*Journal of Artificial Societies and Social Simulation* vol. 9, no. 3. 2006
http://jasss.soc.surrey.ac.uk/9/3/1.html

# Truth and Cognitive Division of Labour
# First Steps towards a Computer Aided Social Epistemology

Rainer Hegselmann
Department of Philosophy, University Bayreuth

Ulrich Krause
Department of Mathematics, University Bremen

**Abstract:** The paper analyzes the chances for the truth to be found and broadly accepted under conditions of cognitive division of labour combined with a social exchange process. *Cognitive division of labour* means, that only some individuals are active truth seekers, possibly with different capacities. The *social exchange process* consists in an exchange of opinions between all individuals, whether truth seekers or not. We develop a model which is investigated by both, mathematical tools and computer simulations. As an analytical result the *Funnel theorem* states that under rather weak conditions on the social process a consensus on the truth will be reached if all individuals posses an arbitrarily small inclination for truth seeking. The *Leading the pack theorem* states that under certain conditions even a single truth seeker may lead all individuals to the truth. Systematic simulations analyze how close and how fast groups can get to the truth depending on the frequency of truth seekers, their capacities as truth seekers, the position of the truth (more to the extreme or more in the centre of an opinion space), and the willingness to take into account the opinions of others when exchanging and updating opinions. A tricky movie visualizes simulations results in a parameter space of higher dimensions.

**Keywords:** opinion dynamics, consensus/dissent, bounded confidence, truth, social epistemology, division of labour.

## 1. Introduction: Opinion Dynamics and Truth

There is an ongoing research on opinion dynamics since almost 50 years.[1] Research started with French (1956). Important early steps did Harary (1959), Abelson (1964), De Groot (1974), Chatterjee (1975), Chatterjee/Seneta (1977), Lehrer (1975) and Lehrer/Wagner (1981). More recent work was done by Galam et al. (1982), Galam / Moscovici (1991), Friedkin / Johnsen (1990), Nowak et al. (1990), Krause (1997), Hegselmann et al. (1999), Deffuant et al. (2000), Dittmer (2001), Weisbuch et al. (2001), Hegselmann / Krause (2002, 2004), Stauffer (2002, 2003, 2004), Lorenz (2003, 2005), Fortunato (2004, 2005) and Fortunato et al. (2005).

Roughly, the following three main lines of study can be distinguished. There is an approach mainly developed by physicists and based on statistical mechanics. A second route emphasizes the social structure underlying opinion formation. A third line analyzes opinion formation as a discrete dynamical system. That is our approach. It provides a formal framework that covers various kinds of opinion formation among a set $I$ of $n$ individuals and a discrete time $t$ = 0, 1, 2, … . Each individual starts in $t$ = 0 with a certain opinion, expressed by a real number $x_i(0) \in ]0, 1]$.[2] An opinion dynamics is then given by

[1] $$x_i(t+1) = f_i(x(t)), \text{with } 1 \le i \le n .$$

[1] For an historical and systematic overview cf. Hegselmann / Krause (2002, ch. 1-3).

[2] It is convenient to scale opinions to have values in the unit interval. To admit for the formation of means like the geometric mean we exclude the value 0.

$f_i$ is the decisive *function* – so far not specified – that operates on the opinion profile

$$x(t) = (x_1(t), \ldots, x_i(t), \ldots, x_n(t)).$$

By specifying $f_i$ a good deal of opinion dynamics mentioned beforehand can be covered.

Over the last ten years or so research on opinion dynamics intensified dramatically. Lots of papers published since then study versions of the so called *bounded confidence*–model (*BC*–model), a label coined by Krause 1998 (see Krause 2000). In that model individuals take into account the opinions of those others whose opinions are not 'too strange', i.e. not too far away from their own opinion.[3] The set of those others that individual $i$ takes seriously is the subset of those others $j \in I$ for which the distance in opinion, i.e. $|x_i(t) - x_j(t)|$, is not greater than a certain $\varepsilon$, the *confidence level.* Updating one's opinion consists in averaging over all opinions within that distance.[4]

In this paper we extend and modify the *BC*–model in a way that it *explicitly* covers a process in which truth seekers go for the truth. We will do that sticking to our macroscopic, non–normative approach guided by the *KISS*–heuristics "Keep it simple, stupid!" As a consequence, we will not even try to model explicitly all the processes and actions in the quest for truth (rational argumentation, reasonable thinking, weighing evidences, making experiments etc.). What we do instead is to cover all which might be important by the assumptions that *there is a true value T* in our opinion space ]0,1] and that this true value $T$ somehow '*attracts'* the individuals – at least some and to some degree. Our decisive equation is

$$[2] \qquad x_i(t+1) = \alpha_i T + (1-\alpha_i) f_i(x(t)), \text{with } 1 \le i \le n\,.$$

In equation [2] the opinion of individual $i$ is given by a convex combination with $0 \le \alpha_i \le 1$. The first part $\alpha_i T$ is the *objective* component. $\alpha_i$ controls the *strength* of the attraction of truth. The second part with the weight $(1-\alpha_i)$ is the *social* component, with $f_i$ being a function of the opinion profile $x(t) = (x_1(t), x_2(t), \ldots, x_{n-1}(t), x_n(t))$. The whole profile of functions in the social component is $f = (f_1, f_2, \ldots, f_{n-1}, f_n)$. To that profile we will refer as the *social process*. For $\alpha_i = 0$ we get again our original dynamics governed by [1], i.e. a situation in which only a social process is at work and the truth does not play any role. The framework offers a fairly

[3] Thus the model assumes agents which somehow and 'their way' solved the problem Isaac Levi emphasizes: "The mere fact that someone disagrees with one's judgement is insufficient grounds for opening one's mind. Most epistemologists forget that it is just as urgent a question to determine when we are justified in opening up our minds as it is to determine when we are justified in closing them" (Levi 1985, 3).

[4] The *BC*–model was formulated and studied for the first time by Krause (1997). Krause (2000) shows for certain conditions the convergence to consensus for a general non autonomous model of opinion dynamics; the BC–model is analysed as a special case. Beckmann (1997) and Dittmer (2000 and 2001) take up the idea and do some further steps towards an analytical understanding of the model. Hegselmann / Krause (2002) combine an analytical *and* simulation based approach. The result is a very comprehensive understanding of the *BC*–model. The understanding includes the effects of different sizes of the confidence intervals, different types of biased confidence intervals (e.g.: leftists 'listen' more to the left, rightists 'listen' more to the right) or multidimensional opinion spaces. Lorenz (2003) presents simulation results of the model's behaviour for up to four opinion dimensions. Lorenz (2003a) analyses the effects of heterogeneous confidence levels. Hegselmann (2004) proposes, describes and partially analyses several extensions and modifications of the *BC*-model, for instance effects of local interactions, different updating procedures, overlaying opinion–formation–mechanisms etc. Since the arithmetic mean is only one of the infinite numbers of possible means Hegselmann / Krause (2004) analyse the often dramatic effects of different ways of averaging for the *BC*-model. Fortunato (2004) uses discrete opinions. Lorenz (2005) proves a stabilization theorem. Urbig / Lorenz (2004) study systematically the effects of different updating procedures. Fortunato (2005) analyses the *BC*–model in a 2-dimensional opinion space.

A model very similar to Krause's original *BC*–model in Krause (1997) was later introduced and investigated in (Weisbuch et al. 2001) and (Deffuant et al. 2000). The difference to the original *BC*–model is that the authors use a pairwise–sequential updating procedure instead of simultaneous updating.

natural (though abstract) understanding of *cognitive division of labour*: By cognitive division of labour we refer to a situation in which *only some* individuals $i$ have an $\alpha_i > 0$.

Since equation [2] is threatened with misunderstandings some clarifying remarks on the interpretation and status of [2] will be helpful:

- *Could* [2] *be a mechanism intentionally followed by our individuals?* No! The simple reason is that whoever understands the concept of truth and somehow knows $T$ would immediately believe that $T$, i.e. update $x_i(t+1) = T$. The interpretation of [2] should be about the following: Individuals with a positive $\alpha$ have access to or generate new data (arguments, evidences, test results etc.) that point in the direction of $T$. Nevertheless their and their fellows' prior opinions are in their mental or cognitive backpack and influence them as well.
- *What is assumed on the nature of truth?* It is assumed, that the truth is one and only one, that it does not change over time, that it is somewhere in the opinion space, and that it is independent of the opinions the individuals hold. The truth influences opinions, but the *opinions don't have any impact on the truth*. Ever since the beginning of philosophy there has been a discussion on the nature of truth, its general concept and status, definition, criteria, indicators etc. The commonsensical assumptions on the truth made above are *not* compatible with *all* known conceptions of truth. They are compatible with a *correspondence theory of truth*, but so called *deflationists* shouldn't have any problem to accept them as well.[5] Thus, the problem with [2] is not so much a built–in partisanship in an ongoing debate on the nature of truth. (More on compatibility problems will follow when some central features of the dynamics driven by [2] are analyzed.)
- *How to make sense of an 'attraction of truth'?* For $\alpha_i > 0$ truth attracts in a *technical* sense. But this technical feature of the equation is used to generate a process which *must not* be interpreted in such a way that truth *attracts* anybody in a literal sense. The *technical* attraction of truth is used to model individuals, which to a certain degree *successfully aim at the truth*.[6] Thus, a positive $\alpha_i$ could be interpreted as the combined effect of education, training, profession, interest, and some epistemic success based on that.[7] A problem is that according to [2] the 'attraction of truth' works smoothly in the direction of $T$. But new evidences might be ambiguous. They may point into different directions or indicate only that the truth is *not* in a certain region of the opinion space. Therefore, [2] should be taken as a *starting point*. Later we (or others) can focus on more complicated epistemic situations.

Equation [2] is a convex combination of what we called an *objective* and a *social* component. The social component may actually be the *BC*–model as characterized above. But obviously other social processes could be brought into play as well[8]. An interesting *alternative* social process is proposed in the book *Rational Consensus in Science and Society*, written by Lehrer and Wagner (1981).[9] The book received major attention especially among philosophers (cf. Loewer 1985; Bogdan 1981). It presents a mechanism driven by *iterated weighted averaging*. The weights reflect the *respect* an individual assigns to other individuals. As in the *BC*–model the opinions are real numbers taken from the unit interval. The weights an individual assigns to others are assumed to sum up to 1.

[5] For an overview see the articles on truth in the web based *Stanford Encyclopedia of Philosophy* at http://plato.stanford.edu/contents.html

[6] For the debate on truth approximation cf. Kuipers 2000. For the philosophical debate on truth directedness cf. Engel 2004 and the literature mentioned in that article.

[7] Following our *macroscopic* approach we do not go into details.

[8] Especially if one thinks – as we do – that there are many different and interfering mechanisms that drive real-world opinion dynamics.

[9] Cf. Lehrer 1975, 1976, 1977, 1981 and 1981a.

We will refer to this model as the Lehrer/Wagner–model (*LW*–model). Formally it is a *dynamical* system, though – and that is important to notice – the authors *do not interpret it as a process over time*. Their *starting* point is a "*dialectical equilibrium*" i.e. a situation *after* "the group has engaged in extended discussion of the issue so that all empirical data and theoretical ratiocination has been communicated. … the discussion has sufficiently exhausted the scientific information available so that further discussion would not change the opinion of any member of the group" (Lehrer/Wagner 1981, 19). The central question for Lehrer and Wagner then is: Once the dialectical equilibrium is reached, is there a *rational* procedure to *aggregate* the normally still divergent opinions in the group (cf. Lehrer 1981b, 229)? Their answer is "Yes". The basic idea for the procedure is to make use of the fact that normally we all do not only have opinions but also information on expertise or reliability of others. That information can be used to assign weights to other individuals. The whole aggregation procedure is then *iterated weighted* averaging with $t \to \infty$ and based on *constant* weights. It is shown that for lots of weight matrices the individuals reach a consensus whatever the initial opinions might be – if they only were willing to apply the proposed aggregation procedure.[10]

Under our interpretation [2] provides a very general and simple formal framework for *the study of truth seeking individuals embedded in a community and a cognitive past or tradition*. The *BC*– and the *LW*–model are different instances for the social component. That makes the article a contribution to what is now called *social epistemology* (cf. Schmitt 1994 and 1999; Goldman 1999, 2001; Fuller 2002). Under the perspectives of accepted methods, relevant questions or common goals social epistemology is more a battlefield than a well established discipline.[11] Nevertheless, topics like the cognitive interaction between truth seeking individuals or the proliferation of the truth in societies only in parts interested in the truth, belong to the set of paradigm topics for the project 'social epistemology'.

In what follows we pursue our studies both ways, by *analytical* means and by means of *simulations*. *Chapter 2* will demonstrate some interesting effects by *single run* simulations of opinion dynamics driven by an objective *and* a social component as outlined in equation [2]. Thereby we open the field of ***C**omputer **A**ided **S**ocial **E**pistemology* (*CASE*) with some first *CASE*–studies. In the simulations the social component will be given by the *BC*–model. *Chapter 3* gives a short and easy to understand rigorous proof for what we call the *Funnel theorem.* The theorem states conditions under which a community will end up at the truth. *Chapter 4* formulates a very general theorem that states conditions under which even a single truth seeker can drive a whole society to the truth. In *chapter 5* the single run *CASE*–studies of chapter 2 are extended to *systematic CASE*–studies of a whole parameter space. We analyze the *chances for truth proliferation* under certain conditions: What if we have a *division of labor* in the sense that *only some* go for the truth, i.e. *not all* have a positive $\alpha_i$? How does the *position of the truth* – more in the centre or more to the extremes of the opinion space – matter? Additionally, our truth seeking individuals may be *more or less 'good' in getting to the truth*, i.e. their $\alpha_i$ may be small or comparatively great. How does that matter? And how is that all influenced by the *size of the confidence intervals*? Finally a tricky *movie* shows under what conditions societies are how fast and good in getting to the truth.

[10] For analytical results cf. as well Hegselmann /Krause 2002, ch. 2.

[11] Goldman (2001) starts his article "*Social Epistemology*" in the *Stanford Encyclopedia of Philosophy* with the following sentences: "Social epistemology is the study of the social dimensions of knowledge or information. There is little consensus, however, on what the term "knowledge" comprehends, what is the scope of the "social", or what the style or purpose of the study should be. According to some writers, social epistemology should retain the same general mission as classical epistemology, revamped in the recognition that classical epistemology was too individualistic. According to other writers, social epistemology should be a more radical departure from classical epistemology, a successor discipline that would replace epistemology as traditionally conceived."

In *chapter 6* we summarize and give an outline for further research questions and perspectives. The detailed proof for the theorem of chapter 4 is put into an *appendix*.

## 2. *CASE*–Studies I: Opinion Dynamics among Truth Seeking Agents[12]

Throughout this chapter we assume a social process driven by the *bounded confidence* mechanism. Thus an agents *i* updates his opinions by averaging over the opinions of all those agents *j* for which $| x_i(t) - x_j(t) | \le \varepsilon$, thereby generating a *symmetric* confidence interval to the left and to the right of agent *i*'s opinion. The confidence level $\varepsilon$ is supposed to be the *same* for all agents (*homogeneity*) and *constant* over time. We assume *simultaneous* updating.

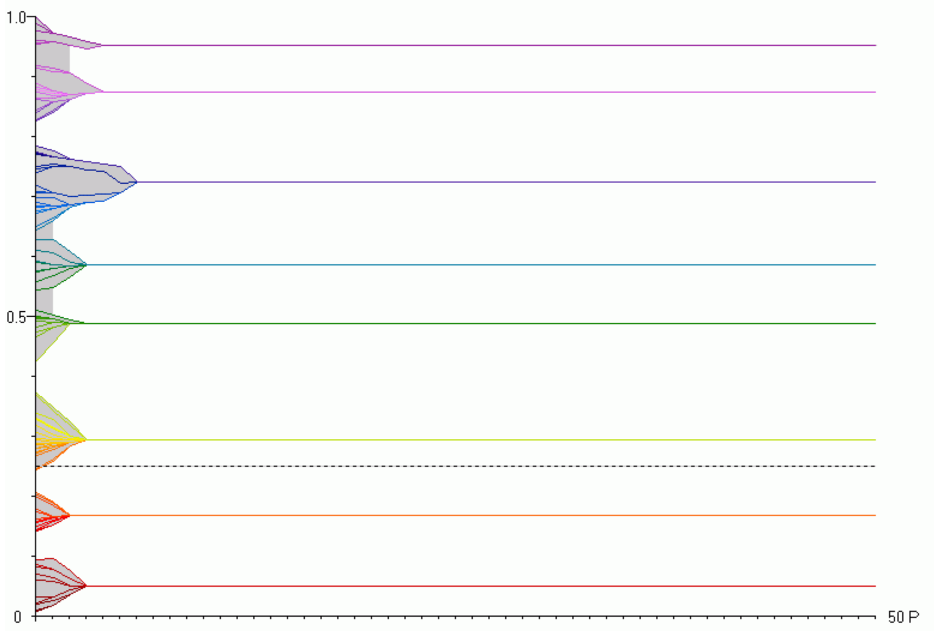


***Figure 1:*** 100 agents, random start distribution, $\varepsilon = 0.05$. $\alpha_i = 0$ for all agents *i*. (We show 50 periods and not only the very first to allow for a better comparison with *figure 2* in which all individuals are truth seekers while the start profile is the same as in *figure 1*.)

*Figure 1* shows the opinion dynamics for 100 agents and a confidence level $\varepsilon = 0.05$. The process starts with an even random distribution. We assume $\alpha_i = 0$ for all agents *i*. Thus *none* of the agents is a successful truth seeker. The ordinate indicates the opinions. Since it is useful for the following analysis the opinions are additionally encoded by colours ranging from red (x = 0) to magenta (x = 1). The abscissa represents time and shows the first 50 periods. The trajectory of an opinion gets a colour according to its start value. The colouring of the trajectories is done following the order of the profile. The dotted line is the assumed position of the truth ($T = 0.25$) nobody is going for. A *grey area* between two neighbouring opinions indicates that the distance between the two is not greater than $\varepsilon$. Note that under simultaneous updating, given the start profile is *ordered* in the sense that $x_1(0) \le x_2(0), \ldots , x_{n\text{-}1}(0) \le x_n(0)$, that order will be preserved over time.[13] If in an ordered profile the distance between two neighbouring opinions is never greater than $\varepsilon$ then we will call such a profile an *ε–profile*. If the area between two neighbouring opinions is *not* grey, that indicates something we refer to as an *ε–split* in the profile. Based on that terminology we can easily describe and understand

[12] In philosophy nobody would refer to individuals as agents. Individuals 'are' individuals or actors. But in the world of modelling and simulating it is common to refer to individuals as agents. Since this chapter is less philosophical and more simulation oriented we switch to the terminology used in that field.

[13] That is *not* true for pair wise sequential updating. The trajectory of an opinion gets a colour according to its start value. The colouring of the trajectories is done following the order of the profile.

what is going on in *figure 1*:[14] Not even at the start the opinion profile is an $\varepsilon$–profile. Split sub–profiles belong to different 'opinion worlds' or communities which *do not (or: no longer) interact*. The *extreme* opinions of a (split sub–)profile are under a *one sided influence* and move direction centre of the (sub–)profile. As a consequence the range of the (sub–)profile *shrinks*. At the extremes opinions *condense*. Condensed regions may *attract* opinions from less populated areas within their $\varepsilon$–reach. As a consequence (sub–)profiles may *split again.* Finally the social process gets to a point where in (split sub–)profiles everybody is within the $\varepsilon$–reach of everybody else. The logical consequence is that all involved opinions 'collapse' into *one* opinion in the next period. As of then this part of the profile is stable for ever. Stability is always reached in *finite* time (cf. Hegselmann/Krause 2002, ch. 3 result 5 and 6). In *figure 1* stability means an *eternal plurality of divergent* views.

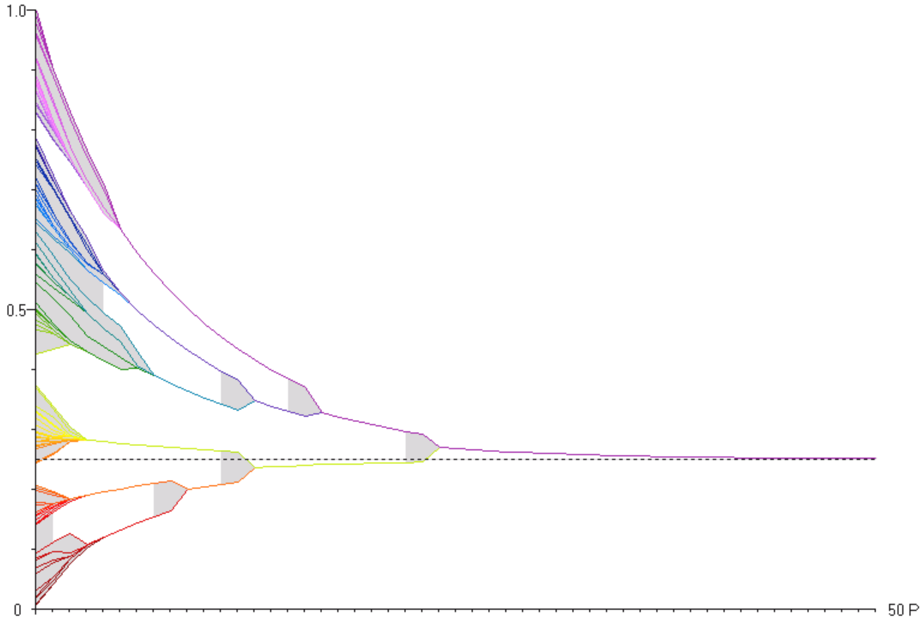


***Figure 2:*** 100 agents, same random start distribution as in *figure 1*, $\varepsilon = 0.05$. $\alpha_i = 0.1$ for all agents *i*, $T = 0.25$.

In *figure 2* we have the same confidence level and exactly the same start distribution as in *figure 1*. The difference is that now *all* agents are *truth seekers* with a homogeneous $\alpha_i = 0.1$ for all agents *i*. So we have a single run example for a dynamics according [2] with $\alpha_i > 0$ and the *BC*–model governing the social exchange process. In *figure 1* the truth did not affect the dynamics at all. By contrast in *figure 2* obviously the whole society is driven direction *T*. Time is the great healer of *$\varepsilon$–splits.* They all close (again) in finite time. Thus, agents in different sub–profiles whose members mutually found their views – in a fairly literal sense – as 'out of range' come closer and start to influence each other again. After times of *plurality* and *polarization* in period 24 a *consensus* is reached. It is *not yet* a consensus on the truth *T*. But *figure 2* suggests that the consensus is moving direction *T,* though the society may not get there in finite time. Thus we can state:

**Observation 1:**

There are conditions under which a society of opinions exchanging truth seekers ends up with a consensus at least fairly close to the truth (*see figure 2*).

Has everybody to be a truth seeker to get a consensus on the truth (or at least on a view close to the truth)? As *figures 3* and *4* demonstrate the answer to that question is: "No!". In *figure 3*

[14] For more details cf. Hegselmann / Krause (2002), chapter 4.

only half of the agents are truth seekers with an $\alpha > 0.1$. *Figure 4* indicates by the colour blue who the truth seekers in *figure 3* are. The trajectories for agents with $\alpha = 0$ are red. Since *not all* agents have a positive $\alpha$ we are facing a situation of a *cognitive division of labour* as defined above.

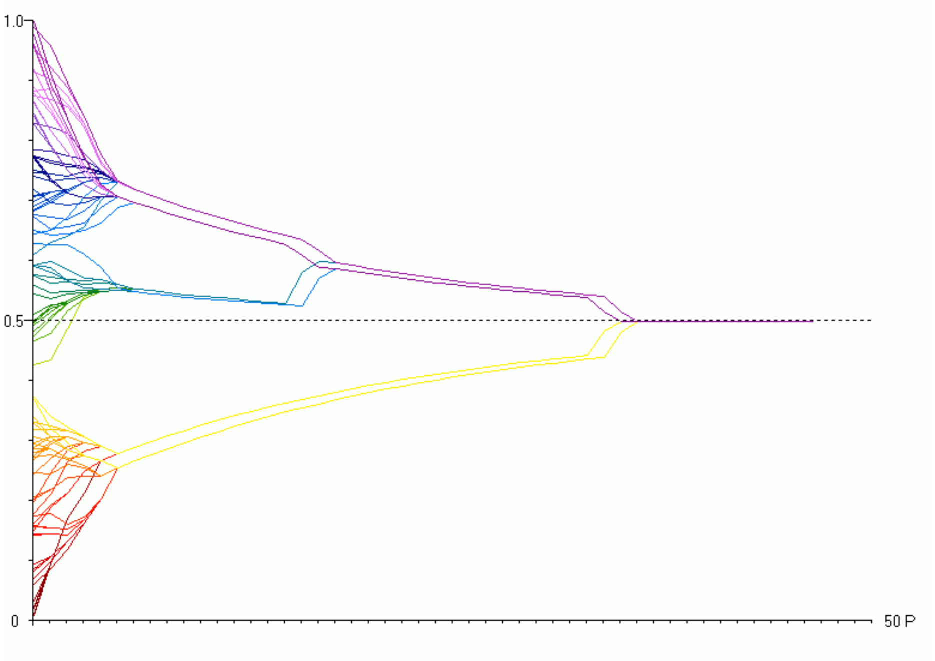


***Figure 3:*** 100 agents, 50% $\alpha = 0.1$, all others $\alpha = 0$, $\varepsilon = 0.1$, $T = 0.5$.

**<u>Observation 2:</u>**
The interplay of both, seeking for the truth by *only some* – i.e. cognitive division of labour – *and* a social exchange process that includes *all,* may finally lead to a consensus at least fairly close to the truth (see *figures 3* and *4*) .

In *figure 5* we have exactly the *same* start distribution as in *figures 3* and *4*. What differs is the *position* of the truth. With $T = 0.5$ in the dynamics given by the former figures the truth was in the centre of the opinion space. In *figure 5* we have $T = 0.05$, i.e. the truth is *extreme*. And that matters. It is still a huge majority of agents, all truth seekers and even almost all with an $\alpha = 0$ that move direction truth. Nevertheless some *non* truth seekers are left behind far distant from the truth.

**<u>Observation 3:</u>**
For the feasibility of the consensus on the truth in observation 2 the position of the truth matters (see *figure 5*). The consensus may become impossible if the truth is extreme.

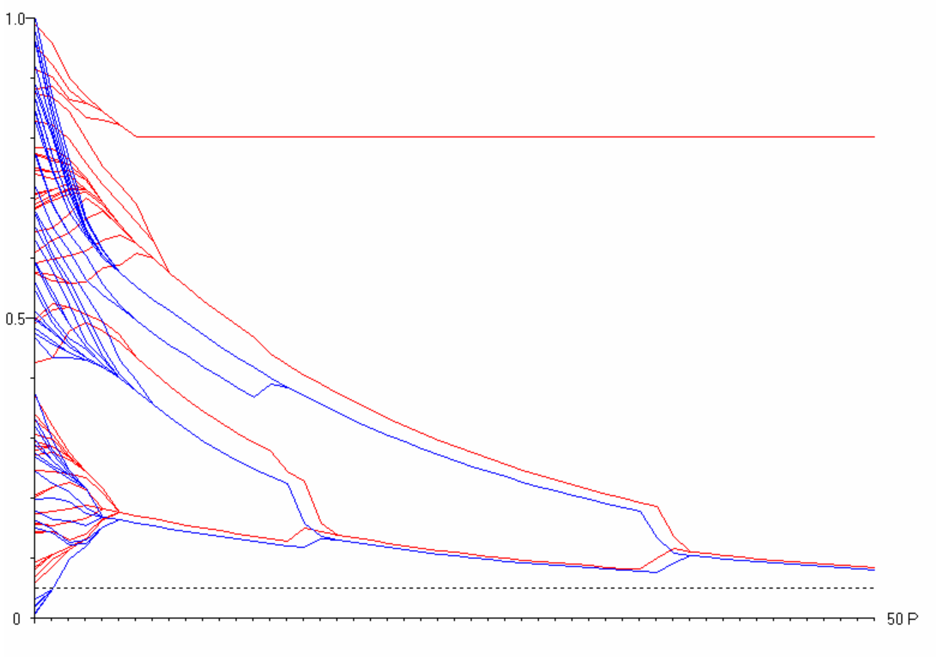


***Figure 4:*** 100 agents, 50% $\alpha$ = 0.1 (blue), all others (red) $\alpha$ = 0, $\varepsilon$ = 0.1, $T$ = 0.5.

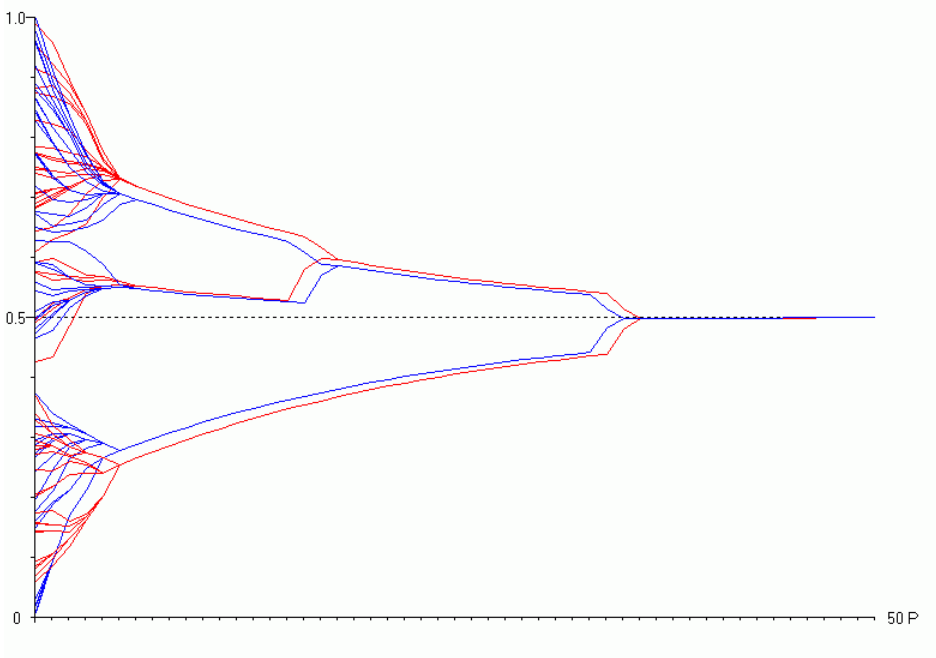


***Figure 5:*** 100 agents, 50% $\alpha$ = 0.1 (blue), all others (red) $\alpha$ = 0, $\varepsilon$ = 0.1, $T$ = 0.05.

In *figure 6* we have again exactly the *same* start distribution and initial conditions as in *figures 3* and *4*. What differs is the exact value of $\alpha$ for the 50% *$\alpha$– positives.* Now we have $\alpha$ = 0.25. As a consequence the *$\alpha$–positives* approach the truth much faster than before. The consequence is that *non*–truth–seekers with start positions more to the extremes of the opinion space are left behind and finally stick to opinions far distant from the truth.

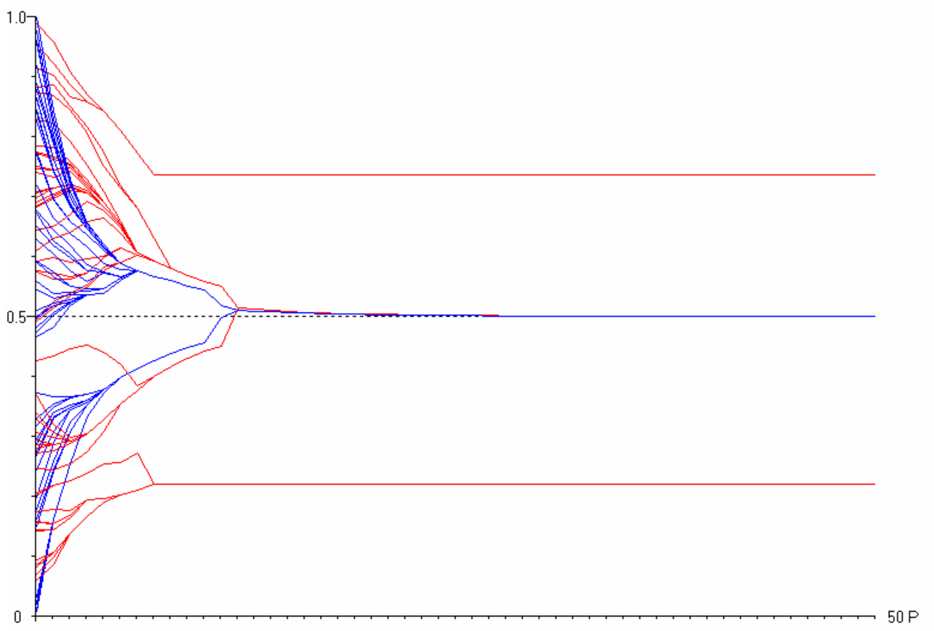


***Figure 6:*** 100 agents, 50% $\alpha = 0.25$, all others $\alpha = 0$, $\varepsilon = 0.1$, $T = 0.5$.

**Observation 4:**

For the feasibility of the consensus in observation 2 the value of $\alpha$ matters. An all including consensus on the truth may become *impossible* if the *$\alpha$–positives* are especially *fast and good* in getting closer to the truth (see *figure 6*).

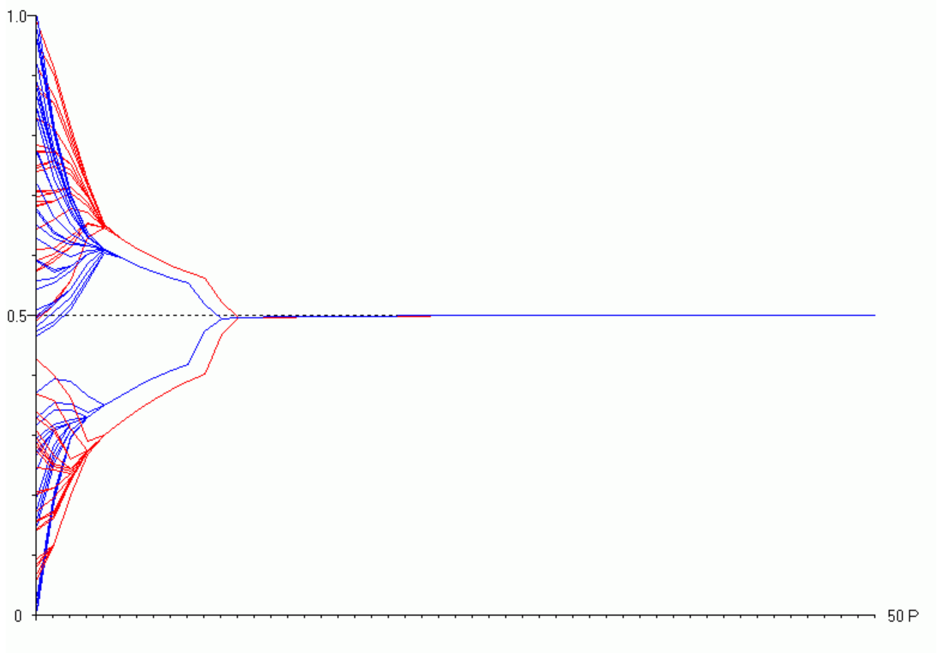


***Figure 7:*** 100 agents, 50% $\alpha = 0.25$, all others $\alpha = 0$, $\varepsilon = 0.15$, $T = 0.5$.

The effect stated by observation 4 may change again if, as *figure 7* indicates, the confidence level is sufficiently high. In *figure 7* start distribution and initial conditions are the same as in *figure 6* except that for the confidence level we now have $\varepsilon = 0.15$. Under that condition the all including consensus on the truth is possible again.

**Observation 5:**
If a consensus on the truth is not possible for a certain value of $\alpha$ then consensus may nevertheless be possible if the *confidence level is sufficiently high* (see *figure 7*).

But again, the consensus on the truth vanishes if there are not enough *$\alpha$–positives.* In *figure 8* we have the same initial conditions as in *figure 7* except that the percentage of *$\alpha$– positives* is reduced to 10%. As a consequence *non*–truth–seekers with start positions in the lower part of the opinion space end up with opinions far distant from the truth. (Though not visible in *figure 8* all opinions in the upper part end up at the truth.)

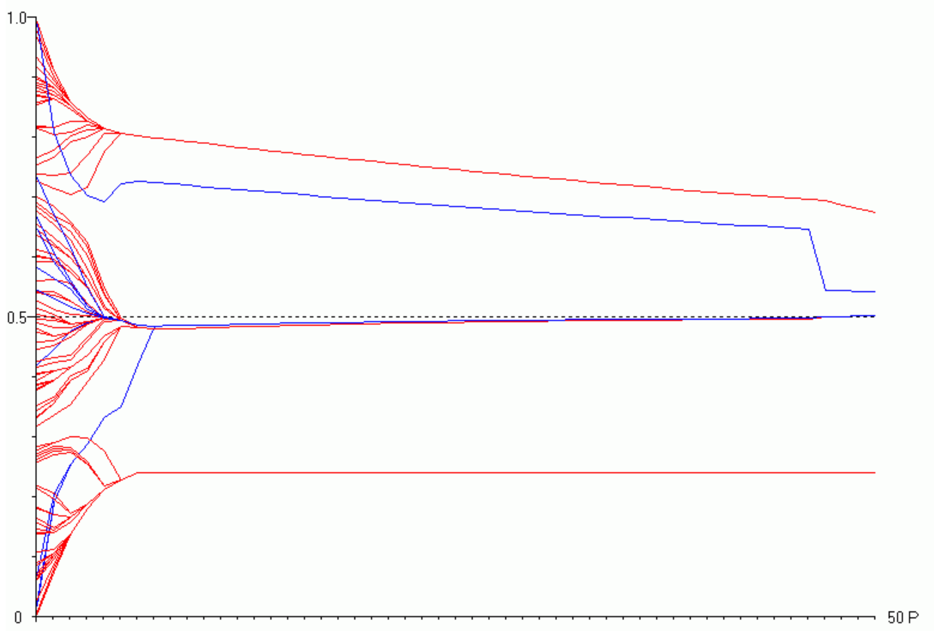


***Figure 8:*** 100 agents, 10% $\alpha$ $\alpha$ = 0.25, all others $\alpha$ = 0, $\varepsilon$ = 0.15, $T$ = 0.5

**Observation 6:**
A consensus on the truth which is reachable based on a certain percentage of *$\alpha$–positives* may no longer be reachable if their frequency drops down. Thus, the frequency of truth seekers matters (see *figure 8*).

Our single run simulations seem to suggest two general ideas: *First*, if all agents in a social exchange process governed by the *BC*–model are truth seekers, then they end up with a consensus on the truth. *Second*, whether under cognitive division of labour, i.e. with not all being $\alpha$–positives, finally a consensus on the truth can be expected, depends crucially on the frequency of $\alpha$–positives, the position of the truth, the confidence level $\varepsilon$ and the value of $\alpha$, i.e. the strength of the truth attraction. In what follows both ideas are analyzed in detail.

## 3. Truth Seeking for Range Preserving Processes: The *Funnel Theorem*

We begin the investigation of truth seeking behaviour as described in the Introduction by equation [2] with quite a general form of the social process $f$. Denote by $K$ the set of all possible profiles $x = (x_1, x_2, ..., x_n)$ with $x_1 > 0, ..., x_n > 0$. Then the social process $f = (f_1, ..., f_n)$ with social components $f_i$ defines a selfmapping of $K$ by $f(x) = (f_1(x), ..., f_n(x))$. The only requirement on $f$ in this section will be that for the set $I = \{1, 2, ..., n\}$ of all agents and for all $i \in I$ and all $x \in K$ it holds that

[3] $$\min\{x_j \mid j \in I\} \leq f_i(x) \leq \max\{x_j \mid j \in I\}.$$

This assumption means that all agents will perform an opinion which remains in the range of opinions of the previous period which is given by the minimum and maximum of opinions in that period. We will refer to an $f$ with this property as a *range preserving social process*. The social process for the *LW*– as well as for the *BC*–model is of this latter kind. The social process for the *LW*–model is given by

[4] $f_i(x) = w_{i1} x_1 + w_{i2} x_2 + ... + w_{in} x_n$ ,

where the weights $w_{ij}$ are fixed values between 0 and 1. For the *BC*–model with confidence level $\varepsilon$ and confidence sets $I(i, x) = \{1 \leq j \leq n \mid |x_i - x_j| \leq \varepsilon\}$ the weights in [4] are allowed to vary according to the rule

[5] $w_{ij} = |I(i, x)|^{-1}$ for $j \in I(i, x)$ and $w_{ij} = 0$ for $j \notin I(i, x)$.

Since the weights in [4] sum up to 1 it follows immediately that the corresponding social process is range preserving. More general, if instead of the weighted arithmetic mean in [4] we would have another mean like the geometric mean, the harmonic mean, or any power mean the same argument as above shows the corresponding social process to be range preserving. (For the social processes driven by various means see Hegselmann/Krause 2004. There a mapping $f$ which satisfies property [3] is called an *abstract mean.*)

Thus we shall consider truth seeking according to the Introduction, equation [2], that is

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

where $f = (f_1, ..., f_n)$ is a range preserving social process. The question whether *agent i approaches the truth T*, that is $\lim_{t\to\infty} x_i(t) = T$, obviously depends heavily on whether $\alpha_i > 0$ or $\alpha_i = 0$. If $\alpha_i = 1$ then agent $i$ will find the truth in the next time step by himself without exchanging opinions with others. If, on the other extreme, $\alpha_i = 0$ then the opinion of $i$ depends solely on the exchange with others. In between, for $0 < \alpha_i < 1$, the opinion formation of $i$ is guided by the truth as well as by exchanging ideas with others. Strictly speaking, approaching the truth depends also on the original opinion profile $(x_1(0), x_2(0), ..., x_n(0))$. For example, if $x_i(0) = T$ for all agents $i$ then all agents approach the truth in a trivial manner without any further assumption. In what follows, approaching the truth is throughout understood to hold for *every* original opinion profile. The following theorem describes how opinions come closer together within a certain 'funnel' and approach finally the truth (cf. Hegselmann 2004b). Though this result is rather general, its proof is simple and short enough to be presented here.

**Theorem 1 (*Funnel theorem*):**
If for truth seeking as modelled by equation [6] the social process is range preserving and all agents go for the truth, that is $\alpha_i > 0$ for all $1 \leq i \leq n$, then *consensus on the truth* holds in the sense that all agents approach the truth.

**Proof:**
Consider the 'funnel' with lower and upper 'walls' $u(t)$ and $v(t)$, both defined inductively by

$$u(t+1) = \min_{1\leq i\leq n}(\alpha_i T + (1-\alpha_i) u(t), u(0)) = \min_{1\leq i\leq n} x_i(0)$$

[7] and

$$v(t+1) = \max_{1\leq i\leq n}(\alpha_i T + (1-\alpha_i) v(t), v(0)) = \max_{1\leq i\leq n} x_i(0).$$

Obviously, $u(0) \leq x_i(0) \leq v(0)$ for all $i$. If $u(t) \leq x_i(t) \leq v(t)$ holds for some $t$ and all $i$ then from equations [3] and [6] we obtain for all $i$ that

$$\alpha_i T + (1-\alpha_i) u(t) \leq x_i(t+1) \leq \alpha_i T + (1-\alpha_i) v(t).$$

Taking the definitions [7] into account we obtain

$$u(t+1) \le x_i(t+1) \le v(t+1) \text{ for all } i.$$

Thus, by the induction principle we conclude that

$$[8] \quad u(t) \le x_i(t) \le v(t) \text{ holds for all } t \text{ and all } i$$

that is, the opinions of all agents are always caught in the funnel as given by [7]. Furthermore, let $\underline{\alpha}$ be the smallest of the values $\alpha_i$ and let $\overline{\alpha}$ be the biggest of the $\alpha_i$. By induction again the funnel can be described as follows, according to the three possibilities of how the truth value relates to the agents initial positions.

**Case 1:** $T \le u(0) \le v(0)$.
Then $T + (1 = \overline{\alpha})^t (u(0) - T) = u(t) \le v(t) = T + (1 - \underline{\alpha})^t (v(0) - T)$.

**Case 2:** $u(0) \le T \le v(0)$.
Then $T + (1 - \underline{\alpha})^t (u(0) - T) = u(t) \le v(t) = T + (1 - \underline{\alpha})^t (v(0) - T)$.

**Case 3:** $u(0) \le v(0) \le T$ (see *figure 9* below)
Then $T + (1 - \underline{\alpha})^t (u(0) - T) = u(t) \le v(t) = T + (1 - \overline{\alpha})^t (v(0) - T)$.

Now, by assumption $0 < \underline{\alpha}$ and, hence, the lower wall $u(t)$ as well as the upper wall $v(t)$ of the funnel converge in all three cases to the truth value *T*. A fortiori, the agents' opinions, which are caught within the funnel, must all converge to *T*. □

Case 3 of the three possibilities above is illustrated below by means of simulations. There, the dotted line depicts the value of *T* and the funnel with walls $u(t)$ and $v(t)$ is in black. The coloured curves within the funnel show the agents' opinions converging to a consensus which approaches *T*.

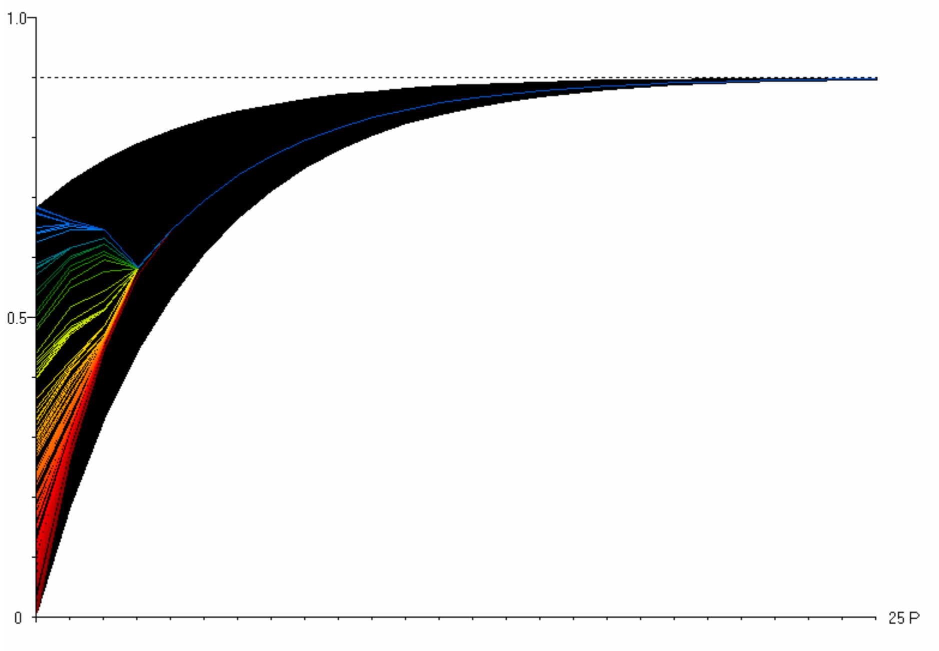


***Figure 9:*** The *Funnel theorem* at work: consensus on the truth.

From the Funnel theorem we get immediately the following consequences.

**Corollary 1:**
Under the assumptions of the *Funnel theorem* with

$$\underline{\alpha} = \min\{\alpha_i \mid 1 \le i \le n\} \text{ and } c = \max\{|v(0)-T|, v(0)-u(0), |u(0)-T|\}$$

the following statements hold:

(a) The distance between the opinions of any two agents is at most △ for all periods after

$$t^*(\Delta) = \frac{\log \frac{\Delta}{c}}{\log(1-\underline{\alpha})}.$$

(b) For the *BC*–model with confidence level $\varepsilon$ and all agents going for truth one obtains consensus on the truth for the time period after $t^*(\varepsilon)+1$.

**Proof:**
(a) For a time period $t \ge t^*(\Delta)$ we have that

$$t\log(1-\underline{\alpha}) \le \log\frac{\Delta}{c}$$

Note that $\log(1-\underline{\alpha})$ is negative, and hence $(1-\underline{\alpha})^t \cdot c \le \Delta.$. From equation [8] and by inspecting the three cases in the proof of the Funnel theorem we obtain that

$$|x_i(t)-x_j(t)| \le v(t)-u(t) \le (1-\underline{\alpha})^t \max\{|v(0)-T|, v(0)-u(0), |T-u(0)|\}.$$

Thus, by definition of $c$, we arrive at $|x_i(t)-x_j(t)| \le \Delta$ for $t \ge t^*(\Delta)$ and all $i$ and $j$.

(b) The assertion follows from (a) for $\Delta = \varepsilon$, because in the *BC*–model a distance of at most $\varepsilon$ between the opinions of any two agents implies consensus in the next time step. □

The effect stated in Corollary 1 is illustrated by *figure 10*.

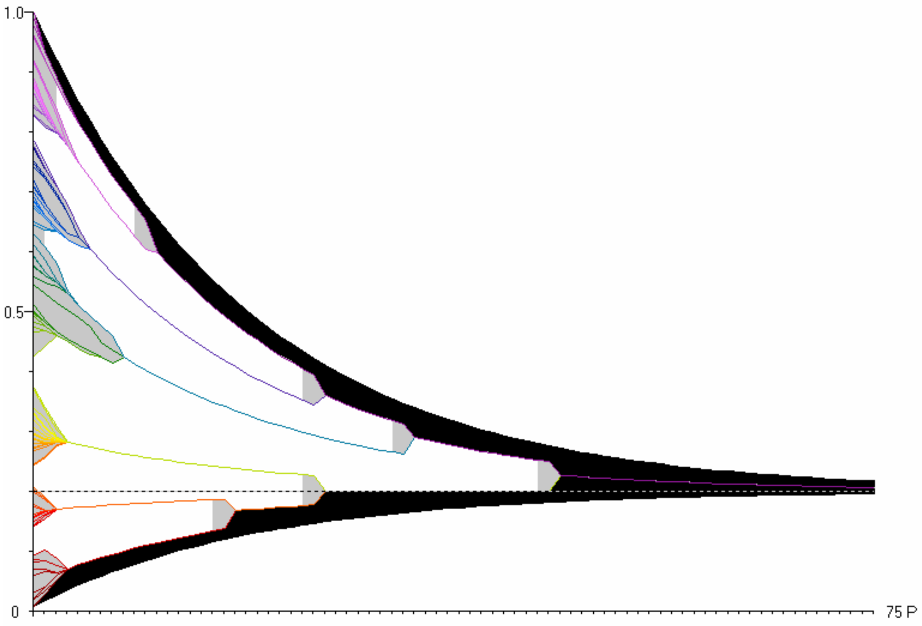


***Figure 10:*** Time as the great healer: All $\varepsilon$–splits close again in finite time.

In *figure 10* the funnel is marked black as in *figure 9*. *Figure 10* shows in addition *grey* areas between two neighbouring opinions if and only if the distance between the two is not greater than $\varepsilon$. Obviously all splits in the opinion profile close again. It is a remarkable feature of the *Funnel theorem* and in particular of the above *Corollary* that agents whose opinions would stay apart forever if there were the social process dynamics only, will come arbitrarily close together if they are $\alpha$–positives, i.e., if they are going for the truth.

**Remark:** The Funnel theorem as well as the Corollary 1 can be extended with essentially the same proofs to a case where not all agents go for the truth. $I$ is now restricted to the set of truth seeking agents, i.e. $I = \{i \in \{1, \ldots n\} | \alpha_i > 0\}$. Suppose $I$ is not empty and that property [3] holds with respect to *this subset I* of $\{1, \ldots, n\}$. Then all agents in the subset $I$ will approach the truth.

## 4. Truth Seeking for Time Variant Processes: The *Leading the Pack Theorem*

The *Funnel Theorem* of the previous section demonstrates that if *all* agents go for truth they will reach a consensus arbitrarily close to the truth. What, however, about the others if only a few agents go for truth? *Figures 3, 4,* and *7* give single run examples in which only half of all the agents are α–positives. Nevertheless the whole society ends up at the truth. In this chapter we will present analytical results which show that even *one* single α–positive agent can lead all others to the truth, provided the latter possess 'enough confidence'. In contrast to the Funnel theorem the results do not hold for range preserving social processes in general, but for certain *time variant social processes* only, which nevertheless still include the *LW–* as well as the *BC*–model. Thus, we will consider a model for truth seeking of the type [6] where the social process is of the form

[9] $$f_i(x(t)) = \sum_{j=1}^{n} w_{ij}(t) x_j(t)$$

where the nonnegative weights $w_{ij}(t)$ add up to 1 for fixed $i$ but are now admitted to vary with time $t$ in such a way that for any two agents $i, j$ and $t \geq 0$ either $w_{ij}(t) = 0$ or $w_{ij}(t) \geq \underline{w}$ for some fixed weight $\underline{w} > 0$. Obviously, the *LW*–model with time–independent weights is of this kind by taking $\underline{w}$ to be the smallest of all strictly positive weights. For the *BC*–model as defined by equations [4] and [5] set

[10] $w_{ij}(t) = |I(i, x(t))|^{-1}$ for $j \in I(i, x(t))$ and $w_{ij}(t) = 0$ otherwise.

Hence for $w_{ij}(t) > 0$ we have that $w_{ij}(t) \geq 1/n$ because of $|I(i, x(t))| \leq n$. Thus, the *BC*–model is of the above kind with $\underline{w} = 1/n > 0$.

Beside the *LW*–model and the *BC*–model there are other interesting examples of time variant social processes. We leave their discussion to another paper.

Before we can state our result on how a few agents may lead the whole pack to consensus–on–the–truth we need the following important concept (which generalizes the corresponding one in Hegselmann/Krause 2002, Appendix D). We say there is a *confidence chain from agent i to agent j for the time interval (s, t)* with $s < t$ if there exists a chain of agents $i_0, i_1, \ldots, i_k$ with $i_0 = i$ and $i_k = j$ such that the weights

$$w_{i_0 i_1}(t-1), w_{i_1 i_2}(t-2), \cdots, w_{i_{k-1}, i_k}(s)$$

are all strictly positive.

Additionally, we have to remind the definition of an irreducible matrix. A nonnegative $n \times n$-matrix $A = (a_{ij})$ is called *irreducible* if for any two indices $i,j \in \{1, ..., n\}$ there exists a sequence of indices such that $a_{ii_1} \cdot a_{i_1 i_2} \cdots a_{i_{k-1},j} > 0$.

Further, we call a path $t \rightarrow (x_1(t), ..., x_n(t))$ an *ε-profile for (i, j) and (s, t)* if there exists a sequence of indices such that

$$|x_i(t-1) - x_{i_1}(t-1)| \leq \varepsilon,\ |x_{i_1}(t-2) - x_{i_2}(t-2)| \leq \varepsilon, \cdots, |x_{i_{k-1}}(s) - x_j(s)| \leq \varepsilon.$$

Now we can state our second analytical result.

**Theorem 2 (*Leading the pack theorem*):**
Let a truth seeking model with a time variant social process given by

[11] $$x_i(t+1) = \alpha_i T + (1-\alpha_i) \sum_{j=1}^{n} w_{ij}(t) x_j(t),\ 1 \leq i \leq n$$

with weights $w_{ij}(t)$ such that either $w_{ij}(t) = 0$ or $w_{ij}(t) \geq w$ for some $w > 0$. Suppose there exists a sequence of time periods $t_m$ with $2 \leq t_m - t_{m-1} \leq r$ for $m = 1, 2, 3, ...$ such that for each agent not going for the truth and each $m$ there exists a confidence chain to an agent going for the truth for the time interval $(t_{m-1} + 1, t_m)$. *Then* for the truth seeking model given by [11] there holds consensus–on–the–truth, i.e. for arbitrarily given initial opinions all agents will approach the truth.

**Proof:**
See the Appendix.

**Corollary 2:**
Consensus–on–the–truth holds
(a) for the *LW*–model if the matrix *W* of weights is irreducible and at least one agent goes for the truth,
(b) for the *BC*–model if there exist time periods $t_m$ as in *Theorem 2* such that for each agent *i* not going for the truth there exists an agent *j* going for the truth such that paths are ε–profiles for (*i, j*) and $(t_{m-1} + 1, t_m)$.

**Proof:**
The statements follow easily from *Theorem 2* by taking into account the given definitions. □

Note that for the results above it is required that at least one agent goes for the truth. If this is not required, the model reduces to the ordinary opinion dynamics for which it is well known that irreducibility is not sufficient for a consensus. An analogue of part (a) of *Corollary 2* has been obtained within a different framework in Friedkin and Johnsen (1990). Notice, however, that according to *Theorem 2* in part (a) a weakened form of irreducibility would be sufficient.

The following example demonstrates that, on the other hand, part (a) may fail for a matrix that is not irreducible.

**Example:** Consider the following LW–model for $n = 3$ with weight matrix

$$W = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & \frac{3}{4} & \frac{1}{4} \end{bmatrix}.$$

This matrix is not irreducible. Consider the truth seeking model with $T > 0$ and $0 < \alpha_1 < 1$ arbitrary and $\alpha_2 = \alpha_3 = 0$. One finds that $\lim_{t \to \infty} x_2(t) = \lim_{t \to \infty} x_3(t) = c$ with

$$c = \frac{x_2(0) + x_3(0)}{2} \text{ and } \lim_{t \to \infty} x_1(t) = \frac{2\alpha_1 T + (1 - \alpha_1)c}{1 + \alpha_1}.$$

Thus, there will be a stable configuration for $t \to \infty$ with a consensus among agents 2 and 3 who do not go for truth. There will be a consensus–on–the–truth if and only if $T = c$. Thus, for most of the original profiles there is no consensus–on–the–truth – though agent 1 is going for truth. Moreover, even the single agent 1 does not approach the truth. This is *not* in contradiction to the above remark concerning the Funnel Theorem because the social process defined by matrix $W$ is not range preserving with respect to $I = \{1\}$.

Now, let the social process still be given by the not irreducible matrix $W$ as above but change the truth seeking model a little bit by assuming that $0 < \alpha_2 < 1$ whereas $\alpha_3 = 0$. Now agent 3 is the only one not going for the truth and there is a chain of confidence to agent 2. Therefore, by *theorem 2* there must be consensus–on–the–truth. It is amazing and the reader is invited to check directly for this example that $\lim_{t\to\infty} x_1(t) = \lim_{t\to\infty} x_2(t) = \lim_{t\to\infty} x_3(t) = T$ – irrespective of the particular values of $\alpha_1$ and $\alpha_2$, as long as they are positive. This result cannot be concluded neither from part (a) of Corollary 2 not from the Funnel Theorem. To apply the latter one would need that the social process is range preserving with respect to {1, 2, 3} which is not the case.

We illustrate part (b) of *Corollary 2* by *Figure 11a* and *11b. Figures 11 a* and *b* both simulate the same dynamics. But whereas *Figure 11a* shows only 20 periods, *Figure 11b* shows 300 periods. A blue path represents an agent going for truth, a red path an agent not going for truth. Though there are only a few 'blue agents' one can find in *Figure 1* for every 'red agent' an $\varepsilon$-profile with a blue agent on an appropriate time interval. Actually, this is possible in many different ways and the reader is invited to find some on his own. For example, one could even choose time intervals starting with 0 as well as profiles involving red agents from 'different groups'. Another example would be to start in *Figure 11a* with period 8. At any period after that each red agent has even a distance of at most $\varepsilon$ to a blue agent. This holds in *Figure 11a* but it does not hold for ever as is clear from *Figure 11b*. There is a split in period 90 between the lowest group of red agents and a group of blue agents because the latter unifies with another group of blue agents due to attraction by the truth. This split will not be 'healed' – in contrast to another split which is 'healed' at about period 160. According to the *Leading the pack theorem* the question of a consensus on the truth (for all agents) depends only on whether finally red agents will be connected to blue ones. In *Figure 11b* the lowest red group will never connect with a blue agent. As a consequence there will be no consensus on the truth that includes all agents.

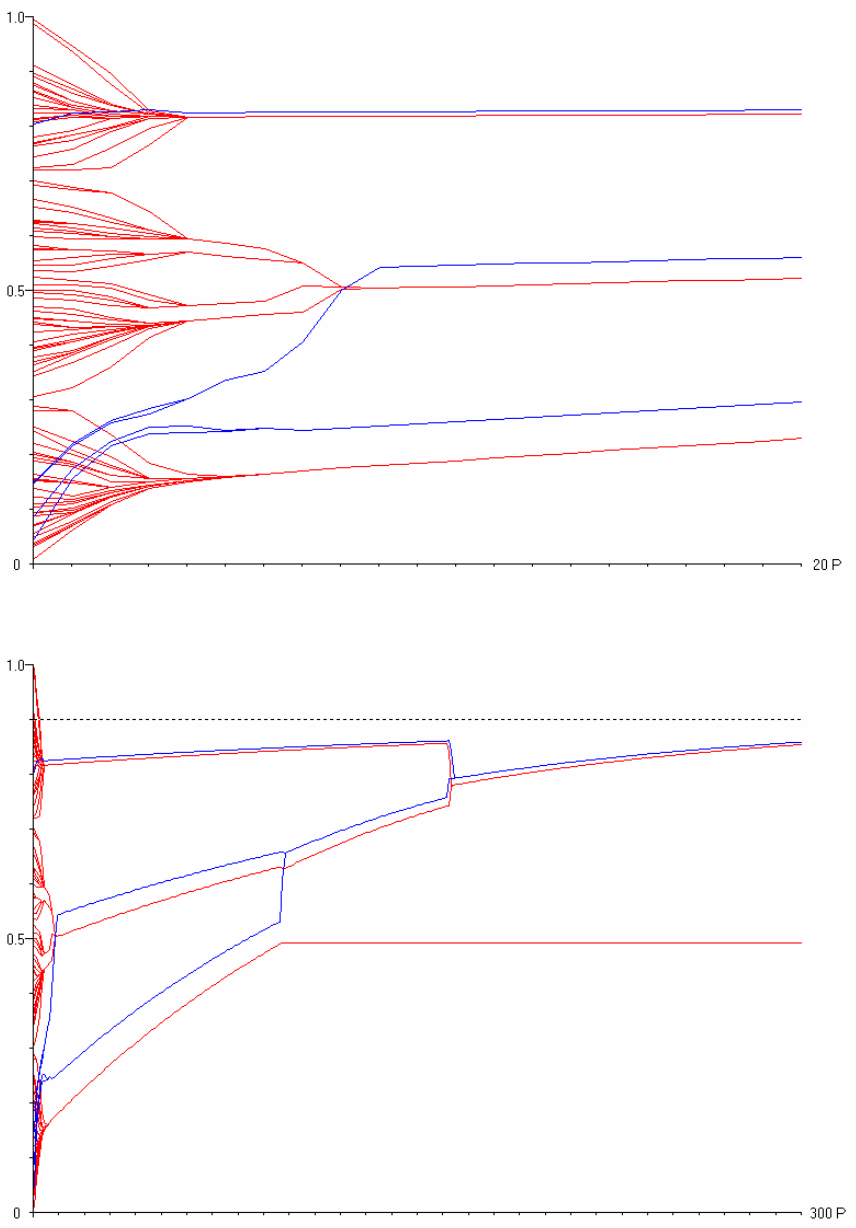


***Figure 11a and b:*** Same dynamics after *(a)* 20 periods and *(b)* 300 periods. Red: α=0; blue: α = 0.1. Confidence level for all *ε* = 0.1.

## 5. *CASE*–Studies II: Truth Deviation in the Parameter Space <*T*, *F*, *ε*, *α* >

The *Leading the pack Theorem* implies for the *BC*–model that under certain conditions few or even one truth seeker may lead the pack to the truth. But when can we expect those conditions to be fulfilled? And where do the dynamics tend to end if such conditions are not fulfilled? By the *CASE*–studies in chapter 2 we know already that the chances of a final consensus on the truth depend crucially on the frequency of *α*–positives, the position of the truth, the confidence level *ε* and the strength *α* of the truth attraction. Thus, we are facing a 4–dimensional parameter space. In this chapter we will analyze that parameter space by systematic simulations. The guiding research question will be: *What average truth deviation do we have to expect in the parameter space after the dynamics has stabilized?* Truth deviation could be measured in different ways. We will do it in a way that closely follows the definition of the *standard deviation*. We simply substitute the mean by *T*. Thus *truth deviation* $\delta(t)$ is defined as follows:

[12] $$\delta(t) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(T - x_i(t)\right)^2}$$

Analysing a 4–dimensional parameter space is hard enough a task. Therefore all other things are kept as simple as possible: We assume, first, *homogeneous, symmetric*, and *constant* confidence levels $\varepsilon$ for all individuals, second, a *homogeneous and constant* $\alpha$ for all $\alpha$–positives, third, a *fixed number* of 100 individuals, fourth, a *uniform* start distribution, and, finally, *simultaneous* updating.

The simulation results on average truth deviation in 4–dimensional parameter space will be *visualized* to allow for a fast and easy understanding of our *CASE*–studies. More in detail, we will vary the parameters as follows:

1. The position of the truth *T*: 0.1, 0.3, 0.5.
2. The frequency *F* of agents with $\alpha = 0$: 10%, 50%, 90%.
3. The confidence level: $\varepsilon = 0.01, 0.02, \ldots, 0.4$; i.e. 40 steps.
4. The strength of the truth directedness: $\alpha = 0.01, 0.02, \ldots, 1.0$; i.e. 100 steps.

As to *T* and *F* the simulation strategy is a *scenario* approach. The truth may be in the centre of the opinion space, the truth may be an extreme or lies somewhere in between. Almost all, half or only some of the agents may be *non* truth seekers. Together that is a total of 3×3 scenarios. As to $\varepsilon$ and $\alpha$ we follow a *grid* approach with 40×100 points. *Figure 12* explains the combined scenario and grid approach.

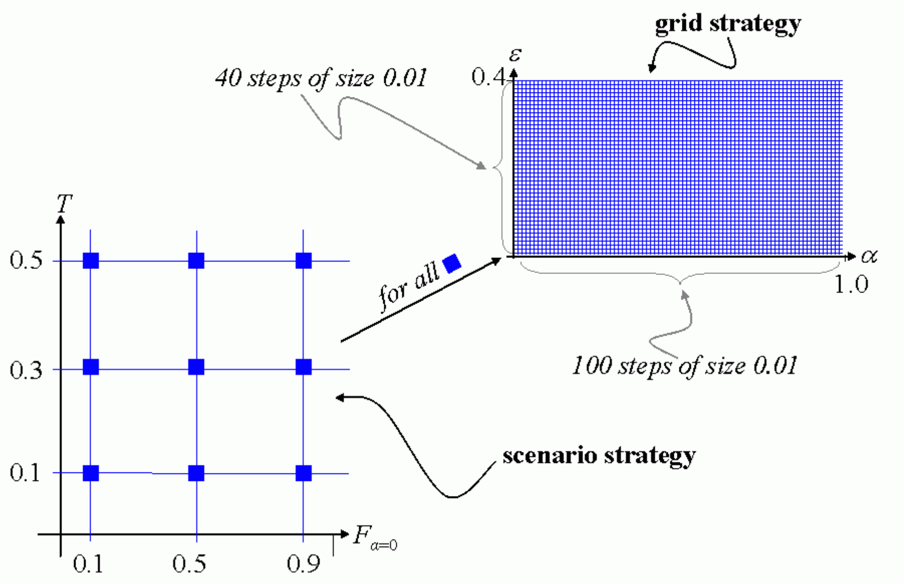


***Figure 12*: The combined grid and scenario approach.**

For each of the 40×100 grid points of the 3×3 scenarios we run 50 simulations until the dynamics is stable. Then we calculate the average truth deviation and encode the numerical value by a colour. As a consequence all simulation results can be presented in the one huge graphics given by *figure 13*.

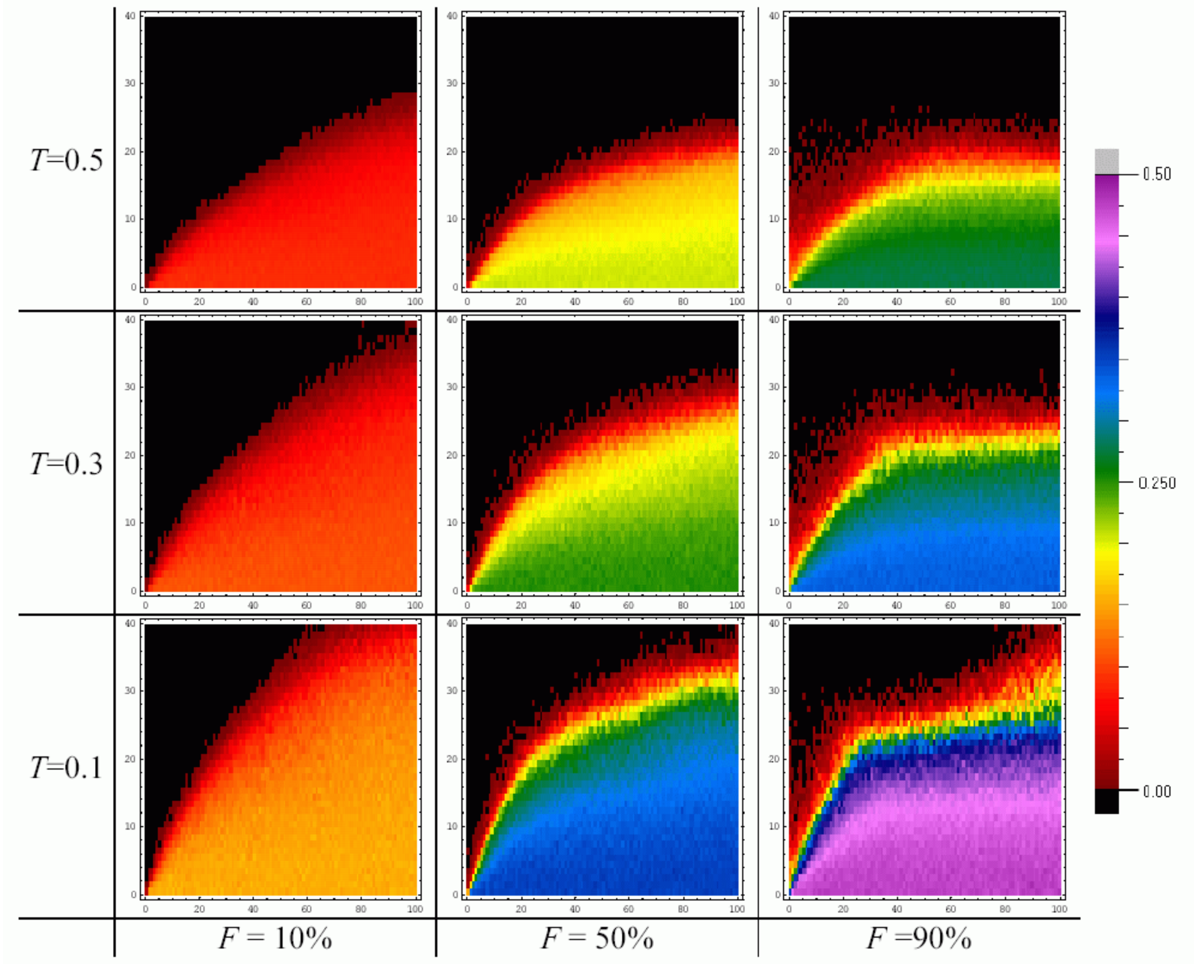


***Figure 13:*** Truth deviation in the parameter space <*T*, *F*, $\varepsilon$, $\alpha$ >.

Most striking in *figure 13* are the huge black areas in all scenarios. In these regions of the parameter space the average truth deviation is 0. Thus the agents finally ended up with a consensus on the truth.

**Observation 7:**
Where ever the truth is located, whether there is a tiny minority or even an overwhelming majority of truth seekers, under *suitable values* for the confidence level and the strength of the attraction of truth the whole society will end up with a consensus on the truth.

For such a consensus on the truth there have to be the *right proportions* between $\varepsilon$ and $\alpha$ as stated in observation 8 and 9.

**Observation 8:**
For almost all *T*, *F*, and $\alpha$ there seems to be a *critical value* $\varepsilon^*$ for the confidence level such that a consensus on the truth is possible only if $\varepsilon > \varepsilon^*$. The critical confidence level that has to be exceeded *increases* as the strength $\alpha$ of truth attraction *increases*. If the truth is in the centre of the opinion space the critical confidence level is lower than for more extreme truths. With exception of the scenario <*T*=0.1, *F*=90%> it seems that for a given position of the truth the critical value $\varepsilon^*$ slightly decreases if *F* increases.

**Observation 9:**

For almost all *T*, *F*, and $\varepsilon$ there seems to be a *critical value* $\alpha^*$ for the strength of the attraction of truth such that for a consensus on the truth it is necessary that $\alpha < \alpha^*$. If the truth is in the centre and the confidence level sufficiently high, then a consensus will be reached with any $\alpha$. For most parts of the scenarios the critical value $\alpha^*$ that must not be exceeded *decreases* as the confidence level $\varepsilon$ *decreases*. It may be that for *F*=90% and low confidence levels consensus on the truth is only possible for values of $\alpha$ much smaller than those we used in our simulations.

Not all parameter constellations finally end up with a consensus on the truth. As it seems there are conditions in which it is *more* and those in which it is *less* difficult to get at least close to the truth:

**Observation 10:**

If for a certain parameter constellation $<\alpha,\varepsilon>$ there is a positive truth deviation in all scenarios <*T*, *F*>, then these truth deviations tend to be higher for higher frequencies *F* and more extreme truths *T*.

For an interpretation of most of these observations it is necessary to understand *two* effects: The *first* might be called the *take–along–effect*: Even a single truth seeker can drive a whole society to the truth by taking along the others. The effect works for small minorities of $\alpha$–positives as well. More general: Via the social exchange process truth seekers can take along those which do not go for the truth or do not succeed in doing so. But for that to be the case the values for $\alpha$ and $\varepsilon$ have to have the right proportion: Given a confidence level the attraction of truth must not be too strong; given the attraction of truth the confidence level has to be high enough. *Figure 14a* (*top*) gives an example. Here the one and lonely truth seeker takes along all others direction *T*. The *second* effect works in the opposite direction: Truth seekers may be 'too good' to take along the others. *Figure 14b* (*bottom*) gives an example. Here a comparatively small group of extremely good truth seekers crosses the whole opinion space very fast in direction of an extreme truth. Though at the beginning scattered over the whole opinion space they leave behind them a polarized society whose opinions they didn't manage to influence significantly. The effect could be called the *leaving–behind–effect.*

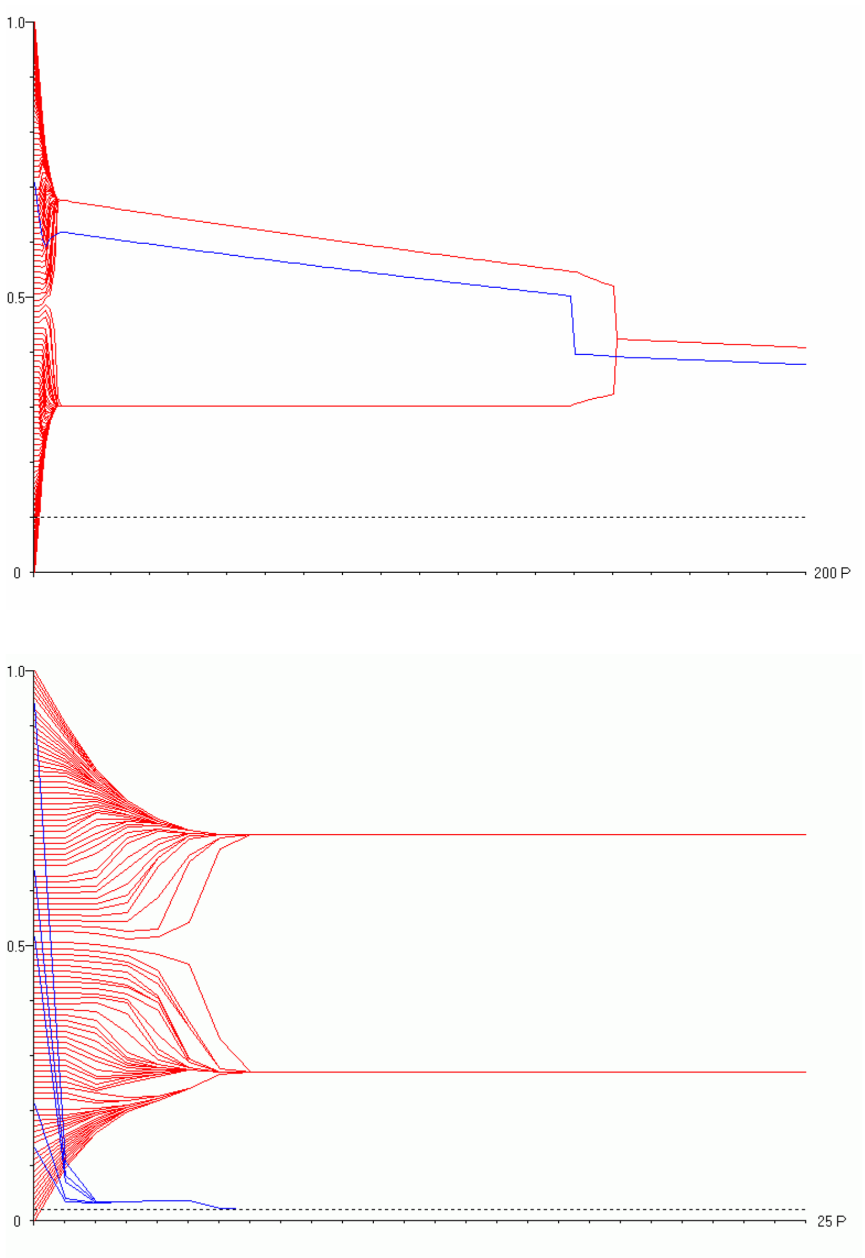
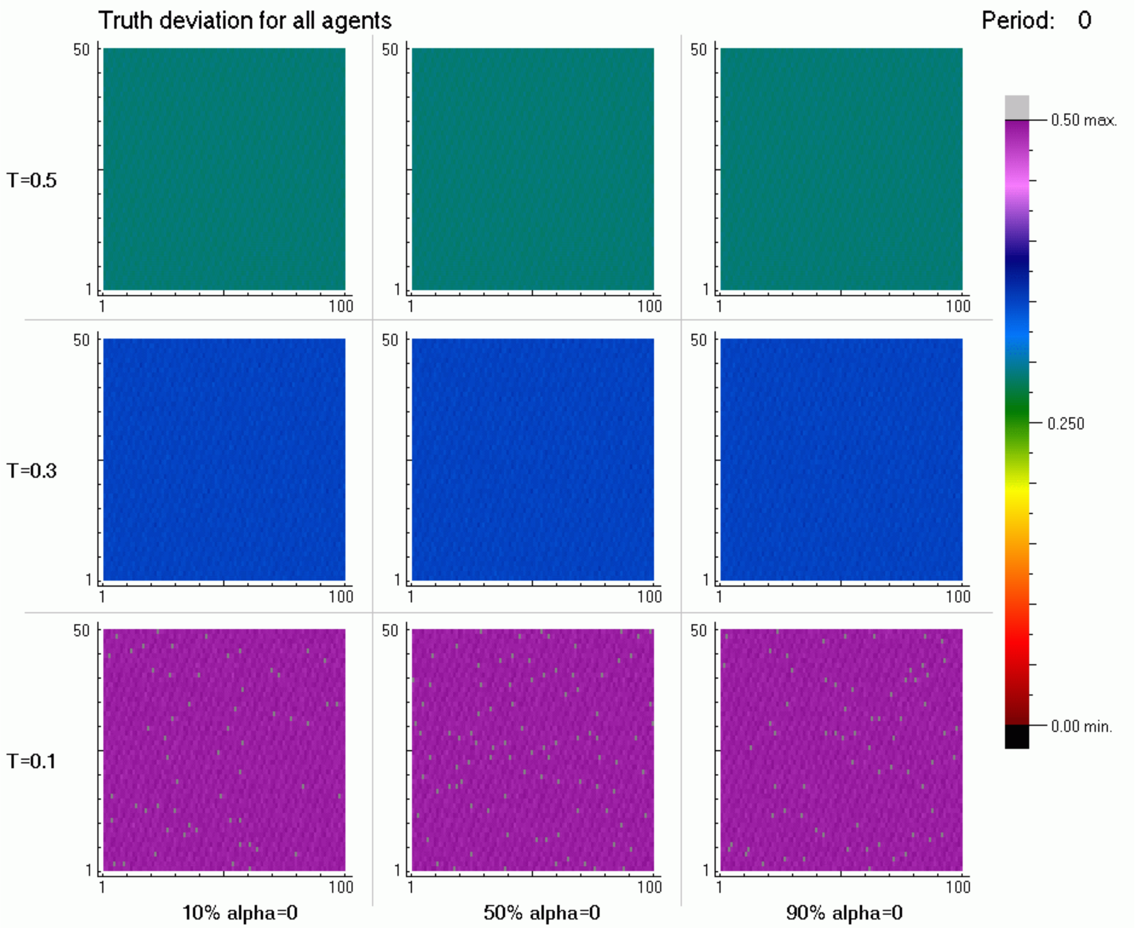


***Figure 14***: (a) *take–along–effect* (T= 0.1, $\alpha = 0.1, \varepsilon = 0.2$). (b) *leaving–behind–effect* (T = 0.02, $\alpha = 0.9, \varepsilon = 0.2$).

*Figure 13* shows the truth deviation *after* the dynamics stabilized. But how does the truth deviation develop over time. A simulation based answer is given by the following *movie*. Each of its 51 pictures has the structure of *figure 13* and shows period wise the truth deviations up to period 50. A difference to *figure 13* is that in the movie the confidence level is analysed in 50 steps up to $\varepsilon = 0.5$.



***Movie:*** Truth deviation up to period 50.

By looking at the movie one can verify at least the following observations:

**Observation 11:**
In all scenarios it is the upper right corner, i.e. the region where high confidence levels are combined with a strong truth attraction, where the most significant *decrease in the truth deviation* starts. The decrease spreads to the left, i.e. in the direction of less strong truth attraction, and to the bottom, i.e. in the direction of lower confidence levels.

**Observation 12:**
The truth deviations goes down especially fast if the truth is in the centre $T = 0.5$ and/or the frequency of agents not interested in the truth is small ($F = 10\%$).

**Observation 13:**
The final structure *after stabilization* as we know it from *figure 13* emerges *very early*. The most significant difference to the final structure can be found for small values of $\alpha$. The difference is more severe if the truth is more extreme and/or the frequency of agents going for the truth is small.

Obviously CASE–studies like the ones above allow studying epistemic constellations under the perspective when and why the spreading of the truth may be more or less difficult.

## 5. Research Perspectives

The most striking result of the analytical and simulation based results presented here is probably that in our model the combined cognitive and social dynamics under lots of circumstances ends up with an all including consensus on the truth. That sounds a bit like assumptions in the wake of the so called consensus theory of truth as it was outlined by Jürgen Habermas and his followers (cf. Habermas 1973). However, one should note that the concept of truth as assumed in our approach, though being compatible with a huge variety of different truth conceptions, is *not compatible* with what is called the consensus theory of truth. We assume the existence of a true value which is totally *independent* of what is believed to be the case, while in the consensus theory of truth a special consensus constitutes the true value. Additionally, in our approach a community may definitely fail to reach a consensus at the truth. The *leaving–behind–effec*t may result in a situation in which the truth seekers end up at the truth but among themselves while all others live (possibly polarized) far distant from the truth somewhere in the opinion space.

The framework we presented here allows the study of social and epistemic processes in which more or less successful truth seekers are at the same time involved in social exchange processes. The framework allows studying the dynamics of *stylized epistemic constellations*. For the research perspective within this framework one should note:

- One can focus on *difficult situation for spreading the truth*, for instance opinion leaders with positions far distant from the truth and not interested in the truth, confidence intervals with a bias against the truth or a peak far distant from the truth in the start distribution.
- The fairly optimistic *idealized assumptions on T could be given up*. For instance, new evidences may point into two or more different directions.
- We took the strength of the attraction of truth as given. Another view could be to look at $\alpha$ as a parameter that can be influenced by intervention. Under such a view one might start thinking on efficient *truth proliferation policies*: Where in the opinion spectrum should one make how much agents to truth seekers in order to make all or at least a significant part of a society believing in the truth? What if there is time pressure? What to do if the social exchange process has a network structure with primarily local interactions?
- In this article the bounded confidence model plays an important role. But the general framework is *not necessarily tied to the BC–model* for the social process. There are other mechanisms as well, for instance mechanisms which are *not* range preserving.

Thus the framework, the methods, and results suggest that there is a realistic perspective for a social epistemology that addresses theoretical and technical question on the spreading of truths.

## APPENDIX: Proof of the *Leading the pack theorem*

To prove Theorem 2, we first reformulate the model given by equation [11], that is

$$x_i(t+1) = \alpha_i T + (1-\alpha_i)\sum_{j=1}^{n} w_{ij}(t)\; x_j(t),\;\; 1 \le i \le n\,,$$

by adding an artificial agent 0 with the dynamics

$x_0(t+1) = x_0(t)$ for all t = 0, 1, 2, ... and $x_0(0) = T$.

Let $x(t) = (x_0(\mathrm{t}), x_1(t), ..., x_n(t))$ (column vector) and define the matrix

[13] $$A(t) = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline \alpha_1 & & & \\ \vdots & & W_\alpha(t) & \\ \alpha_n & & & \end{array}\right]$$

where the elements of the $n \times n$-matrix $W_\alpha(t)$ are given by $(1 - \alpha_i)\, w_{ij}$ (t) for $1 \leq i, j \leq n$.

The models equation [11] can be compactly written as

[14] $x\,(t + 1) = A\,(t)\, x\,(t), \quad x\,(0) = (T, x_1\,(0), ..., x_n\,(0))$

To this discrete dynamical system we will apply a stability theorem from Hegselmann/Krause (2002, Theorem 4) and from Krause (2003, Theorem 1, p. 203) which yields consensus, $\lim_{t\to\infty} x_i\,(t) = c$ for $i = 0, 1, 2, ..., n$, provided the following assumption is satisfied for the entries $b_{ij}(t, s)$ of the matrix $B(t, s) = A\,(t - 1)$ A $(1 - 2)$ ... $A\,(s)$ for s < t: There exist a sequence $t_0 < t_1 < t_2 <$... of time periods and a sequence $\delta_1, \delta_2, \cdots$ in $[0,1]$ with $\sum_{m=1}^{\infty} \delta_m = \infty$ such that for all $m \geq 1$, $1 \leq i, j \leq n$

[15] $$\sum_{k=1}^{n} \min\{b_{ik}(t_m, t_{m-1}),\, b_{jk}(t_m, t_{m-1})\} \geq \delta_m.$$

In the stability theorem as cited above it is assumed that $t_0 = 0$. We may, however, let $t_0$ be arbitrary by choosing $\delta_k = 0$ for finitely many periods.

Before demonstrating this assumption to be satisfied let us check that consensus for model [14] yields consensus–on–the–truth for our truth seeking model. Because of $x_0\,(t) = T$ for all $t$ we have that $c = T$ and, hence, $\lim_{t\to\infty} x_i\,(t) = T$ for $1 \leq i \leq n$.

The main step in proving [15] is the following lemma.

**Lemma:** For $u \in \mathbb{R}_+, v \in \mathbb{R}_+^n$ and $\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}_+^n$ it holds that

[16] $$B(t,s) \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u \\ v' \end{bmatrix} \text{ with}$$

$$v' = u\left(\sum_{k=0}^{t-s-1} W_\alpha(t-1) W_\alpha(t-2) \cdots W_\alpha(t-k)\right)\alpha + W_\alpha(t-1)\; W_\alpha(t-2) \cdots\; W_\alpha(s+1)\; W_\alpha(s)\, v.$$

**Proof:** We prove [16] by induction over $t$ for $s$ fixed. Suppose first $t = s + 1$.

Obviously, $B\,(s+1,s) \begin{bmatrix} u \\ v \end{bmatrix} = A(s) \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u \\ v' \end{bmatrix}$ with $v' = u\,\alpha + W_\alpha\,(s)\,v$ by [11]. Thus, equation [16] holds for $t = s + 1$ (by convention the empty product of matrices is the identity matrix).

Now, by induction hypothesis

$$B\,(t+1,s) \begin{bmatrix} u \\ v \end{bmatrix} = A(t)\, B(t,s) \begin{bmatrix} u \\ v \end{bmatrix} = A(t) \begin{bmatrix} u \\ v' \end{bmatrix} = \begin{bmatrix} u \\ v'' \end{bmatrix}$$

with

$$v'' = u\alpha + W_\alpha(t) v' = u\alpha + u\left(\sum_{k=0}^{t-s-1} W_\alpha(t)\; W_\alpha(t-1) \cdots W_\alpha(t-k)\right)\alpha + W_\alpha(t) W_\alpha(t-1) \cdots W_\alpha(s+1) W_\alpha(s) v.$$

From

$$uW_\alpha(t) \ldots W_\alpha\,(t+1)\;\alpha = u\alpha$$

and

$$\left(\sum_{l=1}^{t-s} W_\alpha(t)W_\alpha(t-1)\cdots W_\alpha(t+1-l)\right) = \sum_{k=0}^{t-s-1} W_\alpha(t)W_\alpha(t-1)\cdots W_\alpha(t-k)$$

we obtain that formula [16] holds for $B\,(t+1, \text{s})$, too. □

The first column of $B\,(t, s)$ is given by $B(t,s)\begin{bmatrix}u\\v\end{bmatrix}$ for $u = 1$ and $v = (0, ..., 0)$. Therefore, by the lemma we obtain for $1 \leq i \leq n$

[17] $b_{i1}(t, s) \geq \sum_{j=1}^{n} m_{ij}(k)\alpha_j$

where $M\,(k) = (m_{ij}\,(k)) = W_\alpha\,(t-1)\; W_\alpha\,(t-2) \ldots W_\alpha\,(t-k)$ for $0 \leq k \leq t-s-1$.

For the sequence $(t_m)$ of time periods assumed in Theorem 2 we shall show that for all $1 \leq i \leq n$

[18] $b_{i1}\,(t_m, t_{m\text{-}1}) \geq \delta_m$ with $\delta_m = \min\{\alpha_i \mid \alpha_i > 0\}\; w^{\,r-1}$.

For, if $1 \leq i \leq n$ with $\alpha_i > 0$ then from [17] for $k = 0$ we have that $b_{i1}\,(t_m, t_{m\text{-}1}) \geq \alpha_i \geq \delta_m$, since $w \leq 1$. If $\alpha_i = 0$ then by the assumption in Theorem 2 there exists a confidence chain $(i_0, i_1, ..., i_k)$ from $i = i_0$ to $i_k = j$ for the interval $(t_{m\text{-}1} + 1, t_m)$ such that $\alpha_j > 0$.Without loss let $\alpha_{i_h} = 0$ for $0 \leq h \leq k-1$. That is, $w_{ii_1}(t_m-1), w_{i_1i_2}(t_m-2), \cdots, w_{i_{k-1},j}(t_{m-1}+1)$ are all strictly positive and, hence, greater or equal to $w$. From [17] we obtain for $t = t_m$, $s = t_{m\text{-}1}$ and $k = t_m - t_{m\text{-}1} - 1$

$$b_{i_1}(t_m, t_{m-1}) \geq m_{ij}(k)\;\alpha_j \geq w_{ii_1}(t_m-1)\cdots w_{i_{k-1},j}(t_{m-1}+1)\alpha_j\,.$$

Because of $w_{ii_1}(t_m-1)\cdots w_{i_{k-1},j}(t_{m-1}+1) \geq w$ and $k = t_m - t_{m\text{-}1} - 1 \leq r - 1$ by assumption we arrive at inequality [18]. The latter implies in particular inequality [15] which together with $\sum_{m=1}^{\infty}\delta_m = \infty$ for $\delta_m = \min\{\alpha_i \mid \alpha_i > 0\}w^{r\text{-}1} > 0$ for all $m$ completes the proof of Theorem 2. □

## References

ABELSON, R P (1964), "Mathematical models of the distribution of attitudes under controversy". In Frederiksen N and Gulliksen H (Eds.), *Contributions to Mathematical Psychology,* New York, NY: Holt, Rinehart, and Winston.

BECKMANN T (1997) *Starke und schwache Ergodizität in nichtlinearen Konsensmodellen*. Diploma thesis: Universität Bremen, 64 p.

BOGDAN R J (Ed) (1981) *Keith Lehrer*, Dordrecht: D. Reidel Publ. Co.

CHATTERJEE S (1975) Reaching a consensus: Some limit theorems. *Proc. Int. Statist. Inst.* pp. 159–164.

CHATTERJEE S and Seneta E (1977) Toward consensus: some convergence theorems on repeated averaging. *J. Appl. Prob*. 14. pp. 89–97.

CHESNEVAR I C, Maguitman A G and Loui R P (2000) Logical Models of Argument. *ACM Computing Surveys* (CSUR) Volume 32, Issue 4 (December 2000) Pages: 337–383.

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

DE GROOT M H (1974) Reaching a consensus. *J. Amer. Statist. Assoc*. 69. pp. 118–121.

DITTMER J C (2000) *Diskrete nichtlineare Modelle der Konsensbildung.* Diploma thesis: Universität Bremen.

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

ENGEL, P (2004) "Truth and the Aim of Belief". In Gillies D. (Ed), *Laws and Models in Science.* London: King's College Publications. pp. 79–99.

FORTUNATO S (2004) The Krause–Hegselmann consensus model with discrete opinions, *Int. J. Mod. Phys.* C15 (7).

FORTUNATO S (2005) On the Consensus Threshold for the Opinion Dynamics of Krause – Hegselmann. Cond-mat/0408648 at www.arXiv.org. To appear in *Int. J. Mod. Phys. C16, issue 2*.

FORTUNATO S, Latora V, Pluchino A and Rapisarda A (2005) Vector Opionion Dynamics in a Bounded Confidence Consensus Model. *International Journal of Modern Physics C* (to appear).

FORTUNATO S and Stauffer D (2005) "Computer Simulations of Opinions and Their Reactions to Extreme Events". In Albeverio S, Jentsch V and Kantz H (Eds), *Extreme Events in Nature and Society*, Heidelberg: Springer. To appear.

FRENCH J R P (1956) A formal theory of social power. *Psychological Review* 63. pp. 181–194.

FRIEDKIN N E and Johnsen E C (1990) Social influence and opinions, *J. Math. Soc*. 15. pp. 193–206.

FULLER S (2002) *Social Epistemology*. Bloomington, Indiana UP, 1st ed. 1988.

GALAM S, Gefen Y and Shapin Y (1982) Sociophysics: A mean behavior model for the process of strike. *J. Math. Soc*. 9, 1–13.

GALAM S and Moscovici S (1991) Towards a Theory of Collective Phenomena: Consensus and Attitude Changes in Groups. *European Journal of Social Psychology* 21, pp. 49–74.

GOLDMAN A (1999) *Knowledge in a Social World*. Oxford: Oxford University Press.

GOLDMAN A (2001) "Social epistemology". In *Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/epistemology-social/

HABERMAS J (1973) "Wahrheitstheorien". In Fahrenbach H (Ed), *Wirklichkeit und Reflexion – Walter Schulz zum 60. Geburtstag*, Pfullingen: Neske. pp. 211–265.

HARARY F (1959) "A criterion for unanimity in French's theory of social power". In Cartwright D (Ed.), *Studies in Social Power*. Ann Arbor: Institute for Social Research.

HEGSELMANN R, Flache A and Möller V (1999), "Cellular Automata Models of Solidarity and Opinion Formation: Sensitivity Analysis". In Suleiman R, Troitzsch K G and Gilbert N (Eds), *Social Science Microsimulation – Tools for Modeling – Parameter Optimization and Sensitivity Analysis*, Heidelberg: Physica–Verlag. pp. 151–178.

HEGSELMANN R and Krause U (2002) Opinion Dynamics and Bounded Confidence – Models, Analysis, and Simulations. *Journal of Artificial Societies and Social Simulation (JASSS)* vol.5, no. 3 http://www.soc.surrey.ac.uk/JASSS/5/3/2.html.

HEGSELMANN R and Krause U (2005) Opinion Dynamics Driven by Various Ways of Averaging. *Computational Economics* 25. pp. 381–405.

HEGSELMANN R (2004) "Opinion Dynamics – Insights by Radically Simplifying Models" In Gillies D (Ed), *Laws and Models in Science*, London: King's College Publications. pp. 19–46.

HEGSELMANN R (2004a), *Das Trichter-Theorem.* Bayreuth: Manuscript.

KRAUSE U (1997), "Soziale Dynamiken mit vielen Interakteuren. Eine Problemskizze". In Krause U and Stöckler M (Eds.), *Modellierung und Simulation von Dynamiken mit vielen interagierenden Akteuren,* Universität Bremen. pp. 37 – 51.

KRAUSE U (2000), "A discrete nonlinear and non—autonomous model of consensus formation". In Elaydi S, Ladas G, Popenda J and Rakowski J (Eds.), *Communications in Difference Equations*, Amsterdam: Gordon and Breach Publ.. pp. 227 – 236.

KRAUSE U (2003), "Positive particle interaction". In Benvenuti L, De Santis A and Farina L (Eds.), *Positive Systems*, Berlin etc.: Springer. pp. 199–206.

KRAUSE U (2005), "Time variant consensus formation in higher dimensions". In Elaydi S, Ladas G, Aulbach B and Dosley O (Eds), *Proceedings of the 8th International Conference on Difference Equations and Applications, 2003,* Boca Raton etc.: Chapman & Hall/CRC. pp. 185–191.

KRAUSE U and Nesemann T (1999) *Differenzengleichungen und diskrete dynamische Systeme*, Leipzig: Teubner.

KUIPERS, T A F (2000) *From Instrumentalism to Constructive Realism. On Some Relations between Confirmation, Empirical Progress, and Truth Approximation*. Synthese Library, Volume 287. Dordrecht: Kluwer Academic Publishers.

LEHRER K (1975) Social Consensus and Rational Agnoiology. *Synthese* 31. pp. 141–160.

LEHRER K (1976) When rational disagreement is impossible. *Nous* 10. pp. 327–332.

LEHRER K (1977) Social Information. *The Monist* 60. pp. 473–487.

LEHRER K (1981), "A Self Profile". In Bogdan R J (Ed), *Keith Lehrer*, Dordrecht: D. Reidel Publ. Co.. pp. 3–104.

LEHRER K (1981a)," Replies". In Bogdan R J (Ed), *Keith Lehrer*, Dordrecht 1981: D. Reidel Publ. Co.. pp. 223–242.

LEHRER K (1985) Consensus and the Ideal Observer. *Synthese* 62. pp. 109–120.

LEHRER K and Wagner C G (1981) *Rational Consensus in Science and Society.* Dordrecht: D. Reidel Publ. Co.

LEVI I (1985) Consensus as Shared Agreement and Outcome of Inquiry. *Synthese* 62. pp. 3–11.

LOEWER B (Ed) (1985) Consensus. *Synthese* 62 (special issue).

LORENZ J (2003) *Mehrdimensionale Meinungsdynamik bei wechselndem Vertrauen*. Diploma Thesis: Universität Bremen. http://www-stuga.informatik.uni-bremen.de/mathematik/-archiv/diplome/jlorenz.zip.

LORENZ J (2003a) *Opinion dynamics with different confidence bounds for the agents.* Preprint, www.janlo.de.

LORENZ J (2005) A stabilization theorem for dynamics of continuous opinions. *Physica* A, 355(1). pp. 217–223.

NOWAK A, Szamrez J and Latane B (1990) From private attitude to public opinion – Dynamic theory of social impact. *Psych. Review* 97. pp. 362-376.

SCHMITT F (1999), "Social Epistemology". In Greco J, Sosa E (Eds.), *The Blackwell Guide to Epistemology*. Oxford: Blackwell. pp. 354-382.

SCHMITT F (1994), *Socializing Epistemology – The Social Dimensions of Knowledge.* Lanham: Rowman & Littlelfield Publishers, Inc.

SZNAJD–WERON K and Sznajd J (2000) Opinion evolution in closed community. *International Journal of Modern Physics* C 11. pp. 1157-1166.

STAUFFER D (2002) Monte Carlo simulations of the Sznajd model. *Journal of Artificial Societies and Social Simulations*, vol. 5, no. 1.
http://www.soc.surrey.ac.uk/JASSS/5/1/4.html.

STAUFFER D (2003) How to convince others? *American Institute of Physics, Conference Proceedings* 690, pp. 147-155.

STAUFFER D (2004) Difficulty for consensus in simultaneous opinion formation of Sznajd model. *J. Math. Soc*. 28. pp. 25-33.

URBIG D and Lorenz J (2004) Communication regimes in opinion dynamics: Changing the number of communicating agents. *Proceedings of the Second Conference of the European Social Simulation Association (ESSA).*

WEISBUCH G, Deffuant G, Amblard F and Nadal J P (2001) Interacting agents and continuous opinion dynamics. http://arXiv.org/pdf/cond-mat/0111494.