A Dialogical Logic-based Simulation Architecture for Social Agents
and the Emergence of Extremist Behaviour

Piter Dykstra[1,2], Corinna Elsenbroich[3], Wander Jager[1], Gerard Renardel de Lavalette[1], Rineke Verbrugge[1]

[1] Groningen University, The Netherlands,
P.Dykstra@rug.nl
[2] Hanzehogeschool Groningen
[3] University of Surrey,
Guildford, United Kingdom

Abstract. In the flourishing research area of agent-based social simulation, the focus is on the emergence of social phenomena from the interactions of individual autonomous agents. There is, however, a relative underexposure of the cognitive properties of agents, as the existing agent architectures often focus on behaviour alone. Cognition becomes particularly salient when the subject under investigation concerns social phenomena where agents need to reason about other agents' beliefs. We see this as a requirement for any communication with some degree of intelligence. In this paper we use concepts and methods from dynamic epistemic logic to build agents capable of reasoning about other agents' beliefs and their own. In dynamic epistemic logic, agents are assumed to be perfect rational reasoners. We break with this unrealistic assumption in order to bridge the gap between the sociological and the logical approach. Our model is based on a minimal set of assumptions representing cognitive processes relevant to modelling the macro-phenomena of group formation and radicalisation.

1 Introduction

Group formation is a social phenomenon observable everywhere. One of the earliest agent-based simulations, the Schelling model [25] of segregation, shows how groups emerge from simple preferences of agents. But groups are not only about segregation according to some differentiating property. Groups have the important feature of possible radicalisation, a group-specific dynamics of opinions.

Radicalisation is a thorn in the side of our liberal society. We have adopted the maxim that difference in opinions enriches society whilst extreme radicalisation, on the other hand, is a threat to society. Terrorist organisations like AlQaeda, ETA or the German RAF result from extreme radicalisation of groups. There are some features that distinguish group radicalisation from mere group formation: (1) exclusion of out-group members (between-group), (2) limitation of opinion

variance (within-group). These two features strongly suggest that group radicalisation is a phenomenon that can occur during agent interaction, i.e. distancing oneself from another agent, or aligning one's opinion to others. Both procedures presuppose the capacity to reason about other agents' beliefs. We want to analyse whether cognitive allegiance, in the form of the alignment of opinions, is sufficient to explain extreme group radicalisation. We investigate this process by using agent-based simulation with cognitive agents.

The development of extremist behaviour has been a topic of investigation in social simulation. There are different approaches, see for example Deffuant [8, 7], Kaufmann et al. [13] and Carley [24]. Existing cognitive architectures do offer an agent model with the capability of representing beliefs, but primarily knowledge is represented in the form of event-response pairs, i.e. perception of the external world and a behavioural reaction to it. We are interested in beliefs in the form of *opinions* rather than in actions. What matters are the socio-cognitive processes leading to a situation in which a group of agents see where opinions radicalise and out-groups are excluded. To model the emergence of a group ideology we need agents who communicate with each other and are capable of reasoning about their own opinions and the opinions they believe other agents to have.

Sherif and Hovland [26], the founders of Social Judgement Theory, describe the conditions under which the change of attitudes takes place, with the intention to predict the direction and extent of that change. An agent's attitude can be understood in terms of what other positions it finds acceptable or not. These sets of opinions are referred to as the "latitudes of *acceptance*, *rejection*, and *noncommitment*". Jager and Amblard [19] demonstrate that the attitude structure of agents determines the occurrence of *assimilation* and *contrast* effects, which in turn cause a group of agents to reach consensus, to bipolarize, or to develop a number of subgroups sharing the same position. In this theory agents engage in social processes. This framework contains a represention of opinions, but there is no logic for reasoning about the different agents and opinions.

2 Social Cognition for Social Agents

Nowadays the mutual interaction between agent-belief and social context is a major topic in the study of Social Network Analysis (SNA), see for example Sun [28] and Breiger [4]. *Construct* is a social simulation architecture developed on the basis of this principle, cf. [21]. There is also research concerning social acquired knowledge directly. Helmhout [17, 16] for example calls this knowledge *Social Constructs*, but the agents in his work, just as the agents in the work of Sun and Breiger, lack real reasoning capabilities, in particular they do not have higher-order social knowledge.

Representing beliefs in the form of a logic-based epistemic theory seems well suited for the kind of higher-order reasoning we need to represent, see for example [11]. The prevalent logical approaches for multi-agent reasoning are either game-theoretic, cf. [29] or based on dialogical logic, cf. [30, 5]. Neither of these logical approaches is usable in agent-based simulation, due to the assumption of perfect

rationality which requires unlimited computational power. This problem was solved by the introduction of the concept of *bounded rationality* and the Belief Desire Intention (BDI) agent, cf. [3]. The BDI agent is used successfully in the computer sciences and AI, in particular for tasks involving planning. See Dunin-Kęplicz et al. for a discussion of such bounded awareness [10]. Learning in a game theoretical context has been studied by Flache and Macy [23]. There are extensions to the basic BDI agent, such as the BOID, integrating knowledge of obligations and work on a socially embedded BDI agent, for example by Subagdja et al.[27] and Dignum et al.[9].

In agent-based simulation, the agent model can be seen as a very simple version of the BDI agent [17, 16]. Agents in simulations make behavioural decisions on the basis of their preferences (desires) and in reaction to their environment (beliefs). Their reasoning capacity is solely a weighing of options; there is no further reasoning. This is partly due to social simulations engaging with large-scale populations for which an implementation of a full BDI agent is too complex. A full implementation might not be helpful either, as these agents are potentially overdesigned with too much direct awareness, for example of obligations and constraints to justify social phenomena as emergent rather than inbuilt to the simulation. The main problem with cognitively poor agents is that, although emergence of social phenomena from individual behaviour is possible, we cannot model the feedback effect of social phenomena on individual behaviour as these agents are unable to reason about these social phenomena. This kind of feedback is very relevant to distinguish group radicalisation from mere group formation.

We propose a model architecture for the simulation of social phenomena based on intelligent social agents with logic-based cognition derived from dialogical logic but dropping the unrealistic assumption of omniscience. To model social beliefs adequately, the agents need to be able to

1. reason about their own and other agents' beliefs,
2. communicate these beliefs to other agents,
3. revise their belief base, and
4. establish and break social ties.

Humans do not acquire their beliefs about the world in a social vacuum and neither should agents in a simulation. In our model agents acquire their beliefs by communication with other agents. So, the truth of a sentence cannot be reduced to the truth of verifiable facts; instead we have a *social construction of beliefs*. We follow Carley's idea of *constructuralism*.

> "Constructuralism is the theory that the social world and the personal cognitive world of the individual continuously evolve in a reflexive fashion. The individual's cognitive structure (his knowledge base), his propensity to interact with other individuals, social structure, social meaning, social knowledge, and consensus are all being continuously constructed in a reflexive, recursive fashion as the individuals in the society interact in the process of moving through a series of tasks. [...] Central to the constructuralist theory are the assumptions that individuals process and communicate information during interactions, and

> that the accrual of new information produces cognitive development, changes in the individuals' cognitive structure." [6, p.386]

On the other hand, agents choose their social network on the basis of the similarity of beliefs. This captures the main principle of Festinger's Social Comparison Theory [12], stating that people tend to socially compare especially with people that are similar to them. This phenomenon of "similarity attracts" has been demonstrated to apply in particular to the comparison of opinions and beliefs, e.g, Suls, Martin, and Wheeler [18]. Lazarsfeld and Merton [22] refer to this process by the term 'value homophily, stating that people tend to associate themselves with others sharing the same opinions and values. To capture this "similarity attracts" process in our model, we need to model the social construction of knowledge bi-directionally. Hence agents prefer to interact with agents having similar opinions.

3 Topic Space and Agent Goals

Let us first characterize our agents, where they are located and what they do to achieve the bi-directionality. Our agents are located in a *topic space*, a spatial representation of opinions. Most models, in particular those mentioned in Sections 1 and 2, are a-spatial. A family of models of opinions that make some use of space are those using the concept of *Bounded Confidence* [15] but the agents are not mobile. In those models opinions vary continuously, and uncertainty and change of opinion are modelled by numerical computations. For our simulation, in contrast, we use mobile agents in a metric space to represent both the dynamics in social contacts and the dynamics of opinion. As agents move about, they change both their social network and their opinions. We provide the model with a realistic communication protocol on the basis of dialogical logic. Communication is made efficient by providing agents with the capability to *generalise*, i.e. to assume a specific opinion to be representative of a group, rather than just a single agent.

A common way for an agent to start a communication in a simulation model is to pick at random another agent to communicate with. In contrast, our agents make announcements that are heard by all agents within a certain range. The topic space serves as the 'postman' for the agents, i.e. a medium through which the messages travel. An agent can use any amount of energy for the utterance of a message. We call that energy *loudness* and it decreases as the message travels through the space. The transmission of the message stops, when the loudness falls below some threshold. So agents can decide whether they will communicate with the agents that are socially close to them or with a larger group.

Now that we have located our agents in an opinion space, we need to give them something to do. The dialogical semantics we introduce in Section 4.1 are based on an argumentation game. An agent, the *proponent*, who *announces* a statement enables another agent, the *opponent*, to *attack* that statement and a dialogue will be the consequence of these two actions. The game is a sequence of *alternating* actions by the proponent and the opponent.

Kamlah and Lorenzen [20] introduced this type of dialogue in order to define a construction for the concept of *logical truth*.[4] A proposition, φ, is logically true iff a proponent of φ has a *winning strategy* (i.e. the proponent is always able to prevent an opponent from winning) for a *formal dialogue* about φ. A formal dialogue is one in which no assumption about the truth of any propositions is made. In our dialogues such assumptions are made: we do not model a formal but a *material dialogue*, transforming the dialogue game to one in which the proof of a complex statement is reduced to the proof of a simple statement. The winner of the proof, and therefore the whole dialogue, is decided by a group vote rather than a *winning strategy*.

But why should agents announce or attack propositions in the first place? Let our agents have a *reputation status* (RS). Further, let RS increase with the amount of information an agent has. Our agents thus want to maximise their information which can only be done via communication with other agents. The game is played as follows: Announcing a statement φ entails that the agent declares itself prepared to offer a payment P of RS-units in case it loses the dialogue on φ. In case it wins it will gain a reward R of RS-units. Agreeing with a proposition is cost-free. Different gaming rules are possible:

1. the proponent pays a certain amount, say $1RS$ to the opponent, in case it loses the dialogue and gets nothing in case it wins. Only when an agent is absolutely sure about φ will it be prepared to enter into such an argument.
2. the proponent offers $1RS$ to the opponent if it loses the dialogue, but it receives $1RS$ in case it wins the dialogue. This opens the opportunity to gamble. If the agent believes that the chance of φ being true is greater than 50 %, it may consider a dialogue about φ as a game with a profitable outcome in the long run.
3. the proponent may specify any amount P it is prepared to pay in case of losing, and any amount R it wants to receive in case of winning the dialogue.

The first rule was proposed by Giles [14] for the Lorenzen type of dialogue to define logical truth for expressions that contain statistical propositions. The other two are our variants to get the communication started. Our choice is the third option because it not only enables a proponent to express the *degree of belief* (a high P/R-ratio)[5] in a proposition, like the second option, it also expresses the *greed* to start a dialogue about that proposition (a high stake expressed by $P + R$). The P/R-ratio reflects an agent's belief in the chances of winning the dialogue. If an agent a believes to win the dialogue about a proposition in 2/3 of the cases, it might consider offering $P_a = 2RS$ to a reward $R_a = 1RS$ as the least ratio that is still reasonable. For an agent b who disagrees with a proposition φ with values, P_b and R_b, attacking that proposition is a rational

[4] We introduce a formal semantics in Section 4.1 using a two-dimensional truth composed of the *evidence* for and the *importance* of a proposition.

[5] The case $R = 0$ is the case an agent is prepared to pay when losing, but does not get a return in case of winning the dialogue. This makes sense only in the case the agent believes that it stated an absolute (logical) truth.

thing to do. $P_b/R_b < P_a/R_a$ assigned to φ means that opponent b expects to win a dialogue about φ more often than the proponent a believes it does.

The information of other agents is the only source of evidence to achieve maximally corroborated information, expressed by the vote on the winner. When a dialogue has come to an end, there is only one simple statement to prove. The agents in the environment of the proponent and the opponent are asked whether they agree on that statement or its negation. In case both environments come to the same verdict the communication is resolved. In case the two environments disagree there is no resolution. In that case the two opponents decrease their mutual trust. For agents it is advantageous to be in agreement with their local environment, i.e. get the vote by those around. There are a few possibilities to achieve this agreement:

1. try to convince the other agents of your position by attacking their opinions.
2. chase agents with deviant opinions away.
3. adopt the opinion of the majority of the environment or try to convince them of the unimportance of that opinion, also known as. preaching freedom of opinion.
4. move to another place where the agents have a more similar state of mind.

Which strategy is applied by an agent is determined by its disposition and its experience in the past, i.e. its memory of received messages. The action choices are constrained by an agent's amount of RS. The first two strategies are aggressive and destructive, probably leading to small, relatively closed groups. The last two strategies are peaceful, probably leading to larger, fuzzy groups that accept new members easily. However, both types have survival value in terms of life points so there is no initial dominant strategy.

Example 1 Anya is part of a group where they have just discussed that schools are terribly underfunded and that language teaching for newly arrived refugees is particularly hard to organise. Bob approaches.

A: If we do not have enough funding and refugees create additional costs, then they should leave this country.
B: Why do you say that?
A: Well, prove me wrong! If you can I give you 1RS but if I am right you have to pay me 0.50RS. [These odds mean that Anya is pretty sure about the truth of its statement.] *Everyone here says that schools are underfunded and that language teaching poses a burden!*
B: Ok, deal. I really disagree with your conclusion. . . .
A: Do you agree then that we do not have enough funding and that refugees create additional costs.
B: Yes of course, that can hardly be denied.
A: Right. . . – people, do you believe refugees should leave this country?

[Clearly, the vote can go either way. They might agree or Anya might have miscalculated the beliefs of the group. Depending on the outcome of the vote, the payment will be made and Anya potentially wants to leave the group setting or reevaluate her opinion.]

4 Epistemic Logic

Agents need to be able to reason about the beliefs of other agents. Defining this concept in predicate logic is possible but cumbersome. Using epistemic logic enables the agents to reason about their own and others' beliefs. So, whereas traditional logic can express that a proposition φ is true, epistemic logic can express that an agent a believes a proposition φ by the formula $B_a\varphi$ or even that a believes that b believes φ by the formula $B_aB_b\varphi$. B is a belief operator. The logical language consists of a first-order calculus plus the belief operator B relating to either an agent or a group of agents and a relation between two agents (or agent groups) we denote by $Trust$. In addition to the epistemic part of the logic the logic is extended by spatial operators, quantification over agent-sets AND actions, like *Announce* and *Ask*, that lead to the dynamics of the logic. The cost and benefit of actions in reputation status is expressed in the values of P and R. For example, the meaning of $Ask_a(b, \varphi)$ is: Agent a asks agent b what its valuation of proposition φ is. The reply is an announcement $Announce_b(\varphi, P, R)$, which means that b is prepared to defend φ in a dialogue against any opponent who is prepared to pay R in the case b wins and b is prepared to pay P in case it loses the dialogue. (See Technical Appendix A for the full definition and examples.) We also treat belief revision in logic; that will save us an extra layer for the implemention of strategies. The Alchourron-Gärdenfors-Makinson-style [1] is however not appropriate for our imperfectly reasoning agents, because they treat belief as a form of agent knowledge that can conservatively be revised on purely rational grounds. They do not deal with trust or degrees of belief, nor with beliefs about beliefs, and they preserve consistency in the agent belief state at all times.

Our variant of epistemic logic has a relatively simple syntax formalising conceptual beliefs about other agents. Given the above syntax, we need to ensure that the semantics of the system express what is needed. As we are not interested in observations of events but in beliefs acquired from the interaction with other agents, we will embed the syntax into a semantic framework of dialogues.

4.1 Two-Dimensional Truth

Given the game our agents play via the dialogue action formulae, there are some initial requirements appropriate semantics have to fulfill. Given the values P and R, the semantics need to allow agents to believe in a statement φ as well as $\neg\varphi$ by some measure, i.e. the semantics needs to allow for paraconsistency. Our logic is related to the relevance logic of Anderson and Belnap [2]; it is a generalization of their 4-valued logic to infinite values. In both systems truth-values can be considered as a composition of two factors: the P and R values an agent assigns to a statement. These are values of the continuum [0,1], so we can speak of a two-dimensional truth-value space represented by the surface of a square, cf. Figure 5.

The corners of the square correspond to the Anderson-Belnap truth values:

1. The points $(1,0)$ and $(0,1)$ correspond to the classical truth values *true* and *false*. They correspond to the game situations in which an agent is willing to pay without receiving a reward and only receive a reward without paying, i.e. maximal belief and disbelief.
2. The point $(0,0)$ represents the total void or the value *unknown*. The agent does not enter a game.
3. $(1,1)$ means a high stake for a proposition of which the agent has no clue about the truth value. It is the paradoxical situation in paraconsistent logic.

The values P and R can be interpreted as the expression of the more intrinsic concepts: the degree of belief or the evidence (E) an agent has in favour of a proposition and the importance (I) of that proposition. Evidence can also be interpreted as *truth* with values in [0,1] and importance as a similar version of *relevance*. We define:

$$E(\varphi) = \frac{P(\varphi)}{P(\varphi) + R(\varphi)} \qquad I(\varphi) = \frac{P(\varphi) + R(\varphi)}{2}$$

We now have a formal system that is dynamic and expressive enough for our purposes. For the formal definition and an example formalisation of Example please refer to Technical Appendix B.

5 The Program RationalAgent

A proof of concept has been implemented in Netlogo to demonstrate the effect of argumentation.[6] At the start of a simulation, a number of agents will be initialised with random opinions between 1(convinced of a proposition) and -1 (convinced of the negation). The agents value their importance with a random value between 0 (unimportant) and 1 (very important, $\frac{1}{2}$ is default). Agents are represented as coloured circles (figure 1); in the middle window the colour

[6] The model can be downloaded at *http://www.math.rug.nl/~piter/ESSA*. The filename is: *RationalAgentXX.nlogo*, where *XX* is the version number.

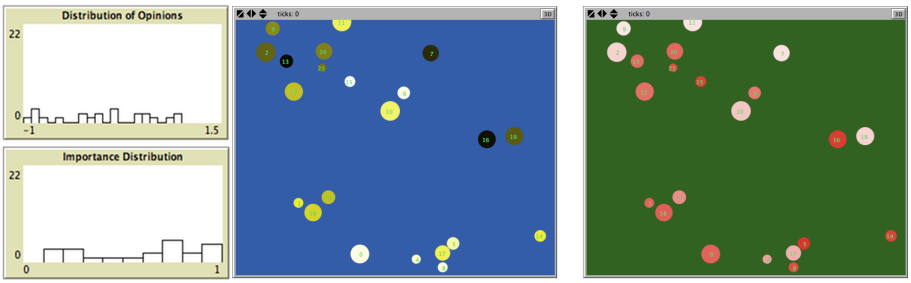

Fig. 1. Screenshots of the RationalAgent program in an initial state.

represents the evidence and in the right window the color represents the importance. The diameter represents in both cases the reputation status. Agents can announce their opinion, attack and defend announcements, move in the direction of the area where their opinion is most often heard. The likelihood of these actions is determined by the user and can be changed while a simulation is running. The loudness of speech and the visual horizon for moving can be changed likewise. The value of parameters are the same for all agent in order to keep the simulation simple.

When an agent a hears an announcement by agent $b, Announce_b(\varphi, P, R)$, it will know that the evidence value assigned to φ by b, is $\frac{P}{P+R}$. If a and b both believe φ, so $E_b(\varphi) > 0.5$ and $E_a(\varphi) > 0.5$ then a increases its belief in φ (asymptotically to 1) and decreases the importance of φ (asymptotically to 0).

What agents in the neighbourhood of b believe becomes visible as a bright cloud around b in the evidence window and as a dark cloud in the importance window. (See figure 2, around agents. 2, 17 and 21). The evidence window represents what the agents think that "The Others" think. For reasons of simplicity, it is assumed that all agents think the same about what "The Others" on a certain spot inside their visual circle think, so there is only one evidence - and one importance window.

We have a similar situation when both agents agree on the falsity of φ. In that case $E_b(\varphi) \leq 0.5$, then a decreases its belief in φ (asymptotically to 0). The importance still decreases too, so in both windows the area around b becomes darker. (See figure 2: around agents 0, 18 and 19).

On the other hand, if a message φ is in contradiction with the belief-base of a, the evidence is decreased and the importance is increased; the truth value approaches the paradoxical value. That means the clouds disappear in the evidence window, while in the importance window bright spots appear (See figure 2: between agents. 21 and 22 there is a conflict zone). Announcements made will be "forgotten" and the cloud will fade away as time passes, unless the statement is repeated.

The behaviour of the agents contributes to their RS, and the RS on its turn effects the loudness of speech. Two methods for the computation of the RS values are implemented: (1) The RS value reflects the success of being in harmony with the environment. Each round the RS value of an agent is incremented/decremented by a value proportional to the gain/loss in similarity with its environment. (2) The RS value is determined by the outcome of the dialogues. On winning an attack/defence of a statement the RS value is incremented by respectively the P/R amount of the attacked/defended statement. On losing an attack/defence, RS is decremented by respectively the R/P value of the statement.

5.1 The Results

In most of the runs a stable configuration is established after a number of rounds. Most parameters only influence the number of rounds needed to reach a stable configuration and the type of the configuration in a lesser degree. The method of

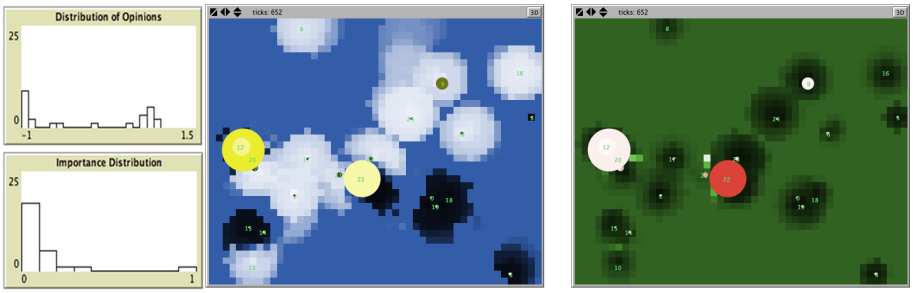

Fig. 2. The simulation after 652 rounds with a RS represents the agents succes of being in an environment that agrees with its opinion.

the RS calculation is the most important factor in the outcome of the simulation runs:

1. If RS measures the agreement of the agents with their environment, small groups of agents with similar opinions will emerge. In some groups the distance between the agents is small and the agents are most similar (see figure 2 the group consisting of the agents 0, 18 and 19). When the mutual distance is greater, the opinions differ more (cf. the group of agent 2, 17 and 21). Differences in RS are not very large. The number of different opinions is reduced; the extreme ones become dominant, but the importance (the need to argue about those opinions) drops.
2. In case RS reflects the balance of paying and receiving according the dialogue rules, only one agent will have almost all the RS points and the other agents have not enough status to announce or to attack a statement (see figure 3). As a consequence, there are hardly groups (no one to go to) except for "the great dictator" in whose environment opinions have no importance at all. The distribution of opinions in the mind of the agents and their importance,

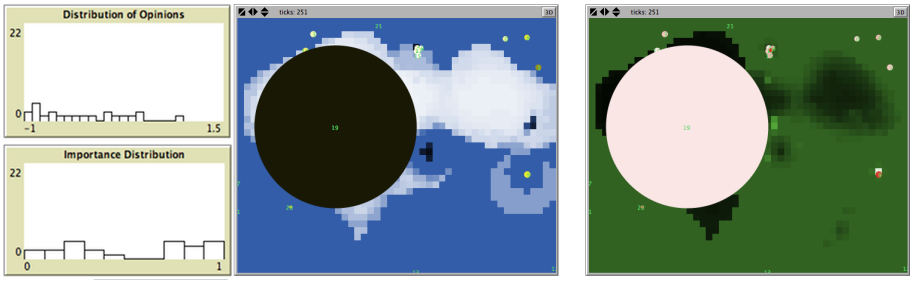

Fig. 3. The state after 251 rounds. RS is the result of winning and loosing dialogues.

has hardly changed since the beginning of the runs however. So in this case the RS values become extreme.

6 Concluding Remarks and Future Work

We have presented a basic syntax and semantics for a cognitive agent-based model of group communication. Agents are involved in a dialogue game in which they gamble reputation status (RS) on the truth of a statement. Dialogue conflicts are resolved by appeal to surrounding agents in such a way that the more supported statement wins the dialogue. Our model will show the process of radicalisation in groups based on three simple ingredients: movement in topic space, communication with others in adversarial or agreeable dialogue, and the ability to reason about other agents' beliefs.

The first future extension of the model is an implementation of the spatial and historic memory of the agents. The history memory is about the evidence for an agent a's own beliefs, the spatial memory is the evaluation of the beliefs of other agents by agent a. The spatial memory is the agent's copy of the topic space and is important for agents' strategic gambles as it contains the information 'where status can be enhanced', i.e. where agents with similar opinions are. The history memory is important for the update mechanism. We need to be careful to implement memory in such a way that agents can forget, as otherwise the cumulative expansion will reduce the impact of new announcements too much. Other future extensions of the model are to incorporate higher-order thinking in such a way that agents can lie, i.e. pretend that their evidence values are higher than they actually are. We also need to implement the reach of a message (its loudness) and a corresponding cost structure for making announcements.

Our model offers the possibility to study the conditions which may cause the emergence of different types of groups and the development of ideas inside those groups in relation to other groups. An extremist agent is one that has strong convictions concerning extreme opinions. Such an extremist attitude is unlikely when an agent has a reasonable number of contacts with different kinds of agents, which will almost be guaranteed when agents have enough freedom of movement. But an unsatisfactory communication between an agent or a group and the rest of the community may cause a vicious circle of alienation and extremization. A final question in our quest for the cause of extremism is: What may cause the loss of freedom of movement or the isolation of groups of agents?

References

1. C.E. Alchourron, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic*, (50):510–530, 1985.
2. Alan R. Anderson and Nuel D. Belnap. *Entailment: The Logic of Relevance and Necessity*, volume 1. Princeton University Press, Princeton, 1975.
3. M. E. Bratman. *Intention, Plans and Practical Reason*. CSLI Publications, Stanford, 1999.

4. R. Breiger, K. M. Carley, and P. Pattison. Dynamic social network modelling and analysis: Workshop summary and papers. *J. Artificial Societies and Social Simulation*, 6(4), 2003.
5. J.M. Broersen, M. Dastani, and L. van der Torre. Beliefs, obligations, intentions and desires as components in an agent architecture. *International Journal of Intelligent Systems*, 20(9):893–920, 2005.
6. K. M. Carley. Knowledge acquisition as a social phenomenon. *Instructional Science*, 14(4):381–438, 1986.
7. G. Deffuant. Comparing extremism propagation patterns in continuous opinion models. *Journal of Artificial Societies and Social Simulation*, 9, 2006.
8. G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. How Can Extremism Prevail? A Study Based on the Relative Agreement Interaction Model. *Journal of Artificial Societies and Social Simulation*, 5(4), 2002.
9. F. Dignum, B. Dunin-Kęplicz, and R. Verbrugge. Agent theory for team formation by dialogue. In *Agent Theories, Architectures and Languages*, pages 150–166, London, UK, 2001. Springer-Verlag.
10. B. Dunin-Kęplicz and R. Verbrugge. Awareness as a vital ingredient of teamwork. In *AAMAS '06: Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1017–1024, New York, NY, USA, 2006. ACM.
11. R. Fagin and J. Y. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34(1):39–76, 1987.
12. L. Festinger. A theory of social comparison processes. *Human Relations*, 7(2):117–140, 1954.
13. D. W. Franks, J. Noble, P. Kaufmann, and S. Stagl. Extremism propagation in social networks with hubs. *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems Archive*, 16(4):264–274, August 2008.
14. R. Giles. Formal languages and the foundation of physics. In *Physical Theory as Logico-Operational Structure*, pages 19–87, Dordrecht, Holland, 1978. Reidel Publishing Company.
15. R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.
16. J. M. Helmhout. *The Social Cognitive Actor*. PhD thesis, University of Groningen, Groningen, November 2006.
17. J. M. Helmhout, H. W. M. Gazendam, and R. J. Jorna. Emergence of social constructs and organizational behavior. *21st EGOS colloquium, Berlin.*, 2005.
18. Suls J., Martin R., and Wheeler L. Social comparison: Why, with whom and with what effect? *Current Directions in Psychological Science*, 11(5):159–163, 2002.
19. W. Jager and F. Amblard. Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory*, 10(4):295–303, 2005.
20. W. Kamlah and P. Lorenzen. B.I.-Wissenschaftsverlag, Mannheim.
21. R. W. Lawler and K. M. Carley. *Case Study and Computing : Advanced Qualitative Methods in the Study of Human Behavior*. Norwood, NJ Ablex, 1996.
22. P. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. In: M. Berger, T. Abel, and C. H. Page, eds. *Freedom and Control in Modern Society*, pages 18–66., New York: Van Nostrand, 2002.
23. A. Flache. Macy, M.W. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences U.S.A.*, 14(99):7229–36, May 2002.

24. K. M.Carley, J. Reminga, and N. Kamneva. Destabilizing terrorist networks. In *NAACSOS conference proceedings*, 2003.
25. T. C. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, (1):143–186, 1971.
26. M. Sherif and C.I. Hovland. *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change.* New Haven: Yale University Press., 1961.
27. B. Subagdja, L. Sonenberg, and I. Rahwan. Intentional learning agent architecture. *Autonomous Agents and Multi-Agent Systems*, 18(3):417–470, 2009.
28. R. Sun. *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation.* Cambridge University Press, New York, NY, USA, 2008.
29. J. van Benthem. Games in dynamic-epistemic logic. *Bulletin of Economic Research*, 53(4):219–48, 2001.
30. D. N. Walton and E.W. Krabbe. *Commitment in Dialogue.* State University of New York Press, 1995.

Technical Appendix

A Formal Language

In addition to *belief* and *trust*, we have spatial operators connecting the agents to the space, quantification over agent-sets and actions that lead to the dynamics of the logic. The cost and benefit of actions in reputation status is expressed in the values of P and R.

Definition 2 (Language) *The language $\mathscr{L}$ of epistemic logic consists of the following formulae:*

$$\begin{array}{lll} \mathscr{L} ::= & p_i \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid B_a\varphi \mid a\ Trusts\ b & (* \text{ epistemic } *) \\ & \forall a(\varphi) \mid \exists a(\varphi) \mid most_a(\varphi) \mid & (* \text{ agent quantification } *) \\ & a\ Moveto\ \vec{x} \mid a\ Atpos\ \vec{x} \mid & (* \text{ spatial formulae } *) \\ & Ask_a(b,\varphi) \mid Announce_a(\varphi,P,R) \mid & (* \text{ dialogue actions } *) \\ & Attack_a(b,F,\varphi) \mid Defend_a(b,\varphi) & \end{array}$$

where:

1. *p_i is an atomic formula.*
2. *φ and ψ are formulae of $\mathscr{L}$.*
3. *a and b are agents from a finite set, or positions, or variables denoting agents or positions.*
4. *$P, R \in [0,1]$ are respectively the pay and the reward of the dialogue about φ after the announcement $Announce_{a_i}(\varphi,P,R)$.*
5. *F stands for the possible moves of attack in a dialogue represented by one of the following: agents, variables denoting agents, or an element from the following set: $\{?, l?, r?\}$.*

The pronounciation of $B_a\varphi$ is: "Agent a believes φ", $a\ Trusts\ b$ reads "Agent a trusts agent b". $\forall a(\varphi)$, $\exists a(\varphi)$ and $most_a(\varphi)$ are pronounced respectively: "For all agents a holds φ", "For some agents a holds φ" and "For most agents a holds φ".
$a\ Moveto\ \vec{x}$ and $a\ Atpos\ \vec{x}$ mean: "Agent a moves to position $\vec{x}$" and "Agent a is at position $\vec{x}$" respectively.

The intuitive meaning of $Ask_a(b, \varphi)$ is: Agent a asks agent b what its valuation of proposition φ is. The reply is an announcement $Announce_b(\varphi, P, R)$, which means that b is prepared to defend φ in a dialogue against any opponent who is prepared to pay R in the case b wins and b is prepared to pay P in case it loses the dialogue.

The dialogue rules for the logical constants ar as follows, an announcement by agent a, $Announce_a(\varphi, P, R)$,$\varphi$, with:

1. $\varphi = \psi \wedge \chi$ can be attacked by $l?$ or $r?$. In response to that a has to defend ψ respectively χ.
2. $\varphi = \psi \vee \chi$ can be attacked by agent b with ?. In response to that a has to defend either ψ or χ to its own choice.
3. $\varphi = \psi \rightarrow \chi$ can be attacked by announcing ψ. In response to that a has to defend by announcing χ.
4. $\varphi = \neg\psi$, can be attacked by agent b, by announcing :$Announce_b(\psi, R, P)$.
5. $\varphi = B_b\psi$ can be attacked by ? and in response agent a has to ask agent b: $Ask(a, \psi)$.
6. $\varphi = \forall x\psi$ can be attacked by presenting an agent b, , and in response to that agent a is supposed to defend the proposition φ applied to agent b: $[b/x]\psi$
7. $\varphi = \exists x\psi$ can also be attacked with ? and in this case agent a has to choose an agent b and defend φ applied to agent b: $[b/x]\psi$

$Proposition$	$\varphi \wedge \psi$		$\varphi \vee \psi$		$\varphi \rightarrow \psi$	$(\neg\varphi, P, S)$	$B_a\varphi$	$\forall x\varphi$	$\exists x\varphi$
$Attack$	$l?$	$r?$	?		φ	(φ, S, P)	?	(agent) $a?$	?
$Defend$	φ	ψ	φ	ψ	ψ		$Ask(a, \varphi)$	$[a/x]\varphi$	$[a/x]\varphi$

Fig. 4. The attack and defence rules for the various forms of propositions.

Example 3 *Examples are:*

1. $B_aB_bp_i$ *means: "Agent a believes that agent b believes* p_i*."*
2. *(Anonymous statement)* $B_{(x,y)}p_i$*: "At position* (x, y) *is an agent who believes* p_i*."*
3. $Announce_a(\varphi, P, R)$*): "Agent a announces that it is prepared to defend in a dialogue proposition* φ *and that it is willing to pay P RS in case it loses the dialogue, to anybody who is prepared to pay R RS in case a wins the dialogue."*
4. $\forall a(a\ Atpos_{(x,y)}) \rightarrow B_xp_i$*: "All agents at position* (x, y) *believe* P_i*."*
5. $B_a(b\ Trusts\ a)$*: "Agent a believes that it has b's trust"*
6. $Attack_{a_i}(b, \varphi)$*: "Agent* a_i *attacks the proposition* φ *uttered by b."*

B Two-Dimensional Truth

Definition 4 *Every agent a has a belief function* $Val_a : \mathscr{L} \rightarrow [0, 1] \times [0, 1]$*. The function* $Val_a(\varphi)$ *is the pair of two functions: the evidence function,* $E_a(\varphi)$ *and the importance function* $I_a(\varphi)$*.*

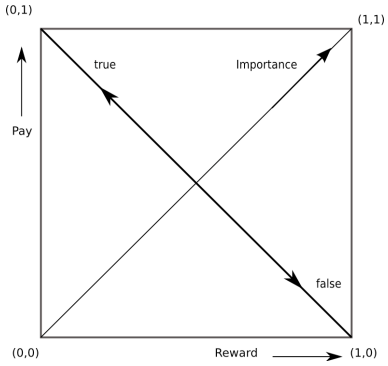

Fig. 5. Two-Dimensional Truth

For complex formulae these functions satisfy the following conditions:

$1: E_a(\varphi \wedge \psi) = min(E_a(\varphi),\ E_a(\psi))$	$I_a(\varphi \wedge \psi) = max(I_a(\varphi),\ I_a(\psi))$
$2: E_a(\varphi \vee \psi) = max(E_a(\varphi),\ E_a(\psi))$	$I_a(\varphi \vee \psi) = \ min(I_a(\varphi),\ I_a(\psi))$
$3: E_a(\varphi \rightarrow \psi) = max(I_a(\varphi),\ E_b(\psi))$	$I_a(\varphi \rightarrow \psi) = min(E_a(\varphi),\ I_b(\psi))$
$4: E_a(\neg\varphi) = I_a(\varphi)$	$I_a(\neg\varphi) = E_a(\varphi)$
$5: E_a(\forall\ i(\varphi(i))) = Min\{E_a(\varphi(i)\|i \in Agents\}$	$I_a(\forall\ i(\varphi(i))) = Max\{E_a(\varphi(i)\|i \in Agents\}$
$6: E_a(B_a\varphi) = E_a(\varphi)$	$I_a(B_a\varphi) = I_a(\varphi)$

But there is no such condition for $Val_a(B_b\varphi)$; agent a has to ask b.

Example 5 We formalise Example 1 about Anya and Bob. *Agent a announces that it believes that everbody who believes that* $\neg\varphi$ *should also believe* ψ *(1). She is prepared to pay* 1 *RS in case it loses a dialogue about this proposition to anybody who is prepared to pay* 0.5 *RS in the case a wins. She is attacked by agent b.*

1.	$\|Announce_a(\varphi \rightarrow \psi, 1, 0.5)$
2. $Attack_b(a, \varphi)$	$\|$
3.	$\|Defend_a(b, \psi)$

At step 3 a could have decided to counterattack step 2 by attacking φ*, but she chooses to defend herself by stating* ψ*.*