ALEXANDER RIEGLER AND IGOR DOUVEN

EXTENDING THE HEGSELMANN–KRAUSE MODEL III: FROM SINGLE BELIEFS TO COMPLEX BELIEF STATES

ABSTRACT

In recent years, various computational models have been developed for studying the dynamics of belief formation in a population of epistemically interacting agents that try to determine the numerical value of a given parameter. Whereas in those models, agents' belief states consist of single numerical beliefs, the present paper describes a model that equips agents with richer belief states containing many beliefs that, moreover, are logically interconnected. Correspondingly, the truth the agents are after is a theory (a set of sentences of a given language) rather than a numerical value. The agents epistemically interact with each other and also receive evidence in varying degrees of informativeness about the truth. We use computer simulations to study how fast and accurately such populations as wholes are able to approach the truth under differing combinations of settings of the key parameters of the model, such as the degree of informativeness of the evidence and the weight the agents give to the evidence.

1. INTRODUCTION

In recent years, a variety of computational models have been developed for studying the dynamics of some clearly circumscribed types of epistemic interaction.[1] In these so-called models of opinion dynamics, truth-seeking agents interact with each other and revise their beliefs either purely on the basis of these interactions or, in the more advanced models, on the basis of the interactions plus independent evidence they receive concerning the truth they are after. The presumably best known model of this sort is the Hegselmann-Krause (HK) model, developed by Rainer Hegselmann and Ulrich Krause, in which agents repeatedly revise their beliefs by averaging (in a specified way), on the one hand, the average of the beliefs of those agents in the model whose beliefs are close to their own and, on the other hand, the truth.

A commonality of the models of opinion dynamics that have been devised so far is that the truth the agents in them are after is the numerical value of some parameter (which is typically left unspecified); accordingly, the belief states of the agents consist of single numerical beliefs. This not only helps to keep the

DOI: 10.3366/E1742360009000616

computational complexity of the models within certain boundaries, it also suggests straightforward ways in which an agent may revise its belief in light of other agents' beliefs and the evidence it receives; for example, it could simply take the arithmetic mean of those beliefs and the evidence (which, in the said models, also comes in the form of a number). But while, as we argued elsewhere (Douven and Riegler 2009), these relatively simple models already help to answer, or at least elucidate, issues directly relevant to social epistemology, their scope is obviously limited in this respect: human agents tend to have numerous beliefs, many of them of a nonnumerical nature, which, moreover, are typically interconnected in certain ways.

If we want to study types of epistemic interaction between agents capable of having such richer belief states – belief states that cannot be adequately characterized by a single numerical value – we will have to look for models that, in all likelihood, are more complex than the ones that have been developed until now. It is far from obvious what such a model should look like. For instance, it is not immediately clear how someone who believes both φ and ψ might reasonably compromise with someone who believes only the disjunction of these propositions (even supposing they have no further beliefs). Surely they cannot "meet in the middle" – as an analogue of taking the arithmetic mean – in any straightforward sense. A belief state dynamics that considers such belief states is certainly less clearly a candidate for computational modelling than the current models of opinion dynamics. Nevertheless, in this paper we aim to take some first steps toward developing such a model, building on Hegselmann and Krause's work and on previous related work of our own. In particular, the model described in the following is a further extension of the two-dimensional variant of the HK model presented in Riegler and Douven (2009), in which agents move in a discrete two-dimensional grid, where they meet with other agents and update their respective (numerical) beliefs in response to these meetings.

In the new model, we study agents of epistemically interacting truth-seeking agents, where the truth is now a theory (a set of sentences belonging to a given language, satisfying certain closure conditions) rather than a numerical value, and where the agents also receive evidence in varying degrees of informativeness about the truth. Computer simulations will be used to determine how fast and accurately such populations as wholes are able to approach the truth under differing combinations of settings of the key parameters of the model, such as the degree of informativeness of the evidence and the weight the agents attribute to the evidence. We will point out parallels between the results obtained in the new model and those obtained in our earlier extensions of the HK model. But we start by briefly rehearsing the basics of the original HK model and of some earlier extensions.

2. THE HEGSELMANN-KRAUSE MODEL AND BEYOND

The basic components of the HK model are, first, a set of discrete time points, second, a population of agents, each of which holds a belief at any given time,

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

and third, an update mechanism for revising the agents' beliefs. The agents are supposed to be looking for the value of a given parameter τ, located in the interval $[0, 1]$. They antecedently know that the parameter lies in this interval and their beliefs are consistent with this knowledge, so that, where $x_i(t)$ is agent x's belief at time t, $x_i(t) \in [0, 1]$ for all i and t. Agents simultaneously revise their beliefs about τ at each time t, where for each i the new belief at time $t + 1$ is given by

$$x_i(t+1) = \alpha \frac{\sum_{j \in X_i(t)} x_j(t)}{|X_i(t)|} + (1 - \alpha)\tau. \tag{1}$$

Here, $|X_i(t)|$ is the cardinality of what we shall refer to as i's *epistemic neighborhood* at t, that is, the set of agents whose beliefs at t are close to i's own belief at that time; more formally, $X_i(t) = \{j : |x_i(t) - x_j(t)| \leq \varepsilon\}$, for some $\varepsilon \in [0, 1]$ (in Hegselmann and Krause's terminology, i's epistemic neighbors at t are those agents that are within i's *bounded confidence interval* at t). The global parameter $\alpha \in [0, 1]$ determines how much weight an agent attributes to its epistemic neighbors relative to the evidence it receives about τ. A useful informal way of thinking about this model is that in it an agent's new belief results from a combination of talking to its epistemic neighbors and performing experiments, where the results of the experiments point in the direction of τ.

In previous papers, we have highlighted various limitations of the above model. For instance, the model assumes that the agents always receive *accurate* evidence regarding τ and also that the agents' beliefs all weigh equally heavily in the updating process. In Douven (2009) and Douven and Riegler (2009), it was noted that neither assumption is particularly realistic. For instance, researchers have to live with measurement errors and other factors that make their data noisy, and it should also be uncontroversial that, both in daily life and in scientific practice, the beliefs of some count for more than those of others. Douven and Riegler (2009) thus proposed an extension of the HK model which has instead of (1) the following as an update rule:

$$x_i(t+1) = \alpha \frac{\sum_{j \in X_i(t)} x_j(t) w_j}{\sum_{j \in X_i(t)} w_j} + (1 - \alpha)(\tau + \text{rnd}(\zeta)). \tag{2}$$

In this equation, $w_j \geq 0$ denotes the fixed reputation of agent j, and $\text{rnd}(\zeta)$ is a function returning a unique uniformly distributed random real number in the interval $[-\zeta, +\zeta]$, with $\zeta \in [0, 1]$, each time the function is invoked.

One main result concerning this model was that in situations in which the evidence agents receive may be noisy, populations of agents that attribute more weight to talking to each other end up on average being closer to the truth over time than populations of agents that give more weight to the evidence, although the latter populations get faster to a value that is at least moderately close to the truth. Another main result was that variations in the assignment of reputations have

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

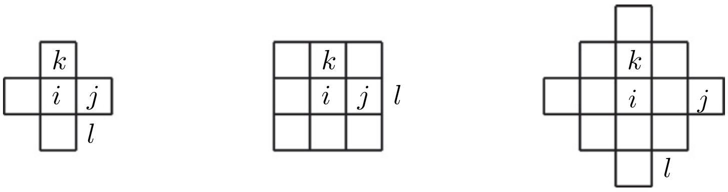

Figure 1. Three neighborhood structures: von Neumann (left), Moore (center), and Gaylord-Nishidate (right). In each case *j* and *k* are spatial neighbors of *i*, whereas *l* is not.

neither a significant influence on average speed of convergence nor on average accuracy of convergence, and that this is so even if some subgroup of agents receives evidence that is considerably less noisy than the evidence the rest of the population receives.

In Riegler and Douven (2009), we drew attention to another limitation of the HK model (shared by the above extension of it), to wit, that it assumes, again quite unrealistically, that all agents know at each point in time the beliefs of all the other agents.[2] To do away with this idealization, we extended the HK model by adding spatial dimensions to it. In the extended model, each agent inhabits a site in a discrete two-dimensional toroidal grid, facing one of the four cardinal points of the compass. The agents can move in the grid according to certain determinate rules. Unlike in the original HK model, an agent in this two-dimensional model does not interact with *all* its epistemic neighbors at each time step, but only with those that are to be found in its *spatial* neighborhood, where this notion can be given various different definitions. In the paper, we made use of the von Neumann, Moore, and Gaylord-Nishidate neighborhood structures, as depicted in Figure 1, which are common in the literature (see, e.g., Gaylord and D'Andria 1998). The first was actually used in two different ways, one in which agents interact with all epistemic neighbors in their spatial neighborhood, and one in which they interact only with those epistemic neighbors that are in their spatial neighborhood *and* that face the agent's position.[3]

Corresponding to the distinction between epistemic and spatial neighborhoods, we characterized the development of agents through time by both a belief update rule and a migration rule. For the former, we proposed this:

$$x_i(t+1) = \begin{cases} \alpha \frac{\sum_{j \in X_i(t)} x_j(t)}{|X_i(t)|} + (1-\alpha)(\tau + \mathrm{rnd}(\zeta)) & \text{if } |X_i(t)| > 1, \\ x_i(t) & \text{otherwise,} \end{cases} \tag{3}$$

where $X_i(t)$ now designates the set of *i*'s epistemic neighbors at t that are also within its spatial neighborhood at that time. (Note that since each agent counts trivially as its own epistemic and spatial neighbor at each time, the upper clause of equation (3) is invoked only when there is at least one epistemic neighbor present in the spatial neighborhood besides the agent itself.) For the migration rule, we proposed that after an agent has updated its belief, it moves one step to the adjacent

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

site it faces if that is free and not faced by at least one other agent, or else it randomly changes its orientation to any of the four directions.

The results of the computer experiments conducted using this model warranted the same conclusion we had been able to draw from our studies carried out by means of the simpler extension of the HK model described four paragraphs back: on the positive side, assuming the evidence to be noisy, talking to others helps the agents to get, on average, closer to the truth – closer, at any rate, than if they disregard the beliefs of others and go purely by the evidence; on the negative side, talking to others decelerates the convergence to the truth.

While the two-dimensional model is clearly less idealized than the HK model, it still has important limitations. As remarked at the outset, people typically have much richer belief states than those the agents considered in the above models are equipped with; they have belief states containing multitudinous beliefs, which are not all numerical in nature and which tend to, or at least are hoped to, obey certain logical principles (such as consistency). While the first of these issues – the number of beliefs – has to some extent been addressed in the work of Lorenz (2003, 2007, 2008), Jacobmeier (2004), and Pluchino, Latora, and Rapisarda (2006), the agents in the models these papers present still only possess numerical beliefs, which, moreover, fail to be interconnected in any interesting sense. In Section 3, we aim to go beyond this by putting forward a two-dimensional model populated by agents equipped with belief states containing not just numerous but also logically related beliefs. In Section 4, we then investigate the basic properties of this model by means of computer experiments.

3. COMPLEX BELIEF STATE DYNAMICS

We take from the two-dimensional extension of the HK model the idea that agents can move in a two-dimensional space and epistemically interact only with those agents that are both their epistemic and their spatial neighbors. However, the belief states of the agents in the new model no longer consist of single numerical beliefs but of theories formulated in a particular language. This requires both a new definition of "epistemic neighbor" and a new belief update algorithm. The definition, or rather definitions, of "spatial neighbor" remain unchanged, as does the migration rule.

The belief states of the agents can be represented in a finitary propositional language, $\mathcal{L}_m$, for some $m \in \mathbb{N}^+$, built up from m atomic sentences $\{\varphi_i\}_{i \leq m}$ and the usual logical connectives. We think of these languages as *interpreted* languages, and we assume that for each of them the classical consequence relation (designated by the symbol $\vdash$) holds.

Let $\mathcal{T}_m$ be the set of theories that can be formulated in $\mathcal{L}_m$. Then, where $B_i(t)$ is the belief state of agent i at time t, we have throughout the remainder that $B_i(t) \in \mathcal{T}_m$, for all i and t. In the following, all theories – including belief states and the truth – are implicitly assumed to be closed under the consequence relation.

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

Note that there are only finitely many different theories that can be formulated in each language – to be precise, 2^{2^m}; in particular, there are $2^{16} = 65,536$ different theories that can be formulated in $\mathcal{L}_4$, which is the language we will work with in our later computer experiments.

To give the new definition of "epistemic neighbor" and state the corresponding belief update mechanism, we first introduce some notational conventions. By elementary logic, each theory in $\mathcal{T}_m$ (for any m) can be stated as an $\mathcal{L}_m$-sentence in disjunctive normal form (DNF) that contains as disjuncts exclusively what (following Carnap) we shall refer to as the *state descriptions* of $\mathcal{L}_m$, that is, instances of the schema $\pm\varphi_1 \wedge \cdots \wedge \pm\varphi_m$, which are the logically strongest consistent sentences of the language; we call this DNF the theory's *canonical* DNF, or CDNF for short. The state descriptions can be ordered so that each theory in CDNF can be represented by a vector of 0's and 1's, where a 1 (respectively, 0) at the n-th place of the vector indicates that the n-th state description occurs (does not occur) in the CDNF. For example, assuming the following ordering of the state descriptions of $\mathcal{L}_2$: $\langle \varphi_1 \wedge \varphi_2, \varphi_1 \wedge \neg\varphi_2, \neg\varphi_1 \wedge \varphi_2, \neg\varphi_1 \wedge \neg\varphi_2 \rangle$, we can represent $\neg\varphi_1$ as the vector $\langle 0, 0, 1, 1 \rangle$, given that the CDNF representation of $\neg\varphi_1$ is $(\neg\varphi_1 \wedge \varphi_2) \vee (\neg\varphi_1 \wedge \neg\varphi_2)$, that is, its disjuncts are the third and fourth elements of the designated ordering; similarly, $\varphi_1 \rightarrow \neg\varphi_2$ can be represented as the vector $\langle 0, 1, 1, 1 \rangle$ and $\varphi_1 \vee \varphi_2$ as $\langle 1, 1, 1, 0 \rangle$. Note that, given this way of representing theories, the inconsistent theory is represented by the vector $\langle 0, 0, 0, 0 \rangle$ and the tautology by $\langle 1, 1, 1, 1 \rangle$; more generally, supposing the same type of coding, the inconsistent theory in $\mathcal{L}_m$ is represented by a vector of 2^m 0's and the tautology by one of just as many 1's. Instead of introducing special notation for the vector representation of theories in CDNF, we simply stipulate that, from now on, whenever we speak of theories (or belief states, or the truth, or evidence), we mean the vector notations of these theories' CDNF's. Also, in the following we shall use $T[k]$ to refer to the k-th element of (the vector notation of the CDNF of) theory T.

We can now define the notion of epistemic neighborhood for this new model in terms of a metric on binary strings known in the literature as the *Hamming distance*.[4] Formally, the Hamming distance $\delta(s_1, s_2)$ between binary sequences s_1 and s_2 is defined as the number of digits in which they differ. Clearly, given the above convention of coding theories as binary vectors, we can speak of the Hamming distance between two theories. Thus, for example, still assuming $\mathcal{L}_2$, if agent i's belief state at t is (the logical closure of) $\neg\varphi_1$ and agent j's is (the logical closure of) $\varphi_1 \rightarrow \neg\varphi_2$, then $\delta(B_i(t), B_j(t)) = 1$ (because $B_i(t)$ and $B_j(t)$ differ just with respect to $\varphi_1 \wedge \neg\varphi_2$). In these terms, we can define agent i's epistemic neighborhood at t as

$$\mathcal{E}_i(t) := \{j : \delta(B_i(t), B_j(t)) \leq \varepsilon\}, \qquad \text{with } \varepsilon \in \mathbb{N}.$$

To give a concrete example in $\mathcal{L}_2$, if $B_i(t) = \varphi_1 \leftrightarrow \varphi_2$ and $\varepsilon = 1$, then $\mathcal{E}_i(t)$ consists of agents j (if any) such that $B_j(t) \in \{\varphi_1 \wedge \varphi_2, \neg\varphi_1 \wedge \neg\varphi_2, \varphi_1 \rightarrow \varphi_2, \varphi_2 \rightarrow \varphi_1, \varphi_1 \leftrightarrow \varphi_2\}$.

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

From the foregoing definition and the fact that δ is a metric, we can derive some important properties of epistemic neighborhoods. First, by the definition of a metric, δ is reflexive – that is, $\delta(B_i(t), B_i(t)) = 0$ for all i and t – and so $i \in \mathcal{E}_i(t)$ for all i, t, and ε. Second, by the same definition, δ is symmetric, so that $j \in \mathcal{E}_i(t)$ iff $i \in \mathcal{E}_j(t)$ for all i, j, t, and ε. However, we do not generally have that if $i \in \mathcal{E}_j(t)$ and $j \in \mathcal{E}_k(t)$, then also $i \in \mathcal{E}_k(t)$.

Agents interact with other agents who are in both their epistemic and spatial neighborhoods, where, as intimated, the latter is understood in one of the ways defined in Section 2. Thus, where $\mathcal{S}_i(t)$ designates i's spatial neighborhood at t (for some given neighborhood structure), we say that i at t interacts with all agents in $\mathcal{N}_i(t) = \{j : j \in \mathcal{E}_i(t) \cap j \in \mathcal{S}_i(t)\}$.

The agents' goal is to determine the truth, which again we designate by τ. This time, however, the truth is not the numerical value of a given parameter, but rather a contingent theory formulated in the agents' language. It is not necessarily the case that τ is one of the state descriptions of $\mathcal{L}_m$; consider that scientists are not typically after a complete description of the actual world, but rather after a nomological characterization of it, that is, roughly, a characterization of the laws our world obeys or, what amounts to the same, a characterization of the set of possible worlds obeying the laws that the actual world obeys (see, e.g., Kuipers (2000, ch. 7) for more on this).

Also, as in the previous models, the agents receive evidence about τ, which this time consists of theories in the agents' language that are entailed by τ. More precisely, if at t agent i meets with one or more epistemic neighbors, then it gets access to a piece of evidence $E_i(t) \in \mathcal{T}_m$ such that $\tau \vdash E_i(t)$. What we shall call the "informativeness" of a given piece of evidence (how much weaker than τ is it?) is constrained by $\zeta = \langle \zeta_{\text{min}}, \zeta_{\text{range}} \rangle$, such that $\zeta_{\text{min}} \leq \delta(\tau, E_i(t)) \leq \zeta_{\text{min}} + \zeta_{\text{range}} < 2^m$. So, assuming $\mathcal{L}_2$, $\tau = \varphi_1 \wedge \neg\varphi_2$, $\zeta_{\text{min}} = 0$, and $\zeta_{\text{range}} = 2$, an agent could for instance receive the evidence $(\varphi_1 \wedge \neg\varphi_2) \vee (\neg\varphi_1 \wedge \neg\varphi_2)$ $(\equiv \neg\varphi_2)$, or the weaker evidence $(\neg\varphi_1 \wedge \varphi_2) \vee (\neg\varphi_1 \wedge \neg\varphi_2) \vee (\varphi_1 \wedge \neg\varphi_2) \vee (\neg\varphi_1 \wedge \neg\varphi_2)$ $(\equiv \neg\varphi_1 \vee \neg\varphi_2)$, or also the very strong evidence $\varphi_1 \wedge \neg\varphi_2$ (i.e., τ).

To come to the belief update procedure, assume again some ordering of the 2^m state descriptions of $\mathcal{L}_m$, and define

$$A_i(t) := \sum_{j \in \mathcal{N}_i(t)} B_j(t) \; + \; \alpha \cdot |\mathcal{N}_i(t)| \cdot E_i(t). \tag{4}$$

Then agent i at t determines its new belief state $B_i(t+1)$ as follows:

$$B_i(t+1)[n] = \begin{cases} 1 & \text{if} \quad A_i(t)[n] > \theta, \\ 0 & \text{if} \quad A_i(t)[n] < \theta, \\ f(0,1) & \text{otherwise,} \end{cases} \tag{5}$$

where θ is a threshold value defined as $(\alpha + 1)|\mathcal{N}_i(t)|/2$, and $f(0,1)$ a function that randomly outputs either a 0 or a 1, with equal probability. A bit less formally, the

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

agent first aggregates into a vector $A_i(t)$ the belief states of his neighbors plus the evidence, weighted in a certain way. Then it compares each element $A_i(t)[n]$ of this vector against a threshold θ and it adopts the corresponding state description as a disjunct of the CDNF characterizing its new belief state iff either the element is above the threshold or it is equal to it and a flip with a fair coin turns out in favor of the element. Putting the last part a bit differently again, if $A_i(t)[n]$ is above θ, then the n-th element of the assumed ordering of state descriptions of $\mathcal{L}_m$ will be a disjunct of the CDNF of $B_i(t+1)$; if it is below the threshold, then it will *not* be among those disjuncts; and if it is equal to the threshold, it will be a disjunct of that CDNF depending on the outcome of a coin flip. As may be noticed, the foregoing leaves open the possibility that $B_i(t+1)$ is the inconsistent theory. But we may assume that an agent would notice if it is about to end up with inconsistent beliefs, and would want to avoid this. We thus add to the belief update mechanism the clause that if, according to (5), $B_i(t+1)[n] = 0$ for all $n \in \{1, \ldots, 2^m\}$, then the agent at $t+1$ randomly adopts a contingent theory in its language. Figure 2 presents the pseudo code of the belief update mechanism.

The parameter α determines the weight the evidence has in the belief update relative to the belief states of the agents that are both epistemic and spatial neighbors; it thereby plays a role in the current belief update mechanism that is in a way analogous to the role the parameter of the same name plays in the HK model (which is, of course, why we have chosen the same name). As a further parallel with that model, we note that while, in the new model, the agents do not literally adopt as their new belief state some (weighted) average of the average belief states of their neighbors and the evidence – as the agents in the earlier models do – they could still be said to average in a more metaphorical sense, by pooling the belief states of their neighbors and the weighted evidence and then distilling something like a majority belief state from this. To make the idea clearer, think of the belief update as a kind of voting procedure in which the agent's neighbors vote about each state description of the language, with the evidence also having a vote – in fact, α times the number of neighbors votes – and where a state description will end up as a disjunct in the CDNF representation of the agent's new belief state if, among these votes, there is a majority of *yea*'s for it. It is not too unnatural, it seems to us, to think of the outcome of this procedure as a sort of average of the individual belief states plus the weighted evidence.

To illustrate the belief update mechanism, still assume $\mathcal{L}_2$, and assume some ordering of its state descriptions. Furthermore, suppose that $\mathcal{N}_i(t) = \{i, j, k\}$, and let $B_i(t) = \langle 1, 0, 0, 0\rangle$, $B_j(t) = \langle 1, 0, 0, 1\rangle$, and $B_k(t) = \langle 1, 1, 0, 1\rangle$. Let $E_i(t) = \langle 0, 1, 0, 1\rangle$, and let $\alpha = 1$. Then

$$\begin{aligned} A_i(t) &= \langle 1,0,0,0\rangle + \langle 1,0,0,1\rangle + \langle 1,1,0,1\rangle + 1 \cdot 3 \cdot \langle 0,1,0,1\rangle \\ &= \langle 1,0,0,0\rangle + \langle 1,0,0,1\rangle + \langle 1,1,0,1\rangle + \langle 0,3,0,3\rangle \\ &= \langle 3,4,0,5\rangle. \end{aligned}$$

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

```
foreach agent of population {
    foreach state-description {
        aggregated-belief-states[state-description] = 0
        }
# loop through spatial neighbors including oneself
    foreach neighbor in spatial-neighborhood(agent) {
        if hamming-distance(belief-state[agent], belief-state[neighbor]) <= epsilon {
            num-neighbors ++
            foreach state-description of belief-state[neighbor] {
                aggregated-belief-states[state-description] ++
                }
            }
        }
# include evidence
    evidence = add random-num(zeta) state-descriptions as disjuncts to tau
    foreach state-description of evidence {
        aggregated-belief-states[state-description] =+ alpha * num-neighbors
        }
    num-neighbors =+ alpha * num-neighbors
# calculation new belief state
    threshold = num-neighbors / 2
    new-belief-state = [ ]
    foreach state-description {
        if aggregated-belief-states[state-description] > threshold {
            add state-description to new-belief-state
            }
        else if aggregated-belief-states[state-description] = threshold {
            if flip-coin = heads {
                add state-description to new-belief-state
                }
            }
        }
    if new-belief-state = [ ] {
        new-belief-state = contingent theory of random length
        }
    }
```

Figure 2. Pseudo code of belief update algorithm.

The threshold $\theta = (1 + 1) \cdot 3/2 = 3$. So, applying (5) yields either $B_i(t + 1) = \langle 1, 1, 0, 1\rangle$ or $B_i(t + 1) = \langle 0, 1, 0, 1\rangle$, depending on how the coin flip turns out for the first state description. Notice that if agent i's evidence at t had been, in vector notation, $\langle 0, 0, 1, 0\rangle$ instead of $\langle 0, 1, 0, 1\rangle$, the agent might have ended up with the inconsistent theory at $t + 1$, were it not for the additional clause that it would then randomly pick a contingent theory.

We have two comments on the foregoing. First, it may have been noticed that (5) leaves open the possibility that an agent's belief state at $t + 1$ is inconsistent with the evidence it received at t. Imagine, in the above example, that instead of the theory $\langle 0, 1, 0, 1\rangle$, agent i had received $\langle 0, 0, 1, 0\rangle$ as evidence at t and that $\alpha = .5$. Then, as one easily calculates, we would have had $B_i(t + 1) = \langle 1, 0, 0, 0\rangle$, which is inconsistent with the evidence $\langle 0, 0, 1, 0\rangle$. This may appear strange, as it would seem to amount to ignoring the evidence – the information that comes directly

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

from the world, so to speak – in favor of the beliefs of the neighbors. However, we should warn here against too literal an interpretation of the phrase "receiving evidence." We think of (4), and of the part the second summand plays in it, exactly as Hegselmann and Krause think of (1), and of the part this equation's second summand plays (see, e.g., their (2006, sect. 1)). Together the summands *capture* how the agents revise their beliefs, partly based on the beliefs of others, partly on evidence they get. But the equations do not state the belief update rules the agents *use* to update their beliefs (these rules are left unspecified). Just as in the HK model, where the agents do not directly perceive the value of τ, in our model the agents do not directly perceive the evidence; that an agent receives a certain piece of evidence is to be interpreted as meaning that that piece of evidence influences the agent's belief update in a way captured by (5).

Second, while the belief update mechanism stated above was motivated by a desire to stay conceptually as close as possible to the belief update mechanism of the HK model, it is not too hard to think of, and it seems worthwhile investigating, variant mechanisms. For instance, one can easily think of different ways of weighing the evidence relative to the belief states of the neighbors, or of differently defined thresholds, or of having the agents resolve tie situations (i.e., situations in which the number of *yea*'s for a given state description is equal to θ) in ways different from flipping a coin. And the Hamming distance is certainly not the only metric that can be defined on the set of theories. Furthermore, we took the evidence to be always entailed by the truth. One could extend the model by allowing for evidence that is, while consistent with, not entailed by the truth, or even for evidence that is inconsistent with the truth (misleading evidence). We experimented with some of the possible variants and extensions, but these yielded results that were not interestingly different from those to be presented below.[5]

4. COMPUTER SIMULATIONS

This section presents selected results from the computer experiments we performed in order to explore the above model of complex belief state dynamics. As intimated earlier, we used in all experiments the language $\mathcal{L}_4$. An experiment invariably starts with randomly selecting an element of $\mathcal{T}_4$ the value of τ, with the restriction that the CDNF of τ was to have at least one disjunct and at most eight (that is, half of 2 to the number of state descriptions of the language; we wanted to avoid truths that were exceedingly weak, though the exact number was chosen arbitrarily). Then agents are randomly placed on the 25 × 25 toroidal grid (see Section 2) and equipped with randomly generated theories whose CDNF's were to have at least as many disjuncts as the CDNF of τ (thus, the agents' initial belief states were never logically stronger than τ). Running the experiment meant applying repeatedly for a fixed number of time steps the belief update and migration algorithms to all agents. Each simulation was iterated a number of times with varying seeds for the random generator to ensure a sufficiently high

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

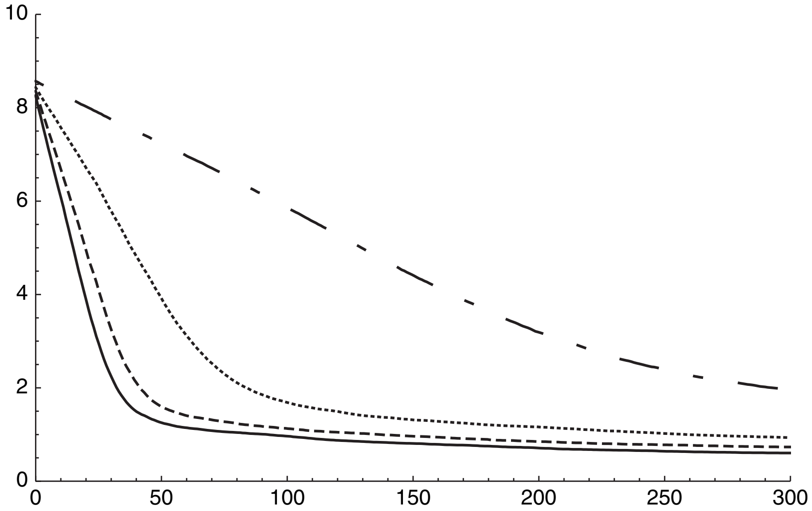

Figure 3. Convergence to τ for various neighborhood structures. The x-axis represents time steps, the y-axis is the average distance from τ.

variation of randomly generated truths, belief states, and evidence sentences. We were interested in the impact the model's core parameters – specifically, α, ε, and ζ – would have on the population's ability to track the truth, in particular on speed and accuracy of convergence to τ. In the experiments, we calculated averages over the different iterations of the average distance from the truth of the agents' belief states at each time. The average distance from the truth at t was simply defined as $(1/n)\sum_{i\leq n}\delta(B_i(t),\tau)$ for a population of n agents.

If not indicated otherwise, the following parameter settings were assumed. Number of time steps: 100; number of iterations: 100; type of neighborhood: Gaylord-Nishidate; number of agents: 100; $\alpha = .1$; $\varepsilon = 6$; and $\zeta = \langle 0, 3\rangle$.

4.1. Spatial neighborhoods and population density

Before focusing on the impact of the core parameters, we did some tests in order to determine, first, which of the various spatial neighborhood structures defined in Section 2 yields the best results in terms of speed and accuracy of convergence to τ, and second, whether population density is a factor in the same regard.

Based on the results obtained in the simpler two-dimensional model, we expected a more extensive neighborhood to do better in the said respects than a smaller one, which would favor the Gaylord-Nishidate neighborhood structure (comprising 13 sites) over Moore (9 sites) and the two von Neumann ones (each 5 sites). This was confirmed by the results of experiments, which are shown in Figure 3: the fastest convergence is achieved by the Gaylord-Nishidate neighborhood (solid), followed by Moore (dashed), von Neumann simpliciter (dotted), and von Neumann facing (dash-dotted). Even after 300 time steps,

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

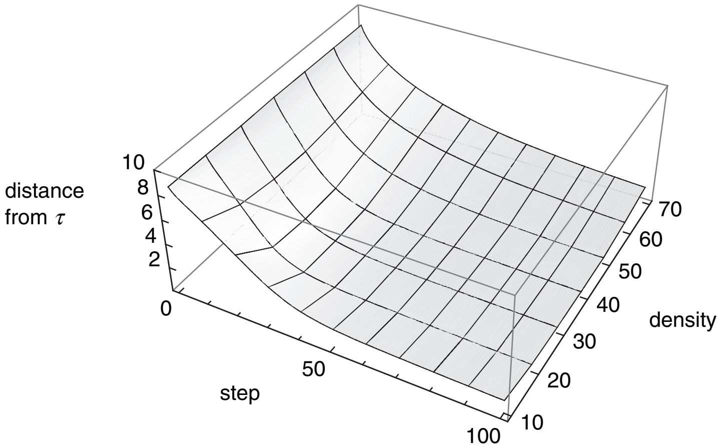

Figure 4. Convergence to τ for different population densities (in percent).

Gaylord-Nishidate keeps the lead in an increasingly decelerated approach toward τ. Because of this, we used the Gaylord-Nishidate neighborhood in all our experiments.

In order to determine the impact of the population density on convergence behavior, we systematically varied the population density between 10% and 70% in steps of 10%. As the environment is a 25×25 grid, thus consisting of 625 sites, this amounts to a spectrum ranging from 62 to 437 agents. As can be seen from Figure 4, this made very little difference in terms of accuracy and speed of convergence. We thus felt justified in keeping the number of agents fixed at 100 in all further experiments.

4.2. *Informativeness of the evidence*

To what extent does the informativeness of the evidence, as determined by the parameter ζ, influence the convergence behavior of populations of truth-seeking agents? To answer this question, we performed experiments in which we systematically varied ζ_{min} and ζ_{range} as discrete numerical values between 0 and 8.

In the experiments concerning ζ_{min}, we kept ζ_{range} at 3. The results, which are visualized by the left-hand graph in Figure 5, indicate that for small values of ζ_{min} there is a speedy and accurate convergence to τ, for middle values of ζ_{min} there is still convergence, albeit not quite as speedy and accurate, and for larger values of ζ_{min} there is no convergence to τ at all, but even a slight moving away from τ. These findings turned out to be largely insensitive to the value of α, at least as long as this remained between 0 and 1. The right graph compares the situation after $t = 100$ for $\alpha = .1$ with the situation after $t = 100$ for $\alpha = 1.1$.

The results clearly suggest that as long as ζ_{min} is relatively low, epistemic interaction helps populations of truth-seeking agents to get closer to the truth

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

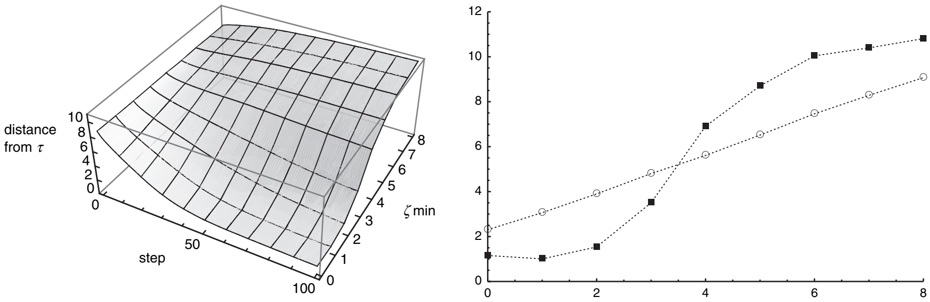

Figure 5. Left: Variation of $\zeta_{\min}$ from 0 to 8, with $\zeta_{\text{range}} = 3$. Right: the situation after $t = 100$ for $\alpha = .1$ (squares) and $\alpha = 1.1$ (circles); the x-axis represents $\zeta_{\min}$, the y-axis distance from τ.

than one would expect them to get purely on the basis of the evidence, given the informativeness of the evidence. For instance, even with $\zeta_{\min} = 0$, the evidence the agents get will still be at a Hamming distance of, on average, $\sum_{i=0}^{2} i \binom{16}{i} / \sum_{i=0}^{2} \binom{16}{i} \approx 1.87$ from τ (given that $\zeta_{\text{range}} = 3$). Nevertheless, after 100 time steps, the agents are only at an average distance of approximately .7 from τ. Even more remarkably, with $\zeta_{\min} = 3$, the evidence the agents get will be at a Hamming distance of, on average, $\sum_{i=3}^{5} i \binom{16}{i} \,/\, \sum_{i=3}^{5} \binom{16}{i} \approx 4.56$ from τ, and still the agents end up, after 100 time steps, at an average distance of only approximately .96 from τ. However, from the right-hand graph in Figure 5, it emerges that for $\zeta_{\min} > 3$, interaction is actually counterproductive: by purely going by the evidence – which is what happens if $\alpha > 1$ (see Section 4.4) – the agents do better in terms of accuracy of convergence.

We strongly suspect that the positive effect of interaction for lower values of $\zeta_{\min}$ is due to a parallel of the "averaging effect" pointed to in Douven (2009) for the case where agents receive noisy data about the numerical value of a given parameter. The noise is, in the variant of the HK model considered there, spread out randomly but evenly around that value. By epistemically interacting and thereby adapting their beliefs to a middle value, the agents to a large extent annihilate the random noise – the randomness gets "averaged out," so to speak. The present results give reason to believe that even though for sets of sentences there can be no averaging in the literal sense of the word, something very much *like* an averaging mechanism is operative in the new model. Evidently, this requires further investigation, which must await another time, however.

As for why at around $\zeta_{\min} = 4$ the convergence gets so much worse, the best explanation would seem to involve the fact that not only will the Hamming distance between τ and the pieces of evidence the agents receive, on average, increase as $\zeta_{\min}$ goes up, the average Hamming distance between these pieces of evidence themselves will, up to $\zeta_{\min} = 7$, also increase.[6] For thereby the evidence may fail

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

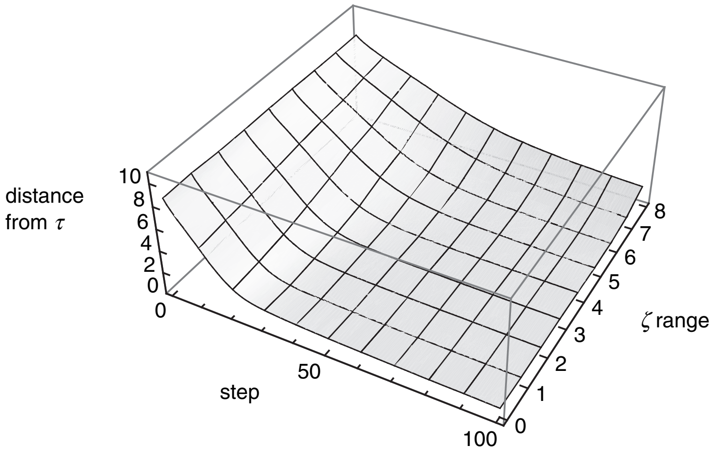

Figure 6. Variation of ζ_{range} from 0 to 8, with $\zeta_{\text{min}} = 0$; $\alpha = .1$.

to bring the agents sufficiently close to each other for them to count as epistemic neighbors, so that the (hypothesized analogue of the) averaging effect peters out. Something similar may explain why for $\zeta_{\text{min}} \geq 6$, there is a slight moving away from τ. Agents with belief states that are initially closer than 6 to τ may gradually be pulled away from the truth by the evidence – which from $\zeta_{\text{min}} \geq 6$ onward is on average at a distance of at least (approximately) 7.15 from τ – without this being sufficiently counteracted by the averaging effect of epistemic interaction, as too few spatial neighbors will also be epistemic neighbors; thereby the population's average distance from τ comes to exceed its average initial distance from τ.[7]

Whereas a high ζ_{min} keeps all agents at a certain distance from τ, a high $\zeta_{\text{max}} = \zeta_{\text{min}} + \zeta_{\text{range}}$ is compatible with agents' receiving evidence very close to τ (they may even receive evidence identical to τ, if $\zeta_{\text{min}} = 0$), such that at any time there may remain sufficiently many agents close to τ to guarantee that subsequent interaction pulls the unfortunate agents who got highly uninformative evidence back toward τ, or in any event that it pulls enough of such agents back to keep the population's average distance from τ low. One might thus expect that, provided ζ_{min} is low, even with $\zeta_{\text{range}} = 8$, the convergence behavior should not be as drastically impaired as in situations in which ζ_{min} is high. The experiments in which we varied ζ_{range} while keeping ζ_{min} at 0 confirmed this expectation. In fact, while the speed of convergence decreases slightly with higher ζ_{range}, in all settings for ζ_{range}, after 100 time steps the agents end up with belief states quite close to τ (see Figure 6 for the results).

The upshot of these experiments is that whereas a higher *maximal* deviation of the evidence from τ delays the convergence but does not make it less accurate in the end, a higher *minimal* value for the deviation (as determined by ζ_{min}) has a dramatic impact on the population's ability to converge toward the truth.

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

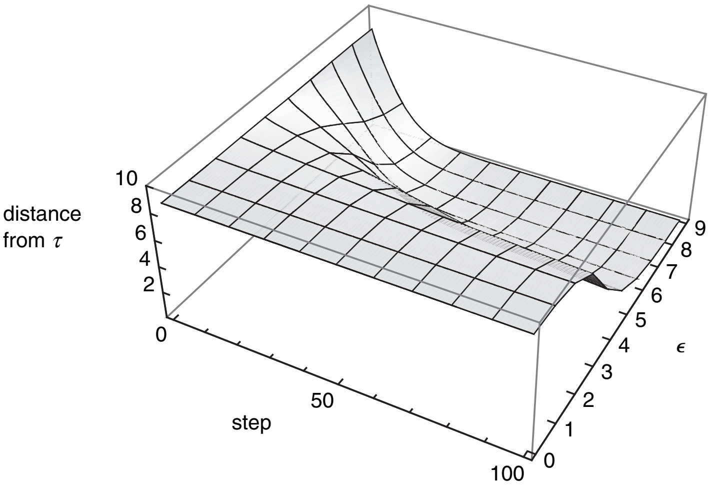

Figure 7. Varying ε; $\alpha = .1$.

4.3. Varying the value of ε

People may embrace a more closed-minded or a more open-minded approach in interacting with others, in the sense that they may be more or less ready to also take into account belief states that are at greater variance with their own. In the HK model, the parameter ε determines how open-minded (or closed-minded) the agents are; in our model of complex belief state dynamics, this is determined by the parameter of the same name. Hegselmann and Krause's work on the former model shows that even small differences in the size of (their parameter) ε can have a significant effect on the long-run development of the agents' beliefs (e.g., on whether these beliefs converge to a common value, or whether they polarize, or whether they group together into various clusters). As Figure 7 suggests, our model is also very sensitive to the size of (our parameter) ε. Populations in which agents interact only with their closest epistemic neighbors ($\varepsilon = 3$) show little or no convergence, whereas populations of more open-minded agents ($\varepsilon = 6$) greatly benefit from interacting. (The Figure shows results only for $\alpha = .1$, but further experiments revealed that the impact of the size of ε is relatively independent of the value for α.)

4.4. Varying the value of α

The value of α determines the weight the agents attribute to the evidence they receive relative to the weight they attach to the beliefs of the epistemic neighbors they happen to meet. We investigated the influence of this parameter on the speed and accuracy of convergence.

Given that the evidence is unable to exert any force on the agents' belief states when $\alpha = 0$, we cannot expect any convergence in this situation; unsurprisingly, that is also what computer experiments show. As for $\alpha < 1$, it is easy to see that

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

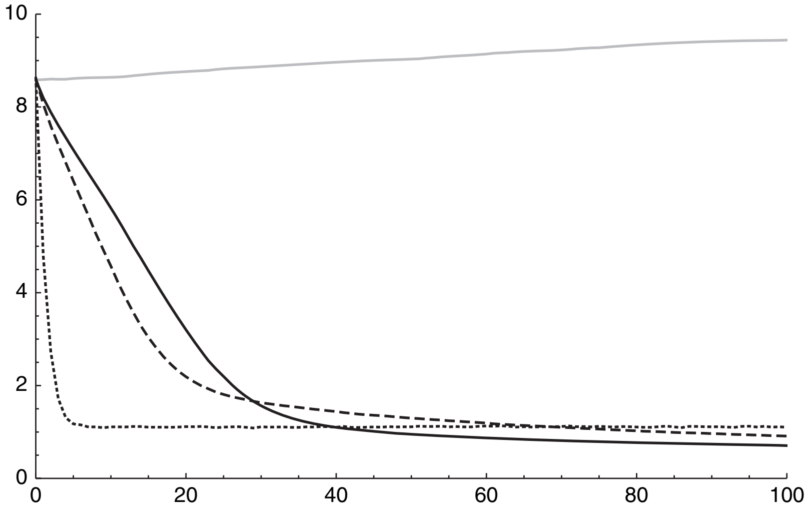

Figure 8. Selected values of α: 0 (gray), .1 (black), .6 (dashed), 1 (dotted).

$B_i(t+1) = E_i(t)$ for all i and t. We would thus expect a rapid convergence to the average distance of the evidence from τ. This was, in effect, already confirmed by the outcomes of the experiments we performed to produce the right-hand graph in Figure 5. For $\alpha = 1$, we would expect something at least very similar. This, too, was confirmed.

It is more interesting (for not quite as obvious) to see that for intermediate values of α, there is a slow but ultimately very accurate convergence – consistently more accurate than for $\alpha \geq 1$ – where the convergence occurs at a slower pace, but is also more accurate (even though here the differences are small), the lower α is. Figure 8 shows the results for some selected values of α. Again, the key to explaining these results may be the presumption that in the present model too, some sort of "averaging out" of the random deviations of the evidence from τ takes place through interaction. In the models studied in earlier papers, the averaging effect proved to be stronger – in the sense that it brings about the most accurate convergence – the more weight the agents, in updating, attribute to the belief states of their neighbors relative to the weight they attribute to the evidence. However, it was also shown that in these models, the more relative weight the agents attribute to the belief states of their neighbors, the slower the convergence occurs. That here we see exactly the same happening – convergence is slower but also more accurate the more relative weight the agents attribute to their neighbors' belief states – is a further strong reason to suspect the presence of an averaging effect.

5. CONCLUSION

The model described in this paper is the outcome of an attempt to generalize the popular approach to opinion dynamics of Hegselmann and Krause to richer and

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

more interesting belief states. We believe that developing and exploring such a model is, to a large extent, a project that is valuable in its own right. Nevertheless, the results so far obtained about the new model also highlight its philosophical relevance. In particular, a major lesson that can be learned from these results, and that should be of immediate concern to social epistemologists, is the following. Whilst being open to interaction with other agents and giving some weight to their beliefs helps agents, at least on average, to track the truth more accurately, it also slows them down in getting within a moderately close distance from the truth as compared to when they go purely by the evidence. Consequently, no general conclusion can be drawn about whether, say, it is a good epistemic strategy to attribute, in updating, relatively much weight to the belief states of one's respected colleagues. It all depends, our studies seem to suggest. After all, whereas sometimes it can be important to come very close to the truth, even if this should take longer, at other times it may be more important to get *somewhat* close to the truth relatively quickly, but it would be of little or no further advantage to get still closer to the truth. This was one of the main conclusions we reached in our earlier papers, on the basis of the simpler models described in Section 2. It is encouraging to see that it could be reconfirmed by studies using a model that in various seemingly important ways is more realistic than the aforementioned ones.[8]

REFERENCES

Deffuant, G., D. Neau, F. Amblard, and G. Weisbuch. 2000. "Mixing Beliefs among Interacting Agents." *Advances in Complex Systems* 3: 87–98.

Dittmer, J. C. 2001. "Consensus Formation under Bounded Confidence." *Nonlinear Analysis* 7: 4615–21.

Douven, I. 2009. "Simulating Peer Disagreements." Manuscript.

Douven, I. and A. Riegler. 2009. "Extending the Hegselmann-Krause Model I." *Logic Journal of the IGPL*, in press.

Fortunato, S. 2005. "On the Consensus Threshold for the Opinion Dynamics of Krause-Hegselmann." *International Journal of Modern Physics C* 16: 259–70.

Gaylord, R. J. and L. J. D'Andria. 1998. *Simulating Society*. New York: Springer.

Hegselmann, R. and U. Krause. 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. and U. Krause. 2005. "Opinion Dynamics Driven by Various Ways of Averaging." *Computational Economics* 25: 381–405.

Hegselmann, R. and U. Krause. 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

Jacobmeier, D. 2004. "Multidimensional Consensus Model on a Barabási-Albert Network." *International Journal of Modern Physics C* 16: 633–46.

Kuipers, T. A. F. 2000. *From Instrumentalism to Constructive Realism*. Dordrecht: Kluwer.

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

Lorenz, J. 2003. *Mehrdimensionale Meinungsdynamik bei wechselndem Vertrauen*. Diploma thesis, University of Bremen. http://nbn-resolving.de/urn:nbn:de:gbv:46-diplom000000564

Lorenz, J. 2007. "Continuous Opinion Dynamics under Bounded Confidence: A Survey." *International Journal of Modern Physics C* 18: 1819–38.

Lorenz, J. 2008. "Fostering Consensus in Multidimensional Continuous Opinion Dynamics under Bounded Confidence." In D. Helbing (ed.), *Managing Complexity*, pp. 321–34. Berlin: Springer.

Pluchino, A., V. Latora, and A. Rapisarda. 2006. "Compromise and Synchronization in Opinion Dynamics." *European Physical Journal B* 50: 169–76.

Ramirez-Cano, D. and J. Pitt. 2006. "Follow the Leader: Profiling Agents in an Opinion Formation Model of Dynamic Confidence and Individual Mind-Sets." *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pp. 660–7.

Riegler, A. and I. Douven. 2009. "Extending the Hegselmann–Krause Model II." In K. Kijania-Placek (ed.), *Proceedings of ECAP6*. London: College Publications, in press.

Weisbuch, G., G. Deffuant, F. Amblard, and J. P. Nadal. 2002. "Meet, Discuss and Segregate!" *Complexity* 7: 55–63.

NOTES

1 See, e.g., Deffuant et al. (2000); Dittmer (2001); Hegselmann and Krause (2002, 2005, 2006); Weisbuch et al. (2002); and Ramirez-Cano and Pitt (2006). For an excellent overview of the main technical results in this area, see Lorenz (2007).

2 Not all extant models of opinion dynamics are idealized in this way. For instance, Deffuant, Weisbuch, and colleagues (Deffuant et al. 2000; Weisbuch et al. 2002) have formulated a model in which agents meet only pairwise in a random manner and then compromise their beliefs if they happen to be (in our terminology) epistemic neighbors. Fortunato (2005) implements a static social network that defines who may talk to whom.

3 For the other two neighborhood structures, this extra "facing" condition obviously makes no sense.

4 This metric is also used by Deffuant et al. (2000) in their model of opinion dynamics.

5 This is, for instance, true of the variant model that adds a clause to (5) similar to the lower clause of (3), that is, the belief update mechanism is invoked only if there are epistemic neighbors in the agent's spatial neighborhood other than the agent itself.

6 Supposing $\mathcal{L}_4$ and $\zeta_{\mathrm{range}} = 3$, for $\zeta_{\mathrm{min}} = n$ the average distance between theories that can be generated as evidence is

$$\Bigg[\binom{16}{n}\left(\sum_{j=0}^{2}\sum_{i=0}^{n}\binom{n}{i}\binom{16-n}{i+j}(2i+j)\right)+\binom{16}{n+1}\left(\sum_{i=0}^{n+1}\left(\binom{n+1}{i}\right.\right.$$

$$\left.\left.\times\binom{15-n}{i+1}(2i+1)+\binom{n+1}{i+1}\binom{15-n}{i}(2i+1)+\binom{n+1}{i}\binom{15-n}{i}2i\right)\right)$$

$$+\binom{16}{n+2}\left(\sum_{j=0}^{2}\sum_{i=0}^{n+2}\binom{n+2}{i+j}\binom{14-n}{i}(2i+j)\right)\Bigg]\Bigg/3\sum_{i=0}^{2}\binom{16}{n+i}.$$

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press

For $n = 0$, this yields (approximately) 2.6; for $\zeta_{\min} = 7$ the function reaches it maximum, which is 8.

[7] Here, of course, only the average distance from τ after 100 time steps is of real significance. That this exceeds the average distance of the agents' initial belief states from τ is an artifact of our working in the language $\mathcal{L}_4$. After all, this initial average depends on the number of atomic sentences.

[8] We are greatly indebted to Christopher von Bülow for very helpful comments on a previous version of this paper.

Alexander Riegler works at the University of Leuven on computational opinion dynamics. He obtained a PhD in Artificial Intelligence and Cognitive Science from Vienna University of Technology in 1995 with a dissertation on Artificial Life. Riegler's interdisciplinary work includes diverse areas, such as knowledge representation and anticipation in cognitive science, post-Darwinian approaches in evolutionary theory, and constructivist and computational approaches to epistemology. He is the editor-in-chief of the journal *Constructivist Foundations*, http://www.constructivistfoundations.info.

Igor Douven is a professor of philosophy at the University of Leuven. His areas of specialization are epistemology, philosophy of science, and philosophy of language. Recent research concerned probabilistic approaches to coherence, the nature of assertion, paradoxes of rational acceptability, and the possibility of rational disagreement. He is the director of the Formal Epistemology Project, which is funded by the Flemish government.

https://doi.org/10.3366/E1742360009000616 Published online by Cambridge University Press