# Confronting the Relative Agreement Model to Social Psychology Experiments on Opinion Radicalisation

Guillaume Deffuant*, Frédéric Amblard**, Gérard Weisbuch**,

*Cemagref-LISC
24, avenue des Landais
63172 Aubière
France
**ENS-LPS
24, rue Lhomond
75230 Paris Cedex 05
France

Contact : guillaume.deffuant@cemagref.fr

## Introduction

We consider several agent-based models (Hegselmann and Krause, 2002; Deffuant et al., 2002; Weisbuch et al. 2002) and confront them to the set of social psychology results on consensus radicalisation or averaging after group discussions. A large set of experiments (see Moscovici and Lécuyer, 1972 and Lécuyer, 1974 for a survey) show that there is a tendency to build radical consensus in warm groups (in which the discussion is free) and a trade-off average consensus in cold groups (with more strict protocols of discussion). S. Galam and S. Moscovici proposed first models of these phenomena, based on variations on the Ising model (Galam and Moscovici, 1991a; 1991b). We discuss some difficulties of interpretation of the Ising model in this context, especially about the extreme individual opinion in a binary state model. The relative agreement (RA) model (Deffuant et al., 2002) seems to fit much closely the considered social-psychology experiments: the opinion states are continuous, and can therefore be more or less extreme, and we showed in previous papers that the dynamics of the model can lead in some conditions to unexpected radicalisation.

However, a closer analysis reveals some inadequacies between the RA model and the results of the psycho-sociology experiments. In particular, the requirement to reach a consensus, which seems important in the experiments, is not considered in the RA model either. Moreover, it can happen in the experiments that all participants get more extreme, which never happens in the RA model.

Therefore, we propose a new model of the opinion influence process through group discussions, which takes these aspects into account. This evolution is based on two main hypotheses:

- The impact of the expressed opinions is related to their potential as a possible consensus, which yields a particular power to each participant to reject some opinions.
- Participants express opinions in a subset of their uncertainty segment (modification of the RA model in this way has yet been suggested by Urbig (2003)).

These new hypotheses allow us to distinguish between warm and cold groups. We suppose that in warm groups, the participants can express more strongly their disagreements than in cold groups. Moreover, they can express opinions in a larger part of their uncertainty zone. On the contrary, in cold groups, they keep expressing opinions close to their average uncertainty, and they do not express so clearly their disagreements. We propose some practical models derived from these hypotheses, and explore their behaviour by the mean of simulations.

## Radicalisation or trade-off in warm or cold groups

Moscovici and Doise (1992) describe a large set of psycho-sociological experiments in which they observe a "radicalisation" of opinions as an effect of group discussions. Radicalisation means here that the initial individual opinions (asked to the subjects before the discussion), become more extreme in the consensus. This effect is observed when the subjects are asked to reach a consensus after the discussion, and if the protocol of discussion is free. The experiment was reproduced in many various conditions since the first one in 1969. As Moscovici and Doise point out, it is rather surprising, because one would expect a consensus to be a compromise between the initial individual opinions, i.e. an average of the initial opinions. This expectation is fully contradicted by the observations. In fact, other experiments showed that an averaging consensus takes place when the subjects must comply to a constrained protocol of discussion (hierarchical for instance).

The opinion radicalisation is therefore very sensitive to the context of the discussion. This led to distinguish between "warm" and "cold" groups. Warm groups favour free expression of opinions and confrontation, and generally lead to new and radical group consensus whereas cold groups regulate the expression and tend to lead to trade-off average group consensus. Moscovici and Doise underline that it is not possible to give any absolute value judgement about the one or the other process. Radicalisation can lead to violent group behaviours, but also to more imaginative and creative solutions to a problem. Average solutions are in general more careful but also more stereotyped and can lead to mistakes (counter-productive solutions) because of a too strong pressure to conformity. The fascinating result of these experiments is that the orientation of the final consensus can be strongly controlled by the discussion procedure.

These social psychology observations offer a very interesting challenge to the social modelling community. What model could explain the radicalisation in warm groups and the trade-off in cold ones ? Could this model suggest new directions of experiments to social-psychology ?

## Ising modelling by Galam

Galam and Moscovici collaborated in a series of papers (Galam and Moscovici, 1991a; 1991b), in which they consider the Ising model from statistical mechanics to describe the interactions between participants of a discussion. In the Ising model, the participants have a binary state of opinion, and an individual bias which induces more or less facility to change their state under the influence of others. The dynamics of the model is based on an updating rule, in which each individual sums up the state of its neighbors, multiplied by a weight. Often, the state is a Gibbs function of this weighted average.

The standard behavior of these models, especially when the participants are totally connected (every one is related to all the others), is to converge toward a consensus on one or the other binary state (+1 or –1). However, with some distributions of individual biases, some participants may change of state less easily, or even always remain in their initial state. When such individuals are distributed in the population, the final result of the discussions can be heterogeneous: some participants having the state 1 and the others the state –1.

Galam and Moscovici interpret the final consensus state of the group as the average of the individual states. With such an interpretation, the model can lead to radical consensus (when every body agrees on one or the other binary state), or trade-off consensus, when different states are found in the population after the convergence. In this perspective, the Ising model can be used to model the experiments of psycho-sociology.

However, in our view, it is not very natural to use binary state individuals to model these social phenomena. In the psycho-sociological experiments, the individuals are asked an opinion which has 7 modalities (from very against to very for). The evolution of individual opinions during the discussions is measured on this scale. The variable playing the role of this individual opinion is not obvious in the Ising model. If the opinion of an individual is considered to be the average of all states in the population at the end of the simulations (as proposed by Galam), then it should be the same in the

previous time steps. But in this case, the population is always in a consensus, which does not fit the experimental data.

It seems therefore interesting to consider continuous opinion models, and explore how they could model the considered experiments.

## The Relative Agreement model

In (Deffuant et al., 2002), we proposed an agent-based simulation model of opinion dynamics, the relative agreement model (RA model). Its main characteristics are that agents' opinions are continuous rather than discrete (e.g. binary) in classical models; that the opinion dynamics of an agent takes into account opinions from others in a limited zone, the agent's uncertainty, around its own opinion as for instance in (Deffuant et al. 2001, Hegselmann and Krause, 2002). Moreover, the influence takes into account the overlap between the two opinions segments and each other's uncertainty. A property of this model is that the less uncertain is the agent, the more convincing or influential it is.

We studied the behaviour of this model when we introduce extremist agents into the population : these agents have opinions located at the extremes of the opinion axis (+1 or –1), and they have a lower uncertainty. The other agents, with opinion distributed uniformly on [-1,+1] have a higher uncertainty, and are called "moderate agents". We observe then three types of convergence, which mainly depend on the initial uncertainty of the moderate agents:

- Central convergence: The moderate agents evolve and form central cluster, being only marginally influenced by the extremists;
- Double extreme convergence: The moderate agents split into two clusters, each one converging to one of the extreme;
- Single extreme convergence: Almost all the moderate agents are attracted by one of the extremes.

We studied this model with the assumption that every pair of agents could interact (Deffuant et al., 2002), and in (Amblard and Deffuant, 2004) we studied the influence of different social network topologies on the model behaviour, especially the single extreme convergence.

This model seems particularly appropriate for modelling Moscovici and Doise findings. First, considering a continuous opinion gives directly the possibility to consider the opinion scales used in the experiments. Moreover, the uncertainty variable finds a natural interpretation, because it is also a measure of conviction (ability to convince and to be convinced). Considering extremists as very convinced agents, with extreme opinions fits very closely the observations of Moscovici and Doise (they note that in general, people with extreme opinions are more convinced).

Moreover, the RA model can easily represent the radicalisation in warm groups and the trade-off in cold ones. Suppose that the groups include more convinced individuals, which are closer to one extreme and that the others are moderate. Suppose also that the more convinced individuals tend to be located at only one of the extremes, because of a fundamental tendency of the society (this hypothesis is explicitly made by Moscovici and Doise), then the behaviour of the RA model is:

- convergence to one extreme (radicalisation) if the extremists are very convinced and the moderate very uncertain.
- central convergence (trade-off) if the difference of uncertainty between extremists and moderate is not too high, and all are uncertain.
- A set of opinion clusters if all are certain of their opinions.

A possibility of interpretation is therefore that :

- in warm groups the individuals more freely express their certainties or uncertainties which results in a perceived higher difference between extremists and moderates. This difference leads to a radicalisation in the RA model.

- On the contrary, in cold groups the expression of certainty or uncertainties is smoothed which tends to lead to trade-off consensus.

Note that the RA model predicts that the convergence to a single extreme is taking place even when there is a similar number of extremists of both sides in the population. In the psycho-sociological experiments, there is always a strong asymmetry of the initial opinions, which indicates the side of radicalisation. Moreover, the RA model, as it exhibits other behaviours, can be considered as a generic model applicable to other cases.

The interpretation of the difference between warm and cold groups seems reasonable in the RA perspective. One can expect that a formal protocol of discussions will limit the expression of opinions, and of very large uncertainties as well as very strong certainties, which on the contrary will freely be expressed in an unconstrained context.

However, a closer analysis shows some weaknesses and inadequacies of the RA model, which invite to design a more refined model. The criticisms are the following:

- the fact that a consensus is required is not taken into account. However, from the conducted experiments, this requirement seems to play a particularly important role in the result.
- Some observations show that all the participants can radicalise during the discussion. This can never happen with the RA model, at most the moderate finally get the opinion of the extremists, but the extremists never get more extreme. An extension of the model currently under study enables such a behaviour (Jager and Amblard, 2004).

## A refined model: the power to say no

We try to design a refined model, inspired from the bounded confidence model family (Hegselmann and Krause, 2002; Deffuant et al., 2001; 2002; Weisbuch et al., 2002), which remains as simple as possible while taking into account deeper aspects of the psycho-sociological observations.

To take into account the requirement of a consensus at the end of the discussion, we propose to includes the following features into the model:

- The influence of an expressed opinion depends on how it is globally appreciated by the group, and the possibility of the opinion to gather a consensus is considered in this appreciation,
- The appreciation of an opinion (in particular its quality as a potential consensus) is publicly discussed, because the consensus is an objective of the discussion,
- The opinions which are strongly disapproved by at least one of the participants are publicly marked negatively, because they are not likely to lead to a consensus,
- In warm groups, the disagreements are expressed more openly and frankly than in cold groups,

To take into account the possibility that all participants radicalise, we propose to include the following features into the model:

- The agents can express opinions chosen at random in a zone located in the middle of their uncertainty segment. In warm groups this zone is larger than in cold groups.
- When they express an opinion which is disapproved by one other participants, the agents increase their uncertainty,
- When they express an opinion which is approved by the other participants and is influential, the agents tend to adopt this opinion and to decrease their uncertainty around it.

More precisely, in the refined model, the agents are still defined by an opinion $x$ and an uncertainty $u$, which evolve during the interaction. We add a parameter $T$, which represents the "temperature" of the group, and rules the width of the opinion expression segment as well as the strength of the disagreement expression. We suppose that $T$ varies from 0 (very cold groups) to 1 (very warm groups). When there is no disagreement (the opinion is located in the segment of all participants), the

interaction rule has similarities with the one proposed in (Weisbuch et al., 2002): the uncertainty systematically decreases. An interaction cycle consists of:

---

For each agent *i*:

Let $x$ be an opinion chosen at random in $[x_i - T u_i, x_i + T u_i]$,

Let $d$ be the largest distance from $x$ to the other opinion segments,

If $d > 0$, then for all opinions segments $j$ containing $x$,

$u_j := u_j.(1+\mu.Td)$

If $d = 0$, then let $d'$ be the largest distance the other opinions $x_j$ (measure of influence), then all the opinions and uncertainties are modified as follows:

$x_j := x_j.+ \mu.d'(x - x_j)$

$u_j := u_j.(1 - \mu.Td')$

Where $\mu$ is a parameter of the model controlling the speed of conforming.

---

The expected behaviour of this model is:

- to give an even higher importance to the more convinced agents located at the extreme, because they will strongly disapprove the moderates. This disapproval will increase the uncertainty of the moderates. The disapproval of the extremists by the moderates will be weaker, because the uncertainty of the latter is larger, which implies a smaller $d$.
- the extreme opinions (higher $d'$) have an advantage, which can result in a radicalisation of even the extremists.

In the full paper, we shall propose an exploration of this model, and an analysis of its adequacy to the psycho sociological phenomenon.

## Discussion - Perspectives

The set of psychological experiments presented by Moscovici and Doise (1992) offers a very interesting challenge to agent-based modellers. Existing models, in particular the RA model, seem to capture important aspects of the observations. However, a closer analysis reveals that some subtleties are ignored by these models.

Refining these models led us to make new hypotheses, in particular the fact that disapproving an expressed opinion has a determinant influence in the process. In this perspective, the distinction between cold and warm groups relies mainly on the openness of disagreement expressions. This hypothesis could be tested in particular psycho-sociological experiments.

## References


Amblard and Deffuant, 2004, "The role of network topology on extremism propagation with the Relative Agreement opinion dynamics" accepted for publication by *Physica A*, (cond-mat/0404574).

Deffuant, G., Neau, D., Amblard, F. and Weisbuch, G., "Mixing beliefs among interacting agents", *Advances in Complex Systems*, vol.3, pp.87-98.

Deffuant, G., Amblard, F., Weisbuch, G. and Faure, T. 2002. How can extremism prevail? A study based on the relative agreement interaction model, *Journal of Artificial Societies and Social Simulation*, vol.5, 4 http://jasss.soc.surrey.ac.uk/5/4/1.html

Galam, S. and Moscovici, S., 1991a, Towards a theory of collective phenomena: consensus and attitude changes in groups. European Journal of Social Psychology. 21 49-74.

Galam, S and Moscovici, S. , 1991b, Compromise versus polarization in group decision making. Defense Decision making, Russ. Psychol. J. 13. 93-103.

Hegselmann, R. and Krause, U. 2002. Opinion Dynamics and Bounded Confidence Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation*, vol.5, 3. http://jasss.soc.surrey.ac.uk/5/3/2.html

Jager, W. and Amblard, F., 2004, «A dynamical perspective on attitude change », accepted to NAACSOS (North American Association for Computational Social and Organizational Science) Conference, Pittsburgh, 2004.

Lécuyer, R., 1974, *Rapports entre l'homme et l'espace,* PhD dissertation, Laboratoire de Psychologie, Paris.

Moscovici, S. and Lécuyer, R., 1972. «Studies on polarzation of judgments », *eruopean Journal of Social Psychology,* vol.2, pp.221-244.

Moscovici, S. and Doise, W., 1992, "dissension et consensus", Seuil.

Urbig, D., 2003. "Attitude dynamics with limited verbalisation capabilities", *Journal of Artificial Societies and Social Simulation,* vol.6, n°1, http://jasss.soc.surrey.ac.uk/6/1/2.html

Watts, D. 1999. Small Worlds. The Dynamics of Networks between order and randomness. Princeton University Press. Princeton.

Weisbuch, G. Deffuant G., Amblard, F., Nadal J.P., 2002, „Meet, Discuss and segregate!", Complexity, vol 7, issue 3, 55-63.