by the value of ε is itself fixed, and not the subject of any updating mechanism.[13] Realistically speaking, however, it is quite reasonable to assume that people may want to adjust their broad- or narrow-mindedness in counting others as peers, if perhaps just by imitation. Being around broad-minded or trusting people, who are willing to take into account a great variety of opinions, one may decide to become, or may perhaps unconsciously become, more broad-minded oneself; analogously if one is around narrow-minded or suspicious people. Indeed, we may all be uncertain to some extent about how liberal we should be in listening to others, and we may well let ourselves be guided in this respect by whoever we recognize as our current peers.

Moreover, adjusting our BCI may be a defense mechanism against the efforts of the campaigners to convert us to their opinion or at least divert us from the truth. After all, campaigners will not let *anyone* influence their opinion, so they effectively have an ε of 0. Being in the neighborhood of such people may make us more selective as regards deeming others worthy of influencing *our* opinion. But that also offers some protection against the campaigners' influence, simply because they are less likely to be in our peer group. Naturally, for reasons already pointed out, we must guard against becoming

[13] Hegselmann and Krause (2005) introduce a type of BCI that is randomly chosen (with an upper bound) at each time step, individually per agent. They analyze the effects of such random BCIs for several types of averaging and compare them to the effects of fixed BCIs, also looking at the effect of different types of averaging (e.g., taking the arithmetic, geometric, or power mean).

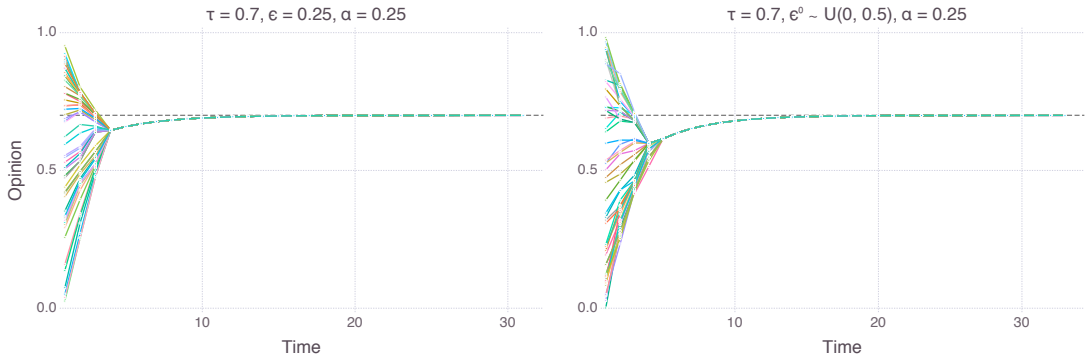

Figure 11: Repeated updating until a fixed point is reached, the left panel showing the updates in the original BC model, the right panel showing the updates in an extension of that model with confidence dynamics.

too selective, as that would also annihilate any positive effects social learning can have. The goal must be to find the *right* level of caution.

For these reasons, we assume that, just as averaging our peers' opinions may, to some extent, guide us in forming our own opinion, averaging their BCIs may guide us in determining our own BCI, and hence in setting our standards for peerhood. Specifically, we introduce a mechanism formally capturing the idea that the BCI may vary from one agent to another and also from one time step to another, and that it may undergo a systematic influence from one's peers. This mechanism implements in the model what we call *confidence dynamics* (CD).[14] The extension amounts to revising the definition of the set of agents within agent x_i's BCI after the u-th update, which previously was given by

$$X_i(u) \;:=\; \{j\colon |x_i(u) - x_j(u)| \leqslant \varepsilon\}.$$

We redefine this as follows:

$$X_i(u) \;:=\; \{j\colon |x_i(u) - x_j(u)| \leqslant \varepsilon_i^u\},$$

with ε_i^0 randomly drawn from U(0, $\hat{\varepsilon}$)—where we may want to specify $\hat{\varepsilon}$ per situation—and, for all $u \geqslant 1$,

$$\varepsilon_i^{u+1} \;:=\; \frac{1}{|X_i(u)|} \sum_{j \in X_i(u)} \varepsilon_j^u.$$

Thus, an agent's current peers determine both her new opinion and her new BCI, where the agents start with a BCI determined by picking, randomly per agent, a real number in an interval ranging from 0 to some contextually set upper bound $\hat{\varepsilon}$.

To develop some initial feeling for this further extension, one can compare simulations with and without CD in a community of only truth-seekers. The left plot in Figure 11 shows the results for running the standard BC model till a fixed point is reached, for the parameter setting with $\tau = 0.7$, $\varepsilon = 0.25$, and $\alpha = 0.25$. The right plot shows for those same values of τ and α the results *with* CD, in particular, where each agent x_i starts with a BCI determined by a randomly picked value for ε_i^0, with $\varepsilon_i^0 \sim$ U(0, 0.5), and then updates her BCI at each time step, in the way just explained.

[14] The idea of confidence dynamics is already to be found in Hegselmann (2014). There, however, all "non-campaigners" start with the same BCI, after which their BCIs shrink more or less, depending on the extent to which they are influenced by the campaigners (which have a BCI of 0).

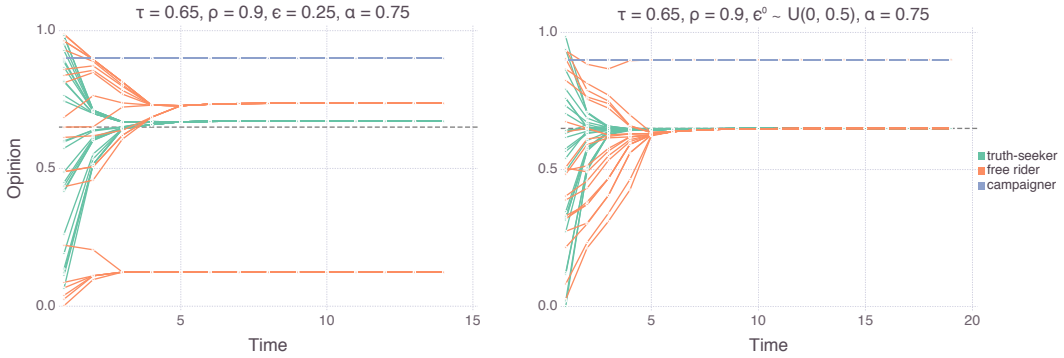

Figure 12: Illustration of confidence dynamics in a community with typed agents; updates are shown until the fixed point: confidence dynamics is absent in the left panel, present in the right.

In the figure, it appears that consensus occurs slightly faster *without* CD, but it is more interesting to know whether this dynamics has any effect on accuracy, again measured in terms of SSEs. A grid search we conducted did not yield a single parameter setting for which there was a significant difference in accuracy between updating without and updating with CD. Clearly, however, what we really want to know is whether there is any effect of CD in (BCC) or (BCCF).

We consider right away communities made up of all three types of agents: truth-seekers, who also attend to incoming evidence; campaigners, who stick to one and the same opinion from the start, ignoring any evidence as well as the opinions of those within their BCI; and free riders, who ignore evidence but do take into account the opinions of the agents in their BCI. We are again interested in both conversion rates and societal costs.

4.1 Conversion

We saw in the previous section that campaigners can keep truth-seekers from believing the truth. We look at the difference CD may be able to make with regard to this.

Example 4.1 Figure 12 gives a first impression of the kind of impact that CD can have in a community with all three types of agents, specifically, here, in a community consisting of 25 truth-seekers, 20 free riders, and 10 campaigners, where $\tau = 0.65$, $\varrho = 0.9$, and $\alpha = 0.75$.

If we imagine that the campaigners are trying to lure away from the truth as many of the others as possible, then we see that they are entirely successful in the situation depicted in the left panel of Figure 12, in which CD is absent: literally no one ends up believing the truth. All truth-seekers do end up believing something that could be said to be close to the truth, but none of them exactly hits the mark—which is especially disconcerting given that they attach three times as much weight in their updating to the worldly part as to the social part. Moreover, the free riders are not even close to τ in this situation.

By contrast, in the right panel of the same figure, where CD *is* assumed, not only do all truth-seekers end up believing the truth, but so does a majority of the free riders. It is also to be noted, however, that in this situation there is a group of free riders who end up believing the falsehood spread by the campaigners. If the latter are meaning to run a misinformation campaign, then they fail completely in the first situation but at least partly succeed in the second. ♦

This example already hints at what will be one of the conclusions of our paper, namely, that from a social-engineering perspective, it is hard to give general recommendations concerning opinion dynamics (broadly speaking, so including CD). After all, so much depends on context—such as, in the example, what the goal of the evildoers is: are they bent on having the public *not* believe the truth or rather on having it believe the falsehood the evildoers are propagating? In Section 2, we gave the example of a politician just trying to divert from the truth voters tending toward her opponent, and not necessarily to make them believe whatever lies she is propagating. In view of Example 4.1, such a politician might actually be happy with CD being operative. On the other hand, the Brexiteers who seriously tried to convince the British electorate that Leave was their best option may not have been helped by CD (supposing there was any among British citizens in the months preceding the referendum).[15]

But let us look more systematically at the effects CD may have. Previously, we saw that the distance between τ and ϱ matters a lot to how much damage the campaigners can do, epistemically speaking; in particular, it was seen that they are able to do more damage by choosing ϱ relatively close to τ. Will that still be true if the possibility of CD is taken into account?

A comparison between Figures 13 and 14 suggests a positive answer. These figures show, for various combinations of numbers of free riders and campaigners, the proportions of truth-seekers ending up believing the truth (top row in both figures) and proportions of free riders ending up believing the truth (bottom row in both figures), where these proportions are averages over 100 simulations per number-of-campaigners–number-of-free-riders combination, and where the truth-seekers and free riders always begin with a random initial opinion. The first figure shows these results for the case where there is a rather moderate gap of 0.1 between τ and ϱ; the second figure shows the parallel results for the case where there is a relatively large gap of 0.25 between τ and ϱ. The comparison suggests that which of lying a lot and lying a little is better may entirely depend on whether CD is operative. If it is not, then by being more moderate one will keep more truth-seekers and free riders from believing the truth, while if CD *is* present, then as far as truth-seekers ending up believing the truth is concerned, there is not much difference between lying a little and lying a lot: neither is very effective in that case. As far as free riders are concerned, lying a little seems to be more effective, although the difference it makes is small.

Now let us look at the situation from the perspective of the truth-seekers. Figures 13 and 14 show that they need not concern themselves much with how many free riders or campaigners there are in their community; these numbers do not appear to matter a whole lot. By contrast, CD *can* make all the difference, depending on how far from the truth the misinformation being spread is. Indeed, Figure 13 shows just how dramatic the difference can be when the misinformation is relatively close to the truth: then virtually all truth-seekers (and also a good portion of the free riders) end up believing the truth if CD is assumed, while virtually no one ends up believing the truth if there is no CD. When there is a relatively wide gap between τ and ϱ, CD hardly makes a difference, as Figure 14 shows; here, the community appears to be even slightly better off without CD.

These results bear on *dis*information campaigns. As for *mis*information campaigns, the situation for truth-seekers is similar to the one described on page 11, where we found that, although campaigners were, under certain circumstances, able to block the truth-seekers from reaching the truth, they were in general unable to convince the truth-seekers of their fixed opinion. The same turns out to be

[15] That the Leave camp won does not mean there was no CD. Example 4.1 only suggests that, if there was, the Leave camp was not helped by it.

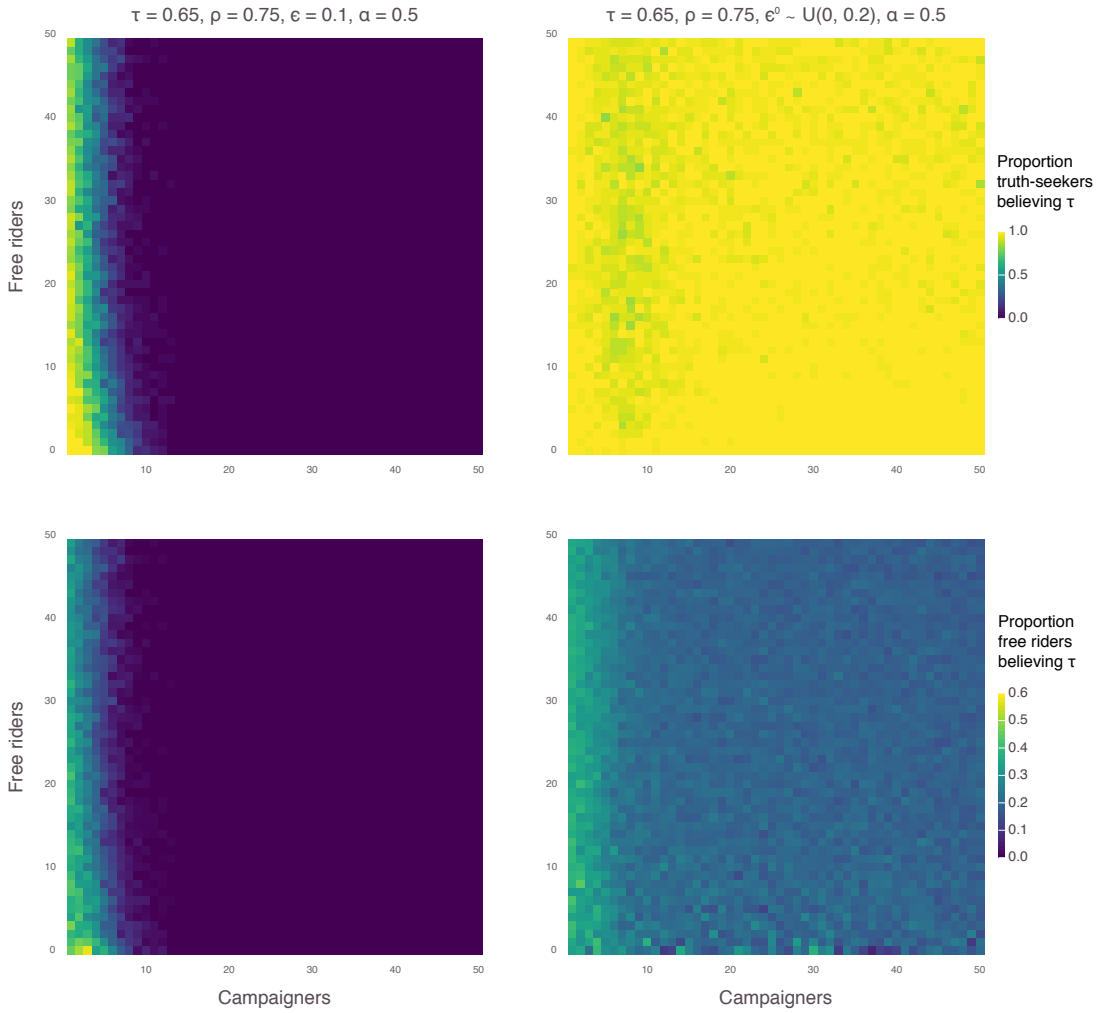

Figure 13: Proportions of truth-seekers (top row) and free riders (bottom row) believing the truth in the fixed point, the left column showing results for communities without and the right column showing results for communities with confidence dynamics. Results are averages over 100 simulations per number-of-campaigners–number-of-free-riders combination. (Color keys apply per row.)

true here, and this irrespective of whether CD is assumed. Indeed, we found that, barring extreme conditions—conditions in which ϱ is very close to τ and/or α very close to 0—conversion of truth-seekers simply does not occur.

For free riders, the picture *can* look very different, in that, depending on the parameter setting (i.e., combination of values for τ, ϱ, ε, and α), CD can have a big impact, but not necessarily a positive one. Figure 15 shows, for three different parameter settings, the effect of CD on the proportion of free riders that side with the campaigners in the fixed point, where the communities consist of 50 truth-seekers and of numbers of campaigners and of free riders that both vary from 1 to 50.

As is clear from the figure, for two of the three parameter settings there is a greater chance for free riders to end up with the campaigners in a community *with* CD than in one without it; for the parameter setting corresponding to the middle row in the figure, adding CD makes hardly any difference.

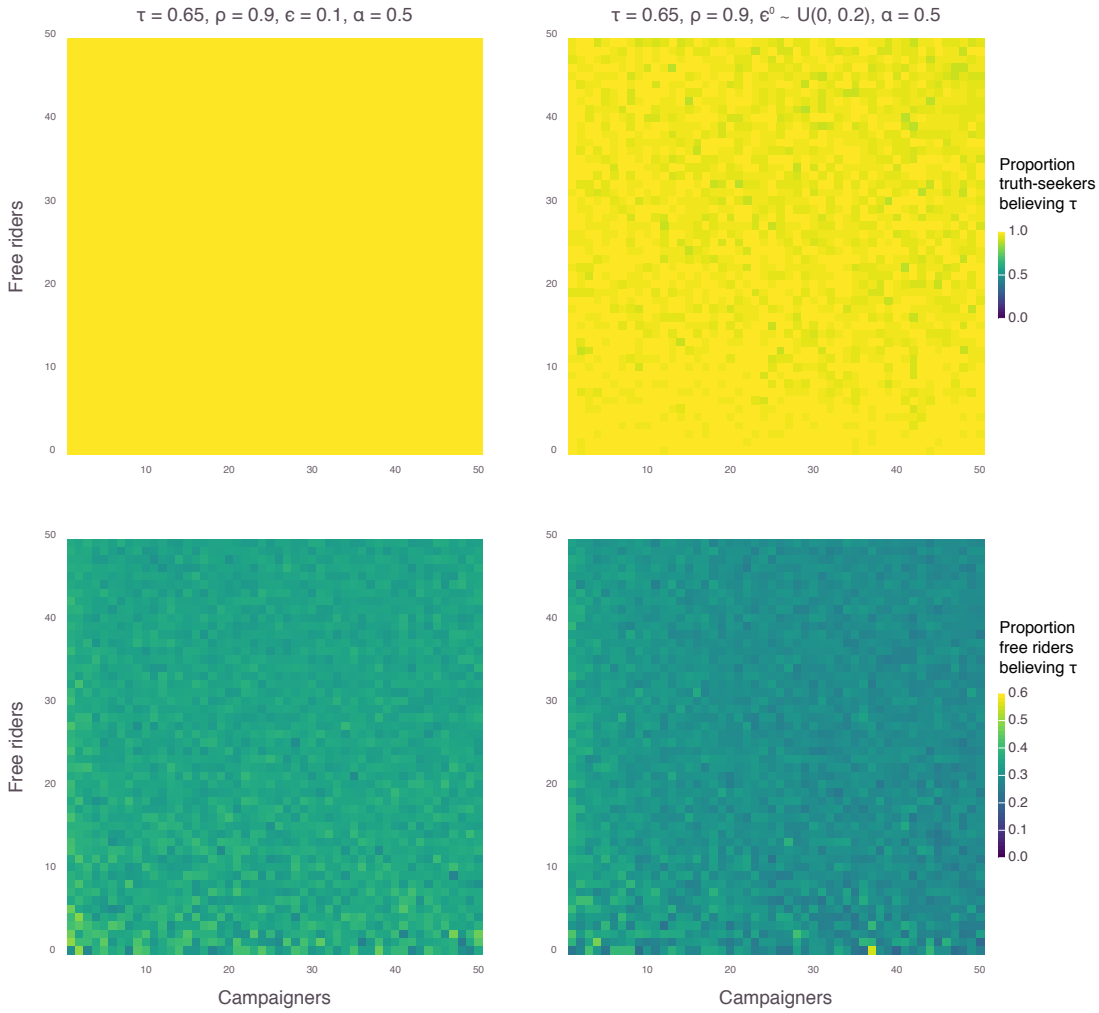

Figure 14: Average proportions of truth-seekers (top row) and free riders (bottom row) believing the truth in the fixed point, as in Figure 13, but with a wider gap between the truth and the campaigners' opinion. (Color keys apply per row.)

So, in a situation in which evildoers are helped enough if they can sway some free riders—perhaps having no hope of diverting any truth-seekers from the truth—they may be more effective if they are operating in a community with CD. By comparing the top and bottom row of Figure 15, we also see that, if this is what the campaigners are after, they better lie blatantly rather than subtly. It is starting to appear again that, once we admit non-truthfulness in the BC model, the only consistent message may be that there *is no* consistent message. What is strategically best from the evildoers' perspective and what is best for us to defend ourselves against them both appear to be highly context-dependent.

4.2 Societal costs

To show the effect of CD on accuracy, we look at what difference it makes if this kind of dynamics is assumed in the situations studied in Examples 3.1 and 3.2. In the former, we compared in terms

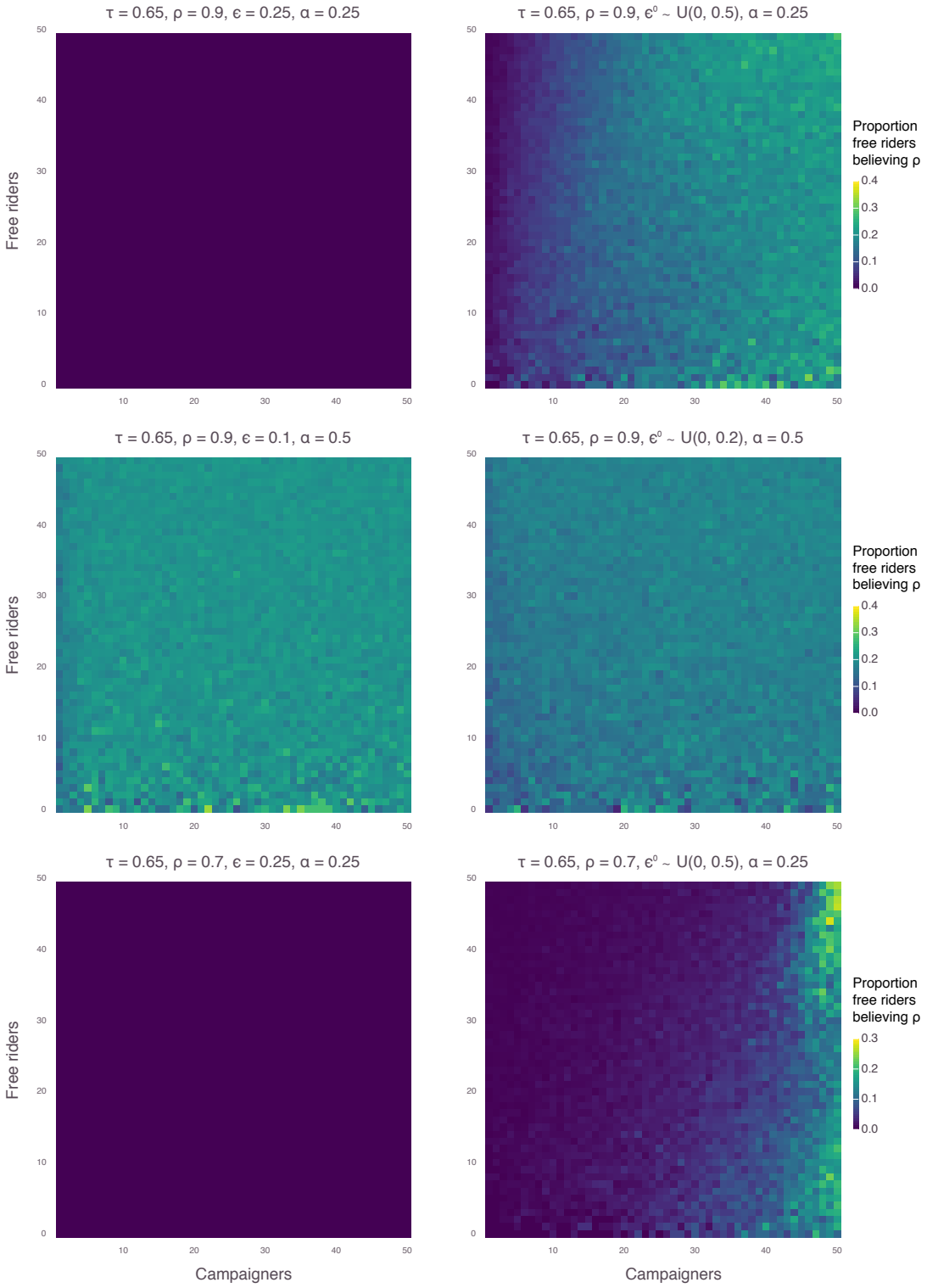

Figure 15: Each row showing for a different parameter settings proportions of free riders that believe the campaigners' opinion in the fixed point, without (left column) and with (right column) confidence dynamics. Results are averages over 100 simulations per number-of-campaigners–number-of-free-riders combination. (Color keys apply per row.)

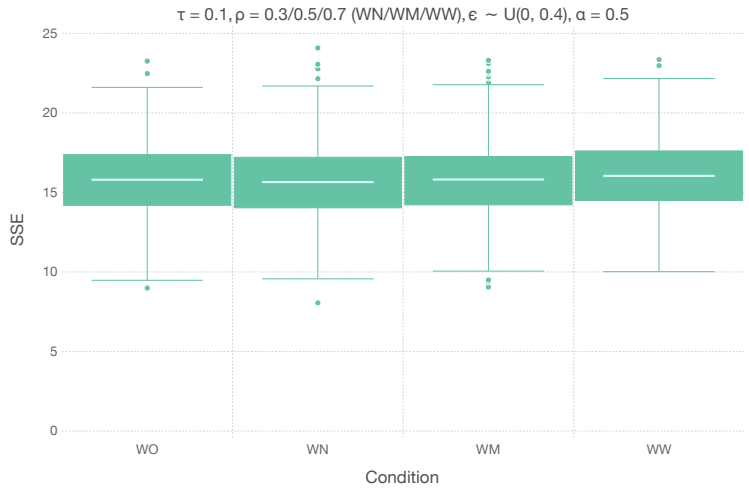

Figure 16: Same as for Figure 8 except that now confidence dynamics has been added.

of accuracy a community *without* campaigners with communities *with* campaigners, where the communities with campaigners differed from each other in how distant from the truth the fixed opinion was. In the latter example, we compared, also in terms of accuracy, communities with different numbers of free riders. In none of the examples did we implement CD. The following example revisits Example 3.1, but now adding CD:

Example 4.2 As in Example 3.1, $\tau = 0.1$ and $\alpha = 0.5$ hold for four communities, the first of which consists only of truth-seekers, and the other three of which also contain 25 campaigners, where these hold a fixed opinion of 0.3 in one community, 0.5 in the second, and 0.7 in the last. We saw that, especially in the community with the most "subtle" campaigners, those agents largely hampered the accuracy of the truth-seekers.

Figure 16 shows the results from running 1,000 simulations for each community and measuring the SSEs, on the assumption that $\varepsilon^0 \sim U(0, 0.4)$, so that the average starting BCI size of the truth-seekers is the same as the (fixed and universal) BCI size of the truth-seekers in Example 3.1. This figure suggests that, if CD is assumed, the campaigners no longer have the ability to dramatically affect the accuracy of the truth-seekers, not even if the former lie only subtly; comparing Figure 16 with Figure 8 makes the difference particularly clear. Whereas running an ANOVA on the outcomes of the new simulations reveals significant differences between the means of the various conditions—$F(3, 3996) = 7.27$, $p < .0001$—and a Tukey post hoc test shows the mean of the WW condition (16.10, ±2.22) to be significantly different from the means of the WO (15.76, ±2.28), the WN (15.64, ±2.31), and WM (15.79, ±2.31) conditions (all ps $< .05$), the size of the effect is negligible ($\eta^2 = .005$). In short, it appears that whether the campaigners lie blatantly or subtly, they are, in a community in which the truth-seekers adjust their BCIs by averaging in the way described in this section, very limited in the amount of damage they can do. ♦

Next let us look at Example 3.2 and rerun the simulations reported in it but now with CD added:

Example 4.3 Here too, we assume the same parameter settings as previously and also the same communities, that is, all consisting of 50 truth-seekers and 15 campaigners, with varying numbers of free

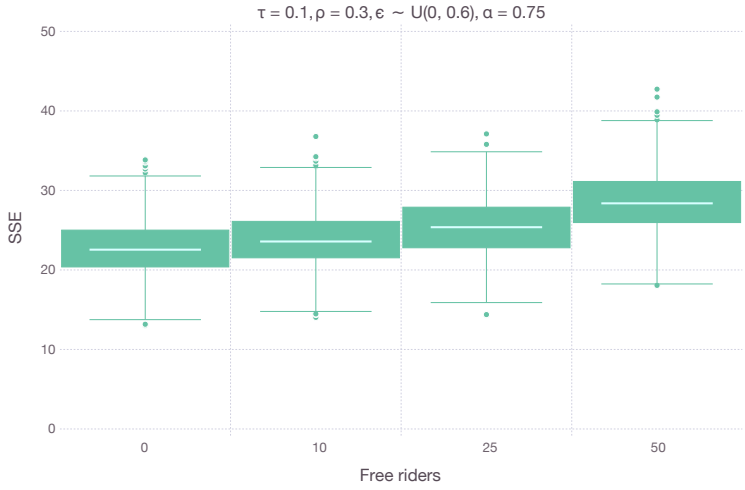

Figure 17: Same as for Figures 9 and 10 except now with confidence dynamics.

riders. We found a small but significant effect in Example 3.2 for $\tau = 0.1$, $\varrho = 0.3$, $\varepsilon = 0.2$, and $\alpha = 0.5$. With the same setting, except that now $\varepsilon^0 \sim \mathrm{U}(0, 0.4)$, we get no effect at all.

By contrast, for the other parameter setting considered in Example 3.2, where we found a large and significant effect, we this time find an even larger effect. Figure 17 shows the outcomes of measuring SSEs in 1,000 simulations for each of the four different communities, clearly indicating that accuracy goes down (SSEs go up) with increasing numbers of free riders. The corresponding ANOVA shows, as expected, that the differences in means are significant—$F(3, 3996) = 502.9$, $p < .0001$—and a Tukey post hoc test further showed that in fact all means are pairwise significantly different; all ps $< .0001$. The effect was, as said, even larger than the one we found for the same setting in Example 3.2: $\eta^2 = 0.27$ now versus $\eta^2 = 0.19$ previously. ♦

We wish we could give a clear and unequivocal answer to the question of whether, in a community which includes campaigners (who are willing to listen to no one and stick to their guns come what may), it helps when the other agents are open to becoming narrow-minded themselves. But once again, our results are a mixed bag, indicating that, under certain circumstances, the answer is positive but that at the same time it would be unwise to recommend CD *generally*.

4.3 Full confidence dynamics

Finally, we would like to present an extension of the BC model that takes the idea of making BC updating more realistic one step further still. Just as, in reality, BCIs will not be the same for all agents and from one time step to the next, the weight agents give to their peers' opinions will also not be the same for all of them, nor remain unchanged through time. There is a vast psychological literature on the so-called conformity bias, which shows that people have a strong tendency to mimic (in all sorts of ways) those around them.[16] In view of this literature, it is reasonable to suppose that, as in the case of the BCI, our peers have some influence on the weight we give to their opinions; for instance, being

[16] See, for instance, Flache et al. (2017), Sunstein (2019), and references given in those publications.

around people who attach much value to the opinions of others may make *us* attach more value to the opinions of others as well.

A first idea is to assume, instead of one fixed α, doubly indexed weights α_i^u, with index i referring to an agent and index u to a time step, where the dynamics of those weights then unfolds in the same fashion as we let the BCI dynamics unfold earlier. More formally, the model would be this:

$$x_i(u+1) \;=\; \begin{cases} \frac{1-\alpha_i^u}{|X_i(u)|}\sum_{j\in X_i(u)} x_j(u) \;+\; \alpha_i^u\,\tau & \text{if } x_i \text{ is a truth-seeker} \\ \frac{1}{|X_i(u)|}\sum_{j\in X_i(u)} x_j(u) & \text{if } x_i \text{ is a free rider} \\ \varrho & \text{if } x_i \text{ is a campaigner,} \end{cases}$$

where $x_j(u)$ and $X_i(u)$ are as defined at the beginning of this section (so as to enable BCI dynamics), and where α_i^0 is randomly chosen from $(0, \hat{\alpha})$ and, for all $u \geqslant 0$,

$$\alpha_i^{u+1} \;:=\; \begin{cases} \frac{1}{|X_i(u)|}\sum_{j\in X_i(u)} \alpha_j^u & \text{if } x_i \text{ is a truth-seeker} \\ 0 & \text{if } x_i \text{ is a free rider} \\ 1 & \text{if } x_i \text{ is a campaigner.} \end{cases}$$

In this model, for truth-seekers, their peers influence their new opinion, their BCI, and also the weight they are going to give to the social part of updating at the next update. The idea behind the other clauses in the above definition is that free riders do not attach any weight to worldly evidence, and so the weight they give to the social part of updating is automatically and invariably 1; mutatis mutandis for the campaigners, who, although they pay no attention to wordly evidence either, also do not give any weight to the social part of updating: they simply stick to their fixed opinion.

But we want a model with full confidence dynamics to allow for a bit more flexibility, by giving agents some control over how *fast* their values for α and ε change under the influence of their peers. For instance, some of us may be much faster in following a trend than others. Being around people who attach great weight to the opinions of their peers, we may want to follow suit but at our own pace. Formally, we may want to move our value for α in the direction of the average of the α values of our peers, but in doing so we may not want to rush things and do not want to abandon our current α value in favor of that average entirely. The same remark applies to ε. Note that this is in fact fully in the spirit of the original BC model, which also gives one control over how fast one's opinion can change. For instance, choosing very small values for α and ε will guarantee that one's opinion changes only very slowly.

To achieve this, we only need to adapt the above model minimally. In particular, we add the assumption that, for each agent i, there are two further parameters, λ_i^α and λ_i^ε, which determine the weight the agent gives to her *current* values for α and ε, respectively, in determining her new values for these parameters. More exactly, we replace the above definition of α_i^{u+1} by this:

$$\alpha_i^{u+1} \;:=\; \begin{cases} \lambda_i^\alpha \cdot \alpha_i^u \;+\; (1-\lambda_i^\alpha)\cdot \frac{1}{|X_i(u)|}\sum_{j\in X_i(u)} \alpha_j^u & \text{if } x_i \text{ is a truth-seeker} \\ 0 & \text{if } x_i \text{ is a free rider} \\ 1 & \text{if } x_i \text{ is a campaigner.} \end{cases}$$

We similarly adapt the definition of ε_i^{u+1} given on page 20, meaning that we make ε_i^{u+1} a mixture of ε_i^u and the average of the ε values of agent i's peers at time u, with λ_i^ε and $(1-\lambda_i^\varepsilon)$ as weights.

Note that all models considered previously can be conceived as limiting cases of this model with full confidence dynamics. For instance, populating the model only with truth-seekers, setting $\lambda_i^\alpha = \lambda_i^\varepsilon = 0$ for all agents i and also fixing ε_i^u and α_i^u for all i and u, we obtain the original BC model we started out with in Section 2.1. There are many ways in which we can relax the aforementioned restrictions. For instance, we could continue assuming that $\lambda_i^\alpha = \lambda_i^\varepsilon = 0$ but then allow all agents to have their own α and ε values; we could add only free riders, or only campaigners, or both, and we could experiment with adding different numbers of both, much in the way we have done in the above. We could then also raise again the questions that were raised in the above, notably, questions concerning conversion and accuracy.

For instance, in the Supplementary Materials (specifically, in the second part of the tutorial—see the Appendix) it is shown that, in a model that is exactly like the simplest version of the BC model except that it allows all agents to choose their individual α and ε values, those values are highly significant predictors of how accurate an agent is over the totality of updates. This was the outcome from running 1,000 simulations with communities of 50 agents all updating 50 times, where at the beginning of each simulation the value of τ was chosen randomly and the agents chose their values for α and ε randomly as well. A linear model with sum of squared errors for the various agents as dependent variable, their α and ε values as independent variables, and value of τ in a given simulation as co-variate, revealed a β-coefficient of -0.66 for α and a β-coefficient of -0.24 for ε (both ps $<$ .0001). The former may not be too surprising: it indicates that the more weight agents give to worldly evidence, the less inaccurate (i.e., the more accurate) they become—which is what one would expect. The result for ε is more interesting, indicating that one also decreases one's inaccuracy by being more inclusive in counting others one's peers.

This result is for a model without campaigners and free riders and so not of immediate interest to our current project. We mention it because it served as a template for the more involved simulations we ran that do bear on the question of what social damage is done by campaigners and free riders. Each of these simulations started with randomly choosing values for τ and ϱ from the unit interval. There were always 50 truth-seekers, and we systematically went through all combinations of n campaigners and m free riders such that $n, m \in \{0, 10, 20, 30, 40, 50\}$, running 1,000 simulations for each combination. Also at the beginning of each simulation, truth-seekers and free riders chose their initial ε values as well as their values for λ^ε, and truth-seekers chose their initial α value as well as their value for λ^α. Each of the aforementioned values was chosen from the unit interval, randomly and independently. In each simulation, the agents updated 50 times, and we registered the sum of squared errors of each of the truth-seekers.

This time, because α and ε values were not kept fixed over time, these could not serve as predictor variables. Instead, we looked at the impact the λ^ε- and λ^α values (which *were* fixed over time) had on inaccuracy (i.e., SSEs). In addition to this, we were interested in the impact the absolute distance between τ and ϱ had on the dependent variable. To find out, we fitted a linear model to the simulation results per combination of number of campaigners and number of free riders and registered the β-coefficients for all three independent variables together with the associated p values.

Figure 18 gives a graphical overview of the findings. The plots in the right column show that, with few exceptions, all outcomes were significant, and typically highly significant. The upper left plot shows that the β-coefficient for λ^α was mostly negative, meaning that the higher an agent's λ^α—the less willing the agent is to change her α value—the lower that agent's SSE tended to be, that is, the more accurate the agent tended to be. More importantly still, the plot shows a pronounced pattern:

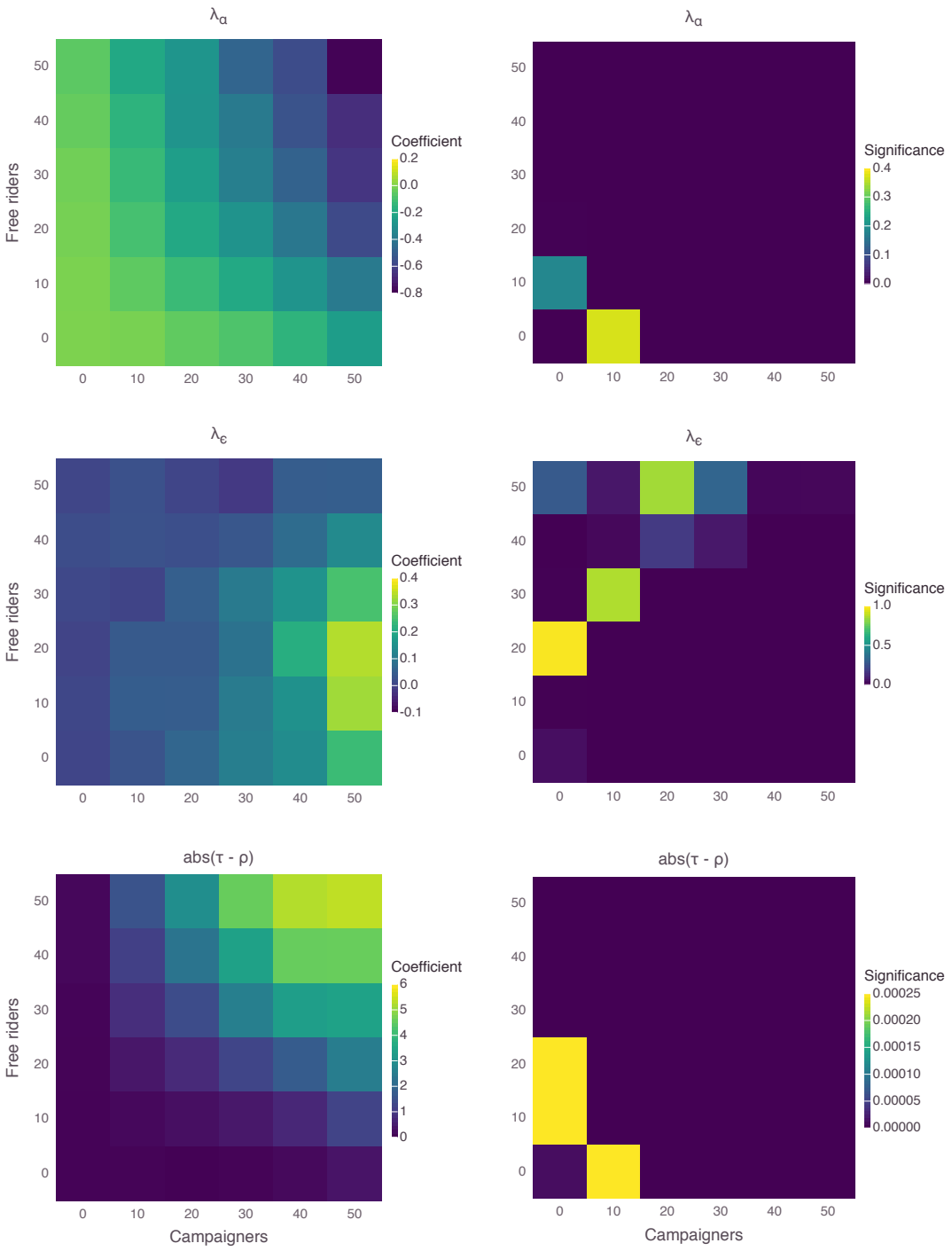

Figure 18: Graphical presentation of the outcomes of the linear models described in the text. The left column shows the coefficients of the predictor variables in those models, and the right column shows the α-levels at which the corresponding results hold.

an agent's λ^{α} value appears to have a stronger impact the more free riders and campaigners there are in the population. This phenomenon is easy to make sense of. For consider that both free riders and campaigners have an α value of 0. The more truth-seekers come in contact with those other types of agents, the more their α values will decrease, meaning that they are pulled toward giving more weight to the social aspect of updating and correspondingly less weight to the worldly evidence, which is likely to make their opinions less accurate. Yet, how strong this pull is depends on their λ^{α} value: higher values lessen the impact the free riders and/or campaigners have on a truth-seeker's α value, simply by making that agent give more weight to its current α value in determining its new one.

Almost the opposite holds for λ^{ε} values. As the middle plot in the left column of Figure 18 shows, the coefficient for λ^{ε} is mostly positive, meaning that larger values tend to increase SSEs (i.e., to make truth-seekers less accurate). We see that this is especially so when the population comprises many campaigners. Again there is no mystery here. The smaller one's ε value, the smaller the chance that any other given agent's opinion is within one's BCI, a fortiori, the smaller the chance that the campaigners' opinion is within one's BCI, and also the smaller the chance that the opinions of free riders held hostage by the campaigners (in the sense explained in Sect. 3.2) are within one's BCI. Consequently, the smaller one's ε value, the smaller the chance that one is detracted from following the lead in the worldly evidence that, as a truth-seeker, one also receives. While confidence dynamics has, in general, the advantage that the presence of campaigners will have a diminishing effect on the size of the ε values of the truth-seekers (and also the free riders) that count them as peers, higher values of λ^{ε} dampen that effect, making one longer vulnerable to the negative influence of campaigners.

Note again the consistency of these results with what has come to appear as the most general finding of our computational experiments, to wit, that there is no single recommendantion to be made to keep us—truth-seekers—safe from the influence of less benevolent epistemic actors. Higher λ^{α} values tend to help, though by how much depends on the number of non-truth-seekers in one's community. Higher λ^{ε} values, by contrast, tend to be *un*helpful, but again, much depends on the community structure. To distill sound advice from this is further complicated by the fact that both free riders and campaigners are most likely to present themselves as truth-seekers, making community structure hard to gauge.

We finally mention that the results concerning the absolute distance between τ (the truth) and ϱ (the campaigners' opinion), as seen in the bottom row of Figure 18, are largely unsurprising in view of what was previously found. When no or very few campaigners are around, then obviously the coefficient is 0 or close to 0. We also saw that free riders act a bit as catalysts when campaigners *are* around, playing the role of intermediaries needed to let the campaigners exert their evil influence. Thus, it is unsurprising to see in the bottom left panel that the absolute distance between τ and ϱ has no or little effect when no or few free riders are around, however many campaigners are around. And that this distance comes to have a larger and larger effect as both the number of free riders and the number of campaigners increase was also to be expected in light of what we found earlier.

5 General discussion

5.1 Summary of main results

Our project, and the idea of BC updating generally, starts from the assumption that social learning is indispensable. At the same time, we realize now (i.e., after the Brexit campaign and the 2016 US presidential election, and amidst the COVID-19 pandemic) more than ever that this dependency on

others may come at a cost, because it can be exploited for purposes whose realization may leave us disenfranchised. Probably, it *has* been exploited for as long as it exists: lying is not a recent invention. But the systematic and strategic form of exploitation that we have witnessed in recent years may well be unprecedented. So, there is practical interest in coming to know more about where our greatest vulnerabilities lie when it comes to communication with others, and also about what measures we may be able to take to protect ourselves against attempts at epistemic manipulation.

With this in mind, we revisited the BC model for studying interacting communities of artificial agents. In its original form, the model makes no provision for representing untruthfulness and hence for representing what we called mis- and disinformation campaigns, where the first is a type of campaign aimed at making people, or perhaps the public as a whole, believe one or more falsehoods, while the second is a type aimed at making them disbelieve or doubt certain truths that they previously endorsed or might be prepared to endorse.

We therefore first proposed an extension of the model in which agents could be to varying degrees epistemically irresponsible, by sticking steadfastly to an opinion different from the truth, ignoring all evidence and also the opinions of their fellow agents, or—in a milder fashion—by being disinterested in any worldly evidence and forming new opinions only by listening to others. In a further step, we zoomed in on the dynamics of counting others as one's peers, and hence as worthy of letting oneself be influenced by, given that becoming less rigid as regards who to rely on appears a natural response to being confronted with untruthful fellow agents. The second extension allowed us to run simulations of possible defense mechanisms against attempted mis- and disinformation efforts, by increasing the agents' selectiveness in their choice of influencers.

There are few general conclusions we can draw at this stage. Naturally, we saw that free riders contribute to the risk of a successful disinformation campaign. So we can generally counsel against free riding. But we would have done that anyway, for independent reasons (as mentioned toward the end of Sect. 3). Apart from that, it is difficult to make any specific recommendations. As we saw in Example 4.1, if CD is part of our belief-forming practices, then by lying blatantly our opponents might fail to divert any responsible person but succeed to convert some free riders—which could be enough to undermine our interests (e.g., it might be enough to hand the victory to our opponents). Under different circumstances, however, the opposite may occur, as we saw in Figure 15.

But even if our results fail to suggest an unequivocal answer to the question of how we might best defend ourselves against mis- and disinformation campaigns, they do shed some light on why it has proven so hard to fight these campaigns, and why we should not expect any global or permanent fixes. Every concrete recommendation one might want to make is likely to play out differently depending on various contextual factors. And recommendations that work may do so only temporarily. Noticing that subtle lies no longer do the job—perhaps because people have adjusted their threshold for deeming others peers—opponents may switch to blatant lying, which may require new adjustments from the public, and so on, initiating a cat-and-mouse game.

5.2 Limitations of our study

While we have made the BC model more suitable to the study of large-scale efforts at deceit, we recognize that it still has limitations which can only be overcome by extending it in ways that go well beyond, for instance, the addition of confidence dynamics that we undertook. For example, the restriction of the agents' opinions to a single issue appears rather serious to us. After all, it is easily imaginable how one and the same interest group might want to run a misinformation campaign with

regard to one issue and a disinformation campaign with regard to another. Also, irrespective of the type of campaign, their options might differ vis-à-vis different issues. With regard to one issue, they may have evidence suggesting that only a blatant lie will rally their own base, even if blatantly lying is likely to alienate free riders, let alone epistemically responsible agents (the truth-seekers, in our terminology). With regard to another issue, their evidence may indicate that there is no special need to drum up their own base—they are enthused enough to stick with them no matter their alleged view on this issue—but a subtle lie may help to sway some free riders and perhaps even some truth-seekers. In other words, the interest group may want to follow different procedures for different issues.

To be sure, this point only buttresses the overall conclusion of our paper—that there is no single general recommendation that one can make to fight mis- and disinformation campaigns—but it would nevertheless be good to model simultaneously such diverse strategies. One way to allow for this possibility is to follow the path of Riegler and Douven (2010), who extended the original BC model to cover communities with agents whose opinions concerned multiple issues, some logically related, some not (see Sect. 2.1). But this is a step we mention only as an avenue for future research.

A second limitation concerns the fact that, in our extensions as in the original BC model, agents adopt information coming from the world in a "black box" fashion, meaning that this part of the actual updating mechanism remains unspecified. This has the advantage of making the model both simple and general. On the other hand, there is at least the possibility that the exact updating mechanism that people employ makes a difference to the degree to which they are susceptible to mis- or disinformation. We mentioned Douven and Wenmackers' 2017 paper, which built on the BC model to compare Bayesian updating with a form of explanatory reasoning in a social setting (see also Douven, 2019, 2021). One notable finding of that work was that the explanation-based update rule they considered was better able than Bayes' rule to detect the signal in the noise. So, supposing that the truth is still somehow dominant in the media, and hence mis- and disinformation can be considered noise of sorts, one might hope that the given kind of explanatory reasoning offers some benefits in protecting us from the efforts of the ill-intending. But to investigate this systematically one would need to unpack the updating mechanism of the extended BC model with typed agents in the way Douven and Wenmackers unpacked the updating mechanism of the original BC model.

5.3 Avenues for future research

The aforementioned limitations already suggest two obvious avenues for future research. Further work on the extended model proposed in this paper should also look at issues of interpretation. In particular, we have introduced a typology of agents in terms of epistemically responsible behavior. What were called "truth-seekers" by us were the epistemically responsible agents, while what we called "campaigners" were the epistemically irresponsible ones. Note, however, that a radical *gestalt switch* is possible here: Equations (BC), (BCC) and (BCCF) are uninterpreted formulas, which get empirical content by dint of an interpretation that—implicitly or explicitly—was given in the explanation of the formulas. We refer to $x_i(u)$ (for given i and u) as an opinion, call τ the truth, speak of agents directly influenced by τ as "truth-seekers," refer to those who stick to ϱ all the time as "campaigners," and call the remaining agents pejoratively "free riders." But we could interpret the central equations quite differently. For instance, think of those who stay with ϱ all the time as scientists who have found the truth and will never be dissuaded from it. All others, however, are only influenced by the truth when it is already in their confidence interval. The agents previously referred to as "truth-seekers" are now epistemic villains, namely, members of a *conspiratorial group* who want to persuade as many

people as possible to believe τ. The conspirators do not reveal their real opinion. At the start of the dynamics, they distribute themselves randomly across the opinion space; and they agree to move toward their favored opinion τ with a speed controlled by the value of α. In another interpretation, the conspirators could simply be *bots* that are used in a "computational propaganda" campaign (see Woolley & Howard, 2019). In the probably darkest interpretation, the equations (BC), (BCC) and (BCCF) describe an opinion dynamics in which two competing manipulation campaigns that use two different approaches try to persuade innocent individuals. One campaign tries to persuade the innocents to believe τ, the other tries to persuade them to believe ς. And the innocents (our former free riders) simply average over all opinions within their confidence interval. The basic equations allow for all these interpretations.[17]

Finally, Hegselmann and Krause (2002) systematically explored the parameter space for the basic BC model which they presented in that paper and which we summarized in Section 2.1. Our main interest in the current paper was to extend that original model in order to shed some new light on the kind of epistemically irresponsible behavior that has lately been attracting a good deal of attention, in particular behavior related to mis- and disinformation campaigns. The extended BC model we developed to that end involves many more parameters than the original BC model, too many to go through the whole parameter space in a systematic fashion. We therefore at various junctures relied on examples, effectively showing slices of parameter space, which jointly helped to make the case that there is no one-size-fits-all solution to problems posed by the spread of mis- and disinformation. In fact, we did better than that, by sampling parameter space and looking for statistically significant effects of centrally important parameters on a centrally important outcome variable (viz., accuracy). Moreover, in the Appendix we explain how interested readers can, without much effort, explore whichever part of parameter space may be of special relevance to their own projects. Still, we do believe that many of the newly introduced features merit further investigation independently of issues of mis- and disinformation, given that by adding these features we arrived at a more realistic model of bounded confidence updating. This, too, is left as a future project.[18]

Appendix

The code for all simulations reported in this paper was written in Julia, a new dynamic language for high-performance computing (Bezanson et al., 2017). Julia code reads almost like pseudo-code, meaning that anyone with some experience in *Mathematica*, MATLAB, NetLogo, Python, R, or similar languages will have little difficulty becoming productive in Julia in a very short time.

For our simulations, this is especially true now that a dedicated Julia package for agent-based modeling has been developed, to wit, the Agents.jl package, information about which can be found here: https://github.com/JuliaDynamics/Agents.jl; see also Vahdati (2019). The package offers a generic framework geared toward making any agent-based modeling easy by requiring from the user only a specification of agent and model properties. The package then takes care of setting up and running the simulations.

The basic BC model is actually one of the examples featured in the Agents.jl package documentation. In this paper, we went far beyond that model. We have therefore written some documenta-

[17] The gestalt switch is discussed for the first time in Hegselmann and Krause (2015).

[18] We are grateful to Christopher von Bülow and to three anonymous referees for valuable comments on previous versions of this paper.

tion to help readers replicate the findings documented in the above using the Agents.jl package. See https://igordouven.github.io/MisDisInformation1, https://igordouven.github.io/MisDisInformation2, https://igordouven.github.io/MisDisInformation3, and https://igordouven.github.io/MisDisInformation4.

The Agents.jl package is geared primarily toward ease of use and generality, and only then toward performance. Given that some of our simulations are computationally highly expensive, we wrote code that did not rely on the Agents.jl package and that is much more performant. We are making this code available as well. It can be retrieved from the following GitHub repository: https://github.com/IgorDouven/Misinformation.

References

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*, 65–98.

Chazelle, B. (2011). The total *s*-energy of a multiagent system. *SIAM Journal on Control and Optimization*, *49*, 1680–1706.

Chen, G. & Lou, Y. (2019). *Naming game: Models, simulations and analysis*. Cham: Springer.

Crosscombe, M. & Lawry, J. (2016). A model of multi-agent consensus for vague and uncertain beliefs. *Adaptive Behavior*, *24*, 249–260.

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *3*, 87–98.

De Langhe, R. (2013). Peer disagreement under multiple epistemic constraints. *Synthese*, *190*, 2547–2556.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*, 554–559.

Dittmer, J. C. (2001). Consensus formation under bounded confidence. *Nonlinear Analysis*, *7*, 4615–4621.

Douven, I. (2010). Simulating peer disagreements. *Studies in History and Philosophy of Science*, *41*, 148–157.

Douven, I. (2019). Optimizing group learning: An evolutionary computing approach. *Artificial Intelligence*, *275*, 235–251.

Douven, I. (2021a) *The art of abduction*. Cambridge MA: MIT Press.

Douven, I. (2021b). Explaining the success of induction. *British Journal for the Philosophy of Science*, in press.

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

Douven, I. & Riegler, A. (2010). Extending the Hegselmann–Krause model I. *Logic Journal of the IGPL*, *18*, 323–335.

Douven, I. & Wenmackers, S. (2017). Inference to the best explanation versus Bayes' rule in a social setting. *British Journal for the Philosophy of Science*, *68*, 535–570.

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. 2017. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, *20*, http://jasss.soc.surrey.ac.uk/20/4/2.html.

Gao, J., Li, B., Schoenebeck, G., & Yu, F. Y. (2017). Engineering agreement: The naming game with asymmetric and heterogeneous agents. *Thirty-First AAAI Conference on Artificial Intelligence*, 537–543.

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

Gottifredi, S., Tamargo, L. H., García, A. J., & Simari, G. R. (2018). Arguing about informant credibility in open multi-agent systems. *Artificial Intelligence*, *259*, 91–109.

Gribbin, J. (2002). *Science: A history*. London: Penguin.

Hahn, U., Hansen, J. U., & Olsson, E. J. (2020). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent. *Synthese*, *197*, 1511–1541.

Hegarty, P. & Wedin, E. (2016). The Hegselmann–Krause dynamics for equally spaced agents. *Journal of Difference Equations and Applications*, *22*, 1621–1645.

Hegselmann, R. (2004). Opinion dynamics: Insights by radically simplifying models. In D. Gillies (ed.) *Laws and models in science* (pp. 19–44). London: King's College Publications.

Hegselmann, R. (2014). Bounded confidence, radical groups, and charismatic leaders. In F. Miguel, F. Amblard, J. Barceló, & M. Madella (eds.) *Social simulation conference advances in computational social science and social simulation*. Barcelona: Autonomous University of Barcelona, DDD repository, http://ddd.uab.cat/record/125597.

Hegselmann, R. (2020). Polarization and radicalization in the bounded confidence model: A computer-aided speculation. In V. Buskens, R. Corten, & C. Snijders (eds.) *Advances in the sociology of trust and cooperation: Theory, experiment, and field studies* (pp. 197–226). Berlin: De Gruyter.

Hegselmann, R., König, S., Kurz, S., Niemann, C., & Rambau, J. (2015). Optimal opinion control: The campaign problem. *Journal of Artificial Societies and Social Simulation*, *18*, http://jasss.soc.surrey.ac.uk/18/3/18.html.

Hegselmann, R. & Krause, U. (2002). Opinion dynamics and bounded confidence: Models, analysis, and simulations. *Journal of Artificial Societies and Social Simulation*, *5*, http://jasss.soc.surrey.ac.uk/5/3/2.html.

Hegselmann, R. & Krause, U. (2005). Opinion dynamics driven by various ways of averaging. *Computational Economics*, *25*, 381–405.

Hegselmann, R. & Krause, U. (2006). Truth and cognitive division of labor: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation*, *9*, http://jasss.soc.surrey.ac.uk/9/3/10.html.

Hegselmann, R. & Krause, U. (2009). Deliberative exchange, truth, and cognitive division of labour: A low-resolution modeling approach. *Episteme*, *6*, 130–144.

Hegselmann, R. & Krause, U. (2015). Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: A simple unifying model. *Networks and Heterogeneous Media*, *10*, 477–509.

Hegselmann, R. & Krause, U. (2019). Consensus and fragmentation of opinions with a focus on bounded confidence. *American Mathematical Monthly*, *126*, 700–716.

Howell, L. (2013). Digital wildfires in a hyperconnected world. *WEF Report*, *3*, 15–94.

Jacobmeier, D. (2004). Multidimensional consensus model on a Barabási–Albert network. *International Journal of Modern Physics C*, *16*, 633–646.

Kavanagh, J. & Rich, M. D (2018). *Truth decay: An initial exploration of the diminishing role of facts and analysis in American public life*. Santa Monica CA: RAND Corporation.

Kitcher, P. (1993). *The advancement of science*. Oxford: Oxford University Press.
Krause, U. (2000). A discrete nonlinear and non-autonomous model of consensus formation. In S. Elaydi, G. Ladas, J. Popenda, & J. Rakowski (eds.) *Communications in difference equations* (pp. 227–236). Amsterdam: Gordon and Breach.
Krause, U. (2015). *Positive dynamical systems in discrete time: Theory, models, and applications*. Berlin: De Gruyter.
Kuipers, T. A. F. (2001). *Structures in science*. Dordrecht: Kluwer.
Kummerfeld, E. & Zollman, K. J. S. (2016). Conservatism and the scientific state of nature. *British Journal for the Philosophy of Science*, *82*, 956–968.
Kurz, S. (2015). Optimal control of the freezing time in the Hegselmann–Krause dynamics. *Journal of Difference Equations and Applications*, *21*, 633–648.
Kurz, S. & Rambau, J. (2011). On the Hegselmann–Krause conjecture in opinion dynamics. *Journal of Difference Equations and Applications*, *17*, 859–876.
Lorenz, J. (2003). *Mehrdimensionale Meinungsdynamik bei wechselndem Vertrauen*. Diploma thesis, University of Bremen, available at http://nbn-resolving.de/urn:nbn:de:gbv:46-diplo00000564.
Lorenz, J. (2008). Fostering consensus in multidimensional continuous opinion dynamics under bounded confidence. In D. Helbing (ed.) *Managing complexity: Insights, concepts, applications* (pp. 321–334). Berlin/Heidelberg: Springer.
Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 9020–9025.
Lu, Q., Korniss, G., & Szymanski, B. K. (2009). The naming game in social networks: Community formation and consensus engineering. *Journal of Economic Interaction and Coordination*, *4*, 221–235.
Mäki, U. (1992). On the method of isolation in economics. In C. Dilworth (ed.) *Idealization IV: Intelligibility in science* (pp. 319–354). New York: Rodopi.
Mason, L. (2018). *Uncivil agreement: How politics became our identity*. Chicago: University of Chicago Press.
Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, *11*, 279–300.
Mason, W. A. & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 764–769.
Mocanu, D., Rossi, L., Zhang, Q., Karsai, M., & Quattrociocchi, W. (2015). Collective attention in the age of (mis)information. *Computers in Human Behavior*, *51*, 1198–1204.
Nichols, T. (2017). *The death of expertise: The campaign against established knowledge and why it matters*. Oxford: Oxford University Press.
Nowak, L. (1980). *The structure of idealization*. Dordrecht: Reidel.
Nunes, D. & Antunes, L. (2015). Modelling structured societies: A multi-relational approach to context permeability. *Artificial Intelligence*, *229*, 175–199.
O'Connor, C. & Weatherall, J. O. (2017). Scientific polarization. *European Journal for Philosophy of Science*, *8*, 855–875.
O'Connor, C. & Weatherall, J. O. (2019). *The misinformation age: How false beliefs spread*. New Haven CT: Yale University Press.

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

Pluchino, A., Latora, V., & Rapisarda, A. (2006). Compromise and synchronization in opinion dynamics. *European Physical Journal B*, *50*, 169–176.

Pomerantsev, P. (2019). *This is not propaganda: Adventures in the war against reality*. London: Faber & Faber.

Proctor, R. N. & Schiebinger, L. (ed.) (2008). *Agnotology: The making and unmaking of ignorance*. Stanford CA: Stanford University Press.

Riegler, A. & Douven, I. (2009). Extending the Hegselmann–Krause model III: From single beliefs to complex belief states. *Episteme*, *6*, 145–163.

Rosenstock, S., Bruner, J., & O'Connor, C. (2017). In epistemic networks, is less really more? *Philosophy of Science*, *84*, 234–252.

Schurz, G. (2019). *Hume's problem solved: The optimality of meta-induction*. Cambridge MA: MIT Press.

Semeshenko, V., Gordon, M. B., & Nadal, J.-P. (2008). Collective states in social systems with interacting learning agents. *Physica A: Statistical Mechanics and its Applications*, *387*, 4903–4916.

Shoham, Y., Powers, R., & Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, *171*, 365–377.

Sunstein, C. R. (2019). *Conformity: The power of social influences*. New York: New York University Press.

Tamargo, L. H., Garcia, A. J., Falappa, M. A., & Simari, G. R. (2014). On the revision of informant credibility orders. *Artificial Intelligence*, *212*, 36–58.

Temin, P. (2017). *The vanishing middle class: Prejudice and power in a dual economy*. Cambridge MA: MIT Press.

Tsang, A. & Larson, K. (2014). Opinion dynamics of skeptical agents. *AAMAS '14: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, 277–284.

Vahdati, A. R. (2019). Agents.jl: Agent-based modeling framework in Julia. *Journal of Open Source Software*, *4*, 1611, https://doi.org/10.21105/joss.01611.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *6380*, 1146–1151.

Weatherall, J. O., O'Connor, C., & Bruner, J. P. (2020). How to beat science and influence people: Policymakers and propaganda in epistemic networks. *British Journal for the Philosophy of Science*, axy062, https://doi.org/10.1093/bjps/axy062.

Weisbuch, G., Deffuant, G., Amblard, F., & Nadal, J.-P. (2002). Meet, discuss and segregate! *Complexity*, *7*, 55–63.

Wenmackers, S., Vanpoucke, D., & Douven, I. (2012). Probability of inconsistencies in theory revision: A multi-agent model for updating logically interconnected beliefs under bounded confidence. *European Physical Journal B*, *85*, https://doi.org/10.1140/epjb/e2011-20617-8.

Wenmackers, S., Vanpoucke, D., & Douven, I. (2014). Rationality: A social-epistemology perspective. *Frontiers in Psychology*, *5*, 581, https://doi.org/10.3389/fpsyg.2014.00581.

Woolley S. C. & Howard, P. N. (eds.) (2019). *Computational propaganda: Political parties, politicians, and political manipulation on social media*. New York: Oxford University Press.

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