Towards Explaining Social and Cognitive Bias

Bruce Edmonds,
Centre for Policy Modelling
Manchester Metropolitan Univeristy

Motivation

Human cognitive 'biases' have generally been conceived of as consistent and persistent deviations from an ideal rationality. Sometimes a ground truth is established by design and then various conditions tested to see what causes subjects to deviate from this – e.g. where subjects are asked to judge which line is longer where others apparently judge wrongly (e.g. Asch 1956). Sometimes it is simply assumed that belief is purely a process of social influence, as in many simulation models – e.g. opinion dynamics models (Deffuant et al. 2002). Sometimes a simple, built-in bias is built into the agents in a simulation – e.g. (Xu et al 2014).

However, I have the feeling that these all miss an important point, namely *how to explain the patterns of belief in socially embedded humans* in a way that makes sense, both from the point of view of the individuals (i.e. bias is not just a matter of being 'pushed' to believe other than the truth by one's peers) and from the point of view of the society they form. This abstract seeks to take tentative steps towards explaining the patterns of such socially embedded belief based on three principles: the preference of individuals for internally coherent belief structures; that the social structure will tend to avoid longer-term connections where individuals disagree; and that the menu of possible beliefs is largely determined by what one's peers believe. I will discuss each briefly in turn.

The *first* base assumption here derives from Thagard's theory of explanatory coherence (Thagard 1989). This explains whether an individual holds a particular belief or not in terms of in terms of its coherence or incoherence with other, already held, beliefs. Under this model, individuals will preferentially accept new beliefs that are coherent with their existing belief structure and tend to reject those that are not coherent with these. In the model presented here, this is extended to go beyond pairwise coherence to the coherence of whole sets of beliefs.

Secondly, the model assumes that people will tend to interact with those with similar, or at least not incompatible, beliefs as themselves. For example (McPherson et al 2001) show that shared beliefs are associated with a higher chance of interaction. Here we use a weak version of this hypothesis, namely that we have a tendency not to interact with those we disagree with.

The *third* major assumption is that most of an individual's beliefs were suggested by someone else. That is only rarely does an individual invent a totally new belief that is not held by others. ??

The Model

I now describe a model that reflects these principles to illustrate the potential of combining them. The model to be described is quite abstract at the moment, *much* simpler than my usual style of model! It is merely a starting point, to point out that the intimate connection between the cognitive and the social can be represented and to stimulate discussion. I am looking for suitable data to enable its assessment and further development, and would welcome the suggestion of any rich data sets or qualitative research that I might compare this to. Other models are available.

In this model:

- There is a network of a fixed set of nodes and arcs (that can change)
- There are, *n*, different atomic beliefs {A, B, ...} circulating between nodes
- Beliefs are copied along links or dropped by nodes according to the change in coherency of the node's belief set that this would result in??
- Links can be randomly made
- Links are dropped when beliefs are rejected for copy between nodes

Node properties

Each node has:

- A (possibly empty) set of the "atomic beliefs" that it holds
- A fixed "coherency" function from possible sets of beliefs to [-1, 1] where 1 is completely coherent, 0 is neutral and -1 is maximum incoherency.

- A fixed scaling function that maps changes in coherency to the probability of a change in beliefs
- A record of the last node it "rejected" a belief from

Initialisation
Beliefs and social structure are randomly initialized at the start according to some global parameters. In the present version there can be up to 3 types of agent, which are distinguished by their coherency and scaling functions.

Coherency function
Key to this model is the model of belief coherency, which is a generalisation of Thagard's pairwise (in)coherence. It gives a measure of the extent to which whole set of current beliefs are coherent. This assumes a background of shared beliefs which are not represented – this is important as the model only captures what might happen to a few foreground beliefs that are changing against all other beliefs. If one chose a different set of 'foreground' candidate beliefs from all possible beliefs then different coherency functions would be needed.

This allows for great flexibility in choices of belief structure, for example we could have the coherency evaluations: {A}→0.5 and {B}→{0.7} but also {A, B}→-0.4 if beliefs A and B are mutually inconsistent, but individually coherent (against the background beliefs). Here the coherency function is set by the programmer for each kind of agent. The probability of gaining a new belief from another or dropping an existing belief in this model is monotonically dependent on whether it increases or decreases the coherency of the node's belief set

Belief change processes
There are basically two belief change processes. Each iteration the following occurs:

- *Copying*: each arc is selected; a source end and destination end selected; a belief at the source is randomly selected; then copied to the destination with a *probability* related to the change in coherency it would cause (due to the scaling function described next).
- *Dropping*: each node is selected; a random belief is selected and then dropped with a *probability* related to the change in coherency it would cause

Scaling the impact of coherency function
There is a variety of ways to map a change in coherence to a probability (of a change occurring). The function that maps from changes in coherency to probability could be any that: (a) is monotonic (b) such that a -1→1 change has probability of 1 (b) a 1→-1 change has probability of 0. Two example such functions are illustrated in Figure 1.

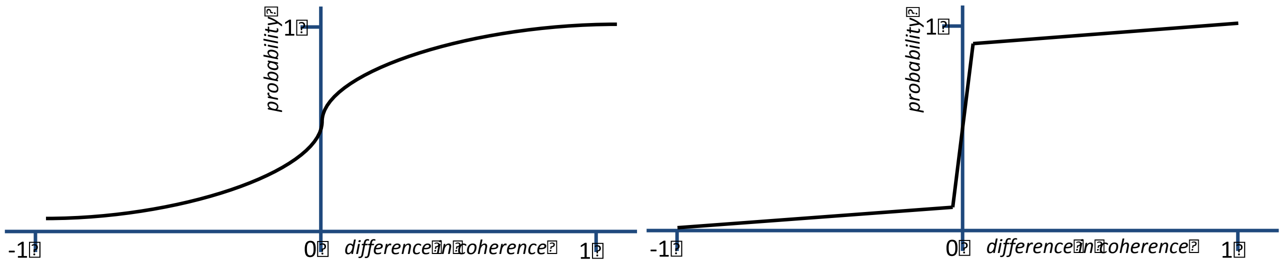

Figure 1. Two example mappings from a change in coherency to a probability (of either a "copy" or a "drop" of a belief): ***left***, A 'weak' mapping – probably changes to increase coherence, ***right***, A 'strong' mapping – almost certainly only changes to increase coherence.

The scaling function thus affects how amenable an agent is to change and the extent to which it may change. E.g. whether only to increase coherency or if it can occasionally decrease.

Network Change Processes
There are two processes for changing the influence network. Each iteration, for each agent:

- *Link Drop*: with a probability: if a belief copy was *rejected* by the recipient, then drop that in-link.
- *New Links*: with another probability, create a new random link with a random other (with a friend of a friend if possible, otherwise any)

Other
In order to maintain the average link density I added the following 'kludge': If there are too many links (as set by arcs-per-node) slightly increase the rate of link drop, if there are not enough, slightly reduce

the rate of link drop. Also, nodes have to have at least one link, or one is added, to stop isolates forming. Finally there is a small probability that a belief is randomly added or dropped, this adds a little bit of extrinsic noise into the system and stops beliefs disappearing (through chance) from the entire population, as discussed in (Edmonds 2012). This is a rare event, and exponentially rare with increasing population size, so the 'forgetting' of beliefs from the population only something that happens in relatively small and isolated populations.

The "opinion" of agents is derived from the belief state of the agents. This is a function from the belief set to [-1, 1]. The global opinion is an average of this function applied to each agent. There is obviously a choice as to how this is done, but we do this uniformly. More about the model, including the source code, can be found in (Edmonds 2016).

Some Results

The particular case explored here is a situation where there is a large neutral population, a substantial minority with a strong view on a particular issue, but where there might also be a small minority who have a fanatically held, but opposing view. One might think of the case of vaccination here, where the minority of scientists are declaring that vaccination is a good idea, but a small minority are convinced that, say, vaccination has very harmful side effects (e.g. the idea that it might trigger autism in a child). However, because I do not have any data about this, I concentrate on the 'Brexit' referendum which can be thought of as a similar kind of situation.

In this example, there are two beliefs (yellow and blue) that are broadly incompatible with each other. The agent population is composed of the following different 'kinds' of agent. 70% 'floaters', these are weakly positive towards having either yellow or blue beliefs, but not both. They have a weak scaling function so they are more open to change and more tolerant of temporarily tolerating moves to lower coherence. 20% are 'remainers' for blue and against yellow, with a medium scaling function. Finally 10% 'leavers' for yellow and against blue with a very strong scaling function. If you run this model 1000 times the final average opinion distribution looks like that in Figure 2 (left).

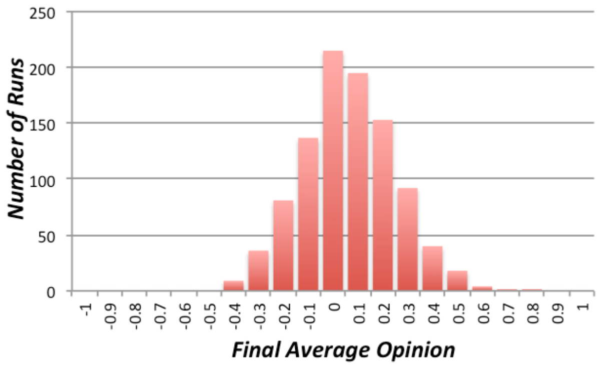

Figure 2. Histogram of final average opinions of 1000 independent runs of this example (tick 1000).

However, individual runs are quite unpredictable, being 'locked into' certain patterns of belief for periods of time (due to the developed social structure), but at other times changing apparently at random. This is illustrated in Figure 3 (right), where a dominant pro-yellow group formed, but eventually broke up.

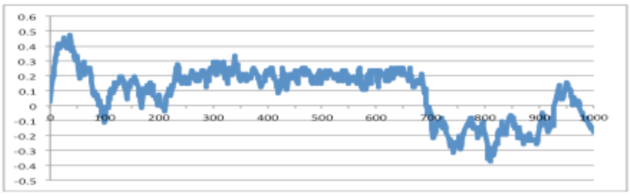

Figure 3. Av. Opinion in one Example Run

When one clusters the final state of runs in two dimensions (how different the beliefs of linked agents are from each other, and whether agents that are linked are of the same kind, one gets the pattern shown in Figure 4 (left). Here one sees two distinct clusters. The bigger cluster (green) is runs where more different kinds of agent are linked together but where beliefs within groups are relatively homogeneous. The smaller (red) is when different kinds tend not to be linked to each other, but there is a greater variety of beliefs between linked agents.

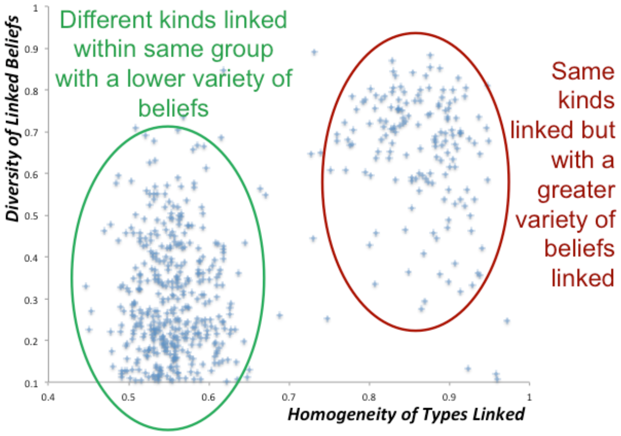

Figure 4. Clustering final state of 1000 runs

Informal observation of the social networks in such runs suggest that the green cluster is where one of 'remainers' or 'leavers' is embedded within the floater agents (influencing them to their own beliefs), and the red cluster is where 'remainers' and 'leavers' are, at best, only weakly connected with the floaters (resulting in diverse beliefs there).

Discussion

The resulting patterns of belief come from a number of sources: (a) what beliefs are available in the social network an individual is connected with; (b) what beliefs predominate in the social network an individual is connected with; (c) who an individual is connected to; and (d) the coherence of beliefs with its existing beliefs

Critical to the outcomes seems to be the conditions under which the individual belief coherence and the social structure can co-evolve – producing recognisably distinct groups with similar beliefs. However, this co-evolution may not occur when this is frustrated by the interconnectedness of the network, or the sheer amount of belief noise (essentially random changes in belief).

Such models can help place explanations of individual and social patterns of belief in terms of the macro- and meso-outcomes. For example, if we accept the social intelligence hypothesis (Kummer at al 1997), that the abilities that were crucial to our survival and evolution were our social abilities, then we would expect our cognitive abilities are more attuned to social functioning than any ideal of 'rationality'. From a macroscopic viewpoint of a group of individuals learning to survive in a particular ecological niche, it may be important that the beliefs of its members are functionally coherent, but in a crisis, that the group may split into factions with differing (but still internally coherent) sets of beliefs. Such a process of group development and selection may allow a species to inhabit and exploit a wide variety of different kinds of niches, and thus have some protection against unpredictable catastrophes.

The model presented here exhibits just such tendencies, whilst at the same time being consistent with a plausible micro-level assumptions (as discussed in the introduction). Here the emergent macro-level outcomes can not be reduced to purely social or individual processes but exhibits social embeddedness (Granovetter 1985). The time series of aggregate opinion in these models exhibit the characteristic unpredictability, turning points and noisiness of observed opinion poll time series. The social networks it produces look plausible. However, at the moment, this is only an illustrative model, to show the potential of this kind of simulation, and we cannot make reliable conclusions about observed social systems from it. I aim to develop this to support an explanation of some observed patterns of belief.

Acknowledgements

I would like to thank everybody with whom I have discussed these, and related, issues. This includes: Edmund Chattoe-Brown, Giangiacomo Bravo, Paul Thagard, Scott Moss, David Hales, Rosaria Conte, Alan McKane, and Guillaume Deffuant as well as those at: the Workshop on *Coherence-Based Approaches to Decision Making, Cognition, and Communication* in Berlin (July 2016), the symposium on *Social Agency and Interactivity* in Odense (Sept 2016) and the *Interdisciplinary Workshop on Opinion Dynamics and Collective Decision* at Jacobs University, Bremen (July 2017). I cannot completely absolve them of any errors in this paper, as is traditional in such acknowledgements, due to the remote possibility that the model presented here is on the right lines and thus that their social influence is an *unavoidable* part of my knowledge construction.

References

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. Psychological Monographs, 70.

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*, 5, (4), 1. http://jasss.soc.surrey.ac.uk/5/4/1.html

Edmonds, B. 2012. Modelling Belief Change in a Population Using Explanatory Coherence. *Advances in Complex Systems*, 15, (6), 1250085. DOI: 10.1142/S0219525912500853

Edmonds, B. 2016. A Model of Social and Cognitive Coherence. CoMSES Computational Model Library. http://www.openabm.org/model/5116

Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American Journal of Sociology*, 91, 481-510.

Kummer, H., Daston, L., Gigerenzer, G., and Silk, J. 1997. The social intelligence hypothesis. In Weingart et. al (eds.), *Human by Nature: between biology and the social sciences*. Hillsdale, NJ: Lawrence Erlbaum, pp. 157-179.

McPherson, Miller, Lynn Smith-Lovin, and J. M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual of Sociology*, 27, 415–44.

Thagard, P. 1989. Explanatory Coherence. *Behavioral and Brain Sciences*, 12, 435-467. http://cogsci.uwaterloo.ca/Articles/1989.explanatory.pdf

Xu, Bo, Liu, Renjing and Liu, Weijiao (2014), Individual Bias and Organizational Objectivity: An Agent-Based Simulation. *Journal of Artificial Societies and Social Simulation*, 17, (2), 2, http://jasss.soc.surrey.ac.uk/17/2/2.html. DOI: 10.18564/jasss.2426