Trust in Communication between Individuals: A Connectionist Approach

Frank Van Overwalle (frank.vanoverwalle@vub.ac.be)
Francis Heylighen (fheyligh@vub.ac.be)
Margeret Heath (mheath@vub.ac.be)
Department of Psychology
Vrije Universiteit Brussel, Belgium

Abstract

How is social information transmitted in a group? Several studies in social cognition documented that communication about groups typically tends to bolster stereotypes and shared beliefs about these groups (Brauer, Judd & Jacquelin, 2001; Klein, Jacobs, Gemoets, Licata & Lambert, 2003; Lyons & Kashima, 2003). A multi-agent connectionist model is proposed that is capable of simulating these stereotype confirmation biases in group communication, as well as the effects of some moderating conditions. The model combines features of standard recurrent models to simulate the process of information uptake, integration and memorization within agents with novel aspects that simulate the communication of beliefs and opinions between agents. A crucial aspect in belief updating from other agents is trust in the information provided. By studying these novel communicative aspects within the framework of standard models of information processing, the unique communicative mechanisms underlying the emergence of a confirmation bias in groups beyond intra-personal factors can be explored.

A Connectionist Approach

There are several characteristics that make connectionist approaches very attractive in comparison with earlier information processing models in social psychology (for an introduction, see McLeod, Plunkett & Rolls, 1998). First, a key difference is that the connectionist architecture and processing mechanisms are based on analogies with properties of the human brain. This allows a view of the mind as encompassing adaptive learning mechanisms that develop accurate mental representations of the world. Learning is modeled as a process of on-line adaptation of existing knowledge to novel information provided by the environment. For instance, in group judgments, the network changes the weights of the connections between the target group and its attributes so as to better represent the accumulated history of co-occurrences between the group and its attributes. Most traditional algebraic and activation-spreading models in social psychology are incapable of learning.

Second, connectionist models assume that the development of internal representations and the processing of these representations occurs in parallel by simple and highly interconnected units, contrary to traditional models where the processing is inherently sequential. As a result, these systems do not need a central executive, which eliminates the requirement of centralized and deliberative processing of information. This suggests that much of the information processing within agents is often implicit and automatic. This does not, of course, preclude people from becoming aware of the outcome or end result of these preconscious processes.

Third, and perhaps more crucially in the present context, based on the principle that activation in a network spreads automatically to interconnected units and concepts and so influences their processing, connectionist models exhibit emergent properties such as pattern completion and generalization, which are potentially useful mechanisms for an account of the confirmation bias of stereotype information dissemination within and between agents in a group.

A Recurrent Model

In this paper, we apply and extend the recurrent auto-associator model developed by McClelland and Rumelhart (1985). Let us first focus on the standard recurrent model. This model has already been applied in social psychology to study, for instance, person and group impression formation (Smith & DeCoster, 1998; Van Rooy et al., 2003; Van Overwalle & Labiouse, 2004), attitude formation and change (Van Overwalle & Siebler, 2005), and causal attribution (Read & Montoya, 1999). We apply this model here to emphasize the theoretical similarities that underlie these diverse social phenomena with the present findings of confirmation bias in collective judgments and stereotypes emerging during group communication.

A recurrent network can be distinguished from other connectionist models on the basis of its (a) architecture (how information is represented in the model), the (b) manner in which information is processed and (c) its learning algorithm (how information is consolidated in the model).

(a) In a recurrent architecture, all units within an agent are interconnected with all of the other units of the agent. Thus, all units send out and receive activation. In the present context, the units in the network represent a target group and its various attributes.

(b) Information is represented by *external activation*, which is automatically spread among all interconnected units within an agent in proportion to the weights of their interconnections. The activation coming from the other units within an agent is called the *internal activation*. Typically, activations and weights have lower and upper bounds of approximately –1 and +1.

(c) The short-term activations are stored in long-term *weight changes* of the connections. Basically, these weight changes are driven by the difference between the internal activation received from other units in the network and the external activation received from outside sources. This difference, also called the "error", is reduced in proportion to the learning rate which determines how fast the network changes its weights and learns. This error reducing mechanism is known as the *delta algorithm* (McClelland & Rumelhart, 1985; McLeod, Plunkett & Rolls, 1998).

TRUST: An Extended Recurrent Model of Communication

The standard recurrent model was augmented with a number of features, which enabled it to realistically reproduce communication between agents. This extension assumes that information about persons and groups and their attributes is represented in broadly the same manner among different agents. Communication is then basically seen as transferring the activation on person and group attributes expressed by *talking* agents to *listening* agents. This is accomplished by activation spreading between agents in much the same way as activation spreading within the mind of a single agent, with the restriction that activation spreading between agents is (a) limited to identical attributes and (b) in proportion to the connection weights linking the attributes between agents. A crucial aspect of this between-agents dissemination of information is *trust*, or the degree to which the information on a given attribute or concept by a given agent is deemed reliable and valid. The connection weight held by agents on the same concept reflects this degree of trust, and is therefore the cornerstone of the extended recurrent model. We therefore termed the extended model TRUST.

Maxims of Quality and Quantity

Because agents can play the role of speaker or listener, the trust connections in the model go in two directions for each agent: Sending connections for a speaking agent and receiving connections for a listening agent. These two trust connections implement Grice's (1975) maxims of *quality* and *quantity* of communication.

First, the maxim of quality suggests that in order to communicate efficiently, communicators generally try to transmit truthful information. In the model, this maxim of quality is implemented on the side of the receiving agent. Communication is more efficient if the information is believed to be trustworthy. This is implemented in the trust connection from an agent expressing his or her ideas to the receiving agent. When trust is maximal (+1), the information expressed by the talking agent is unattenuated by the listening agent. To the degree that trust is lower, information processing by the listener is attenuated in proportion to the trust weight. When trust is minimal (0), no information is processed by the listening agent. This mechanism is schematically represented in Figure 1.

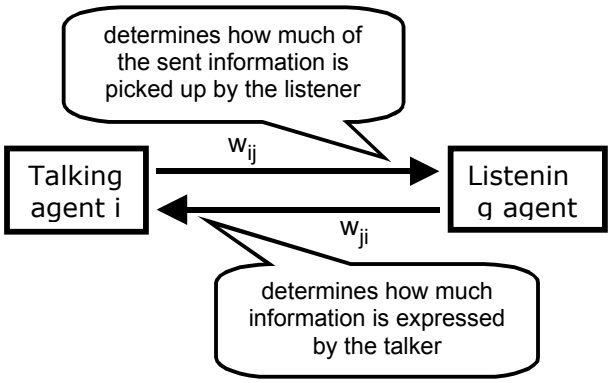

Figure 1: The role of trust weights in communication between agents.

Thus, the listener sums all information received from other talking agents in proportion to the trust weights, and then processes this information internally (according to the standard recurrent approach). Or, in mathematical terms:

$$ext\_a_j = \sum_i (w_{ij} * a_i)$$

where ext_a_j represents the external activation received by the listening agent j or the degree to which the external information is ready for processing; w_{ij} is the trust weight from the talking agent i to the listening agent j; and a_i denotes the final activation (which combines the external and internal activation received) expressed by the talking agent i. Given that this mechanism of trust spreading between agents is a straightforward extension of activation spreading of connectionist models within a single agent, this suggests that, except for the transmission of information by speech, the acceptance of the information by the listener is most probably a relatively automatic process.

Second, Grice's (1975) maxim of quantity suggests that communicators transmit only information that is informative and adds to the audience's knowledge. In addition, research on group minority suggests that communicators tend to increase their interaction with an audience that does not agree with one's position. This is implemented in the model by the trust weights from the listening agent to the talking agent. These weights indicate the degree of trust by the talking agent in the listening agent, and are determined by earlier communications in which the listening agent expressed judgments on an issue that were largely (in)congruent with the talking agent's ideas. To the extent that these trust weights are high (above trust starting weights), knowledge and agreement on an issue is assumed and the talking agent will restrain him- or herself from expressing these ideas further. In contrast, when these weights are low (below trust starting weights), the talking agent tends to express and defend his or her ideas on this issue more strongly. In mathematical terms:

if ($\max(w_{ij})$ < trust starting weight)
then $a_i = a_i * [1 + \max(w_{ij})]$
else $a_i = a_i * [1 - \max(w_{ij})]$

where a_i is the final activation expressed by the talking agent i, and $\max(w_{ij})$ represents the maximum trust weight from all listening agents j to a talking agent i. In contrast, for all other issues k agent i is talking about, the reverse change of activation occurs:

if ($\max(w_{ij})$ < trust starting weight)
then $a_k = a_k * [1 - \max(w_{ij})]$
else $a_k = a_k * [1 + \max(w_{ij})]$

Because this boosting and attenuation of activation (or expression of information by a talking agent) is a rather novel idea, it is unclear whether this is either a largely automatic process or a more controlled strategy used by the speaker.

Adjustment of Trust Weights

Given that perceived trust plays a crucial role in the transmission of information, it is important to describe how trust is developed and changed in the model. Like the standard delta learning algorithm which is used to adjust memory traces within individual agents, the degree of trust depends on the error between external beliefs expressed by a talking agent and a listening agent's own internal beliefs. If the error is below some *trust tolerance*, the trust weight between the concepts held by the two agents is increased towards 1; otherwise, the trust weight is decreased towards 0. In mathematical terms, trust weight change between agents or Δw is implemented as follows:

if $|ext\_a_j - int\_a_j| <$ trust tolerance
then $\Delta w_{ij} =$ trust rate $* (1 - w_{ij}) * | a_i |$
else $\Delta w_{ij} =$ trust rate $* (0 - w_{ij}) * | a_i |$,

where ext_a_j represents the external activation received (from the talking agent i) by the listening agent j and int_a_j the internal activation generated independently by the listening agent j; and trust rate is the rate by which trust is adjusted.

Because the trust change mechanism is a straightforward extension of the basic delta learning algorithm in that it is also error-driven and attempts to reduce the error between the listening agent's internal representation and external information, we assume that this mechanism is largely automatic.

Summing up, the larger w_{ij} becomes, the more the listening agent j will trust the talking agent i on the issues communicated, and the more influential the talking agent will become (maxim of quality). In turn, this will restrain the just-listening agent in expressing his or her ideas on this issue (maxim of quantity). Note that when a listening agent's own beliefs are changed as a result of the feedback from some agents, this will have an effect on the listener's own internal activation (int_a_j) and so on his or her perceived trustworthiness of all other agents.

Group Communication: A Case Study

Maxim of Quality

Experiment. The maxim of quality suggests that communication is more efficient when the information is trustworthy. To illustrate the working of the maxim of quality as implemented in the trust weight from the communicating agent to the receiving agent, we will now apply the TRUST model to an empirical study undertaken by Lyons and Kashima (2003, Experiment 1). In this study, information was communicated through a serial chain of 4 people. In this paradigm, one person begins to read a set of information before reproducing it from memory to another person. This second person then reads this reproduction before then reporting it verbally to a third person and so on. The information in the study involved a story depicting a member of a fictional group of Jamayans. Before disseminating the story along the chain, general stereotypes were induced about this group. In one of the conditions, all 4 participants in the chain were given the same stereotypes about the Jamayans (*actual shared condition*). In another condition, 2 participants were given stereotypes about the Jamayans that were opposite to that given to the other 2 participants, so that each subsequent participant in the chain held opposing group stereotypes (*actual unshared condition*). The story given afterwards always contained mixed information that both confirmed and disconfirmed the stereotype.

As can be seen in Figure 2, when the stereotypes were shared, the reproduction became more stereotypical further along the communication chain (see left side). The story was almost stripped of stereotype inconsistent (SI) information, whereas most of the stereotype consistent (SC) information had been retained. In contrast, when the stereotypes were not shared (see right side), the differences between SC and SI story elements were minimal.

Simulation. We simulated a simplified version of the original experimental procedures by Lyons and Kashima (2003). As can be seen in Table 1 (panel 1a), for the actual shared condition, we provided 10 stereotypical trials indicating that the Jamayans were smart (i.e., by activating the Jamayans and the smart unit), and 10 stereotypical trials indicating that they were honest (i.e., by activating the Jamayans and the honest unit) for each of the 4 agents. For the actual unshared condition, two of the agents received contradictory information indicating that the Jamayans were stupid and dishonest (i.e., by activating the stupid and liar units; panel 1b). Next, the first agent received 5 SC trials reflecting story elements indicating that the member of the Jamayans was smart and 5 SI story elements indicating that this member was a liar (panel 2). This story was reproduced

by this agent and received by the next agent. That is, the Jamayans unit in agent 1 was activated and, together with the resulting activation of the other smart/stupid and honest/liar units in agent 1 (panel 3a), was then transmitted to agent 2 (panel 3b), and so on along the chain. After each talking phase, we measured how much the talking agent had expressed or communicated the notion that the Jamayans were smart, stupid, honest or liar (i.e., by averaging the activation of the relevant "i" units during that phase).

Table 1: Simplified Simulated Learning History on the Jamayans Story (Lyons & Kashima, 2003, Exp. 1)

	Jamayans	Smart	Stupid	Honest	Liar
1a. SC Information on Jamayans ***(all Agents)***					
# 10	1	1			
# 10	1			1	
1b. SI Information on Jamayans ***(all Agents)***					
# 10	1		1		
# 10	1				1
2. Mixed Story given to ***Agent 1 only***					
# 5	1	1			
# 5	1				1
3a. Reproducing Story by ***Agent 1, 2, and 3*** [a]					
# 5	1	i	i		
# 5	1			i	i
3b. Listening by ***Agent 2, 3, and 4*** [b]					
# 5	?	?	?		
# 5	?			?	?

Note. Cell entries represent external activation and empty cells reflect 0 activation. SC=Stereotype Consistent, SI=Stereotype Inconsistent, #=Number of trials, based on the actual study. Each experimental condition (shared versus unshared) was run separately, and always preceded by the SC Information in Phase 1a (in the shared condition) or by the SC and SI Information in Phase 1b (in the unshared condition), followed by the Mixed Story Phase 2, Reproducing and Listening Phase 3 (together for talking and listening agents 1-2, 2-3, and 3-4 respectively). Trial order was randomized in each phase and condition for 50 runs, and the results were averaged. The results in Figure 2 represent the mean internal activation during the talking phase (phase 3a)

[a] i = internal activation (generated after activating the Jamayans unit) is taken as external activation; [b] ? = internal activation transmitted to the listening agent.

We ran 50 simulation runs (each with a different random trial order), and averaged the results. The parameters were intra-individual learning rate = 0.30, and inter-individual trust rate = 0.40, trust tolerance = 0.50 and trust starting weight = .40. The other standard recurrent parameters were the same as in earlier simulations by Van Overwalle and colleagues (E = I = Decay = number of internal Cycles = 1, intra-individual starting weights = 0, and a linear summation of int_a and ext_a; see also Van Overwalle & Labouise, 2004; Van Overwalle & Siebler, 2004).

Results and Discussion. As can be seen in Figure 2, the simulation closely matched the observed data (r = .94, p < .001). Given that the story was told in a predetermined order along the communication chain, boosting or attenuation of belief expression (maxim of quantity) did not play a role here as the agents had no opportunity to hear their communication partners before being told the story, and so were unable to test whether they agreed on the Jamayans' attributes. This strongly suggests that for these experimental results, only the trust in a talking agent's statements (maxim of quality) was sufficient for creating a stereotype confirmation bias during group communication.

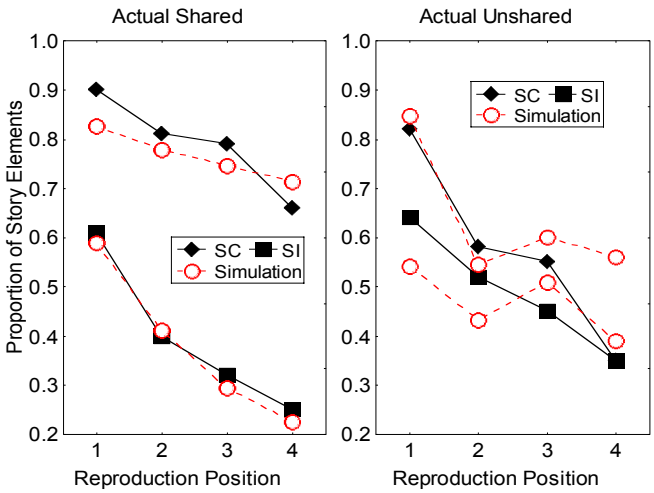

Figure 2: Mean proportion of stereotype-consistent (SC) and stereotype inconsistent (SI) story elements in the actual shared and unshared conditions, and simulated values from the TRUST model (projected on the observed data by linear regression). The human data are from Figure 2 in Lyons and Kashima (2003, p. 995), averaged across central and peripheral story elements.

Maxim of Quantity

Experiment. The maxim of quantity suggests that when the audience is knowledgeable and agrees with the communicator's position, no information is transmitted. In the model, the maxim of quantity is implemented by the trust weight from the listening agent to the talking agent. A high weight indicates that the listener is to be trusted and that expression of the same information can be attenuated, while the expression of other information can be boosted. To illustrate the working of the maxim of quantity, we will now apply the TRUST model to another data set from the same empirical study by Lyons and Kashima (2003), described earlier. In this study, Lyons and Kashima provided half of their participants with the false information that the other participants in the chain had received completely similar general information on the Jamayans (*perceived complete knowledge*), while the other half were given the false information that the other participants were completely ignorant (*perceived complete ignorance*).

Figure 3 depicts the results. It was found that given the belief of complete knowledge, both SC and SI story elements were reproduced and no substantial stereotype bias emerged. In contrast, in the complete ignorance condition, a stereotype bias became apparent in that SI story elements were strongly suppressed.

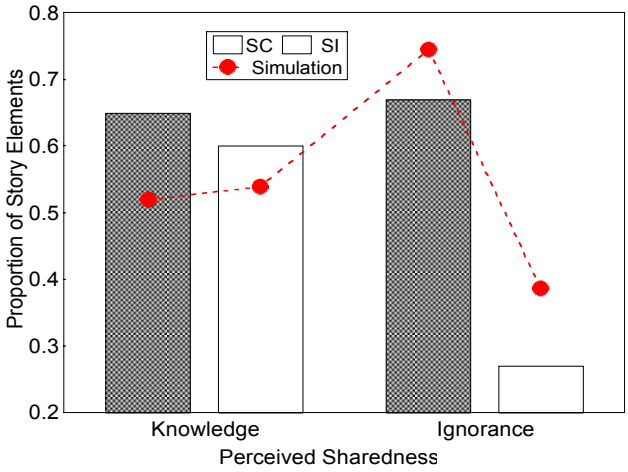

Figure 3: Mean proportion of stereotype-consistent (SC) and stereotype inconsistent (SI) story elements in the perceived complete knowledge and perceived complete ignorance conditions, and simulated values from the TRUST model (projected on the observed data by linear regression). The human data are from Figure 1 in Lyons and Kashima (2003, p. 995).

Simulation. We ran the same simulation as before, with the following modification. In order to obtain high trust weights from the listening agents to the talking agents, (a) we included only the actual shared condition, and (b) we set the initial trust weights from listening to talking agents 0.20 above the trust starting weight for the units involved in the transmission of SC information (Jamayans, smart, honest). These high trust weights directly simulate that the listening agents were to be trusted because they largely agree with the speaker.

Results and Discussion. As can be seen in Figure 3, the simulation matched the observed data although not above conventional levels of significance (r = .80, p = .21) due to lack of data points (only 4). The inclusion of the maxim of quantity (which depends on the trust in the listener) in the simulated complete knowledge condition, suggests that it can counteract the maxim of quality (trust in the talker), and so neutralizes the stereotype confirmation bias in group communication.

Conclusion

The proposed TRUST connectionist model combines all elements of a standard recurrent model of impression formation with additional elements reflecting communication between individuals. Specifically, Grice's (1975) maxims of quality and quantity were implemented on the basis of the experienced trust in the other individuals' position on similar issues. This implementation seems adequate, as it was capable of replicating the main patterns in the observed data from a study by Lyons and Kashima (2003). In particular, it replicated the role of actual sharedness of information (maxim of quality) and perceived sharedness (maxim of quantity). Other simulations (not discussed here) which explored additional moderating factors of group communication and stereotyping were also successful, such as the communication of shared versus unique information (e.g., Larson, Christenson, Abbott & Franz, 1996).

Perhaps one of the major advantages of the model that makes this possible is its dynamic nature. It conceives communication as a coordinated process that transforms the beliefs of the agents as they communicate. Through these belief changes it has a memory of the social history of the interacting agents. Thus, communication is at the time a simple transmission of information about the internal state of the talking agent, as well as a coordination of existing opinions and emergence of novel beliefs on which the conversants converge.

An obvious limitation of the present model is that it largely ignores by what means information is communicated. In most social psychology experiments, this is simply accomplished by speech. How exactly the outputs of the semantic units in our connectionist system are transformed to speech by the talking agent, and how speech is again transformed to input for the connectionist system of the listening agent is left out of our model. At least at the moment, this seems a sensible simplification, since social communication may be driven by other acts than speech, such as non-verbal behavior or deaf sign language. However, it is an interesting avenue for further research and modeling.

Well-known phenomena such as "group think", mass hysteria, the spreading of false rumors, and the failure to consider all relevant information or possibilities point us to the danger that at least under some circumstances, the processes of communicating information among the members of a group seems to make their collective cognition and judgments less reliable. The present paper helps us to illuminate and tear apart some basic mechanism in the creation of group biases and misperceptions.

References

Brauer, M., Judd, C. M., & Jacquelin (2001). The communication of social stereotypes: The effects of group discussion and information distribution on stereotypic appraisals. *Journal of Personality and Social Psychology, 81*, 463—475.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Speech acts* (pp. 41—58). New York: Academic Press.

Klein, O. Jacobs, A., Gemoets, S. Licata, L. & Lambert, S. (2003). Hidden profiles and the consensualization of social stereotypes: how information distribution affects

stereotype content and sharedness. *European Journal of Social Psychology, 33*, 755—777.

Larson, J. R., Christensen, C., Abbott, A. S., & Franz, T. M. (1996). Diagnosing Groups: Charting the flow of information in medical decision-making teams. *Journal of Personality and Social Psychology, 71*, 315—330.

Lyons, A. & Kashima, Y. (2003) How Are Stereotypes Maintained Through Communication? The Influence of Stereotype Sharedness. *Journal of Personality and Social Psychology, 85*, 989-1005.

McClelland, J. L. & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology, 114*, 159—188.

McLeod, P., Plunkett, K. & Rolls, E. T. (1998). *Introduction to connectionist modeling of cognitive processes*. Oxford, UK: Oxford University Press.

Read, S. J., & Montoya, J. A. (1999). An autoassociative model of causal reasoning and causal learning: Reply to Van Overwalle's critique of Read and Marcus-Newhall (1993). *Journal of Personality and Social Psychology, 76*, 728—742.

Smith, E. R. & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology, 74*, 21—35.

Van Overwalle, F. & Siebler, F. (2005). A Connectionist Model of Attitude Formation and Change. *Personality and Social Psychology Review*, in press.

Van Overwalle, F., & Labiouse, C. (2004) A recurrent connectionist model of person impression formation. *Personality and Social Psychology Review, 8*, 28—61.

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C. & French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review, 110*, 536-563.