Interacting Brains: The Dynamics of Belief among Reasoning Agents

John Mahoney and Kai Spiekermann*


People talk to each other to change each other’s beliefs. Sometimes changing each other’s beliefs is easy: When John says, “Kai, we’re meeting for Kick-ball at 7”, Kai is likely to add the belief “There will be Kick-ball at 7” to his set of beliefs. Oftentimes, changing each other’s belief is not that simple. Whether we accept the proposition another person suggests to us depends on the beliefs we already hold. For example, if someone tries to convince us that the world is flat, it is rather unlikely that we add this belief to our set of beliefs. The attempt to convince us fails because it directly contradicts our belief that the world is a sphere, and it indirectly contradicts many other beliefs about astronomy, geology, cartography, travel, etc. More generally, the success of convincing a person of a belief depends on the beliefs this person holds and on the logical connections between these beliefs and the new, proposed belief.

To explore the dynamics of belief change, it is necessary to model agents as having logically interdependent beliefs. When humans reason and argue, they do not only disagree on one isolated proposition. Rather, they revise their stance towards one proposition in the light of many other propositions. The notions of “deliberation” and “public reason”, both core concepts in political theory, are about this exchange of reasons (Rawls, 1999; Gutmann and Thompson, 1996; Habermas, 1992). The importance of structured, logically interdependent beliefs is also of prime importance for modelling cultural dynamics. A culture is not a vector of independent, isolated beliefs. Culture should rather be seen as a highly interdependent net of beliefs supporting each other. Nevertheless, many models of belief or culture dynamics work without any logical connections between beliefs. Robert Axelrod (1997) and Bednar et al. (2006) choose a vector of integers as a representation of a culture. Simplifying representations in the interest of building parsimonious models is often a good choice (and these models have other merits), but in this case the extreme simplification does not have much in common with real human reasoning, real deliberation, or real culture. More importantly, ignoring the logical dependence of different beliefs leads to entirely different dynamics, in particular to a much stronger tendency towards homogeneity of beliefs, compared to dynamics with structured beliefs.

*This work was partially supported by the Santa Fe Institute through NSF Grant No. 0200500 entitled “A Broad Research Program in the Sciences of Complexity.”
We propose to represent the belief structures of agents as directed graphs. This way of modelling has two advantages. Firstly, we can show that many logical constraint expressible in propositional logic can be represented in our model. Secondly, the graph structure provides an intuitive approach to classify the different structural properties of logical interdependencies. For instance, one can distinguish between hierarchical belief structures, where top-level beliefs influence the lower-level beliefs, and circular belief structures, where beliefs are linked in a more holistic manner.

This paper is in five parts. Section 1 gives a more detailed description of our modelling strategy. Section 2 analyses a basic dynamics for populations with identical belief structures. In section 3 we look at a different, circular belief structure. We then proceed to look at possible extensions. Section 5 concludes.

1 Brains as Graphs

Let there be $N$ agents. The agents in the game debate on an agenda of $k$ propositions $P_1, P_2, P_3, ..., P_k$. We mark each proposition with a tilde to denote that the agenda allows agents to believe either $P_x$ or its negation $\neg P_x$. Each agent has beliefs over all propositions, that is the agent assigns a truth value to each proposition. This could be written down as the set of all beliefs the agent holds (for example $P_1, \neg P_2, \neg P_3, P_4, ...$), but for simplicity we represent the beliefs of agent $i$ as a vector $b_i$, with $b_{ix} = 0$ if agent $i$ thinks that proposition $P_x$ is false, and $b_{ix} = 1$ if the agent thinks that it is true (the example would translate into $(1, 0, 0, 1, ...)$).

We distinguish between an agent’s beliefs and her world view. The world view is the set of logical rules which determine the dependency between different propositions. The world view and beliefs taken together are called a “brain”. We use the term “agent” and “brain” interchangeably.

We represent an agent’s world view and beliefs as a graph. Each node represents a proposition and has a status. Status 1 means that the agent believes the proposition is true, status 0 means that the agent believes the proposition is false. We display nodes with status 1 (true) as green, nodes with status 0 (false) red. The directed edges in the graph represent the logical interdependencies between propositions. These logical dependencies determine the susceptibilities of nodes to change their status. Whether a vertex is susceptible to change its status during deliberation depends on the status of its children. The children of node can be found by looking at the directed edges (arrows): The arrows point from a node $n$ to other nodes. These nodes are $n$’s children. Not all nodes necessarily have children.

The rules for the logical dependencies between beliefs are coded in a “key” for each node. The key determines which constellation of the children must be given for the node to be susceptible to change. The intuition behind the key metaphor is the hierarchical structure of beliefs. Some of our beliefs change easily when we take up new information. Other beliefs are more central and

---

1We use the terms “node” and “vertex” interchangeably.
more dependent on other beliefs. We are only willing to change a core belief when we have been convinced to change a number of other beliefs. These other beliefs are the key to the core belief. If all the other beliefs are in the right combination, then we are willing to change our view on the core belief. We will describe the key more formally below.

![Figure 1: A simple tree graph with key.](image)

Take figure 1 as an example. There are three nodes. Each node is labeled in the form “[Node number]: [status of node]”. Each node $x$ represents a proposition $P_x$. Node 1 has node 2 and 3 as children (and node 2 and 3 have no children). Node 1 has status 1, which is equivalent to proposition 1 believed to be true; node 2 and 3 have status 0 which signifies that proposition 2 and 3 are believed to be false.

The key in figure 1 is the key for node 1 (the other two nodes do not have children and therefore no key). It enumerates the rules for the status of node 1, given the status of its children 2 and 3. Here the first line translates into: “If node 2 has status 1 and node 3 has status 1, node 1 is susceptible to have status 1” (we explain what we mean by “susceptible” below). For the second line, this is: “If nodes 2 has status 0 and node 3 has status 0, then node 1 is susceptible to have status 0”. In a notation similar to propositional logic, we write $(P_2 \land P_3) \implies P_1$ and $(\neg P_2 \land \neg P_3) \implies \neg P_1$. The curly arrows indicate that these are not logical implications, but mere susceptibilities. It is perfectly possible that a brain remains in a state of inconsistency, that is a state where one or more key rules are not satisfied. Keys do not need to state rules for all children combinations. In the example, the combinations $\{0,1\}$ and $\{1,0\}$ do not have a rule. This means nothing follows for the status of node 1 for these combinations.

In the example, node 1 is susceptible to change from status 1 to status 0 because its two children both have status 0 and the key rule $\{0,0\}$ applies. When this change occurs depends on the update rules used. In this paper we consider only one update rules: Interaction with other agents. Other update rules are conceivable, in particular forms of “reflection” where agents update...
their beliefs to become more consistent.

The interaction update rule concerns changes of node status through the interaction between two brains. We distinguish between the “attacker” brain and the “target” brain. The attacker brain tries to convince the target brain to change the truth value for exactly one proposition.\(^2\) This means that the status of only one node is at stake. In the example in figure 2 the attacker targets node 1. Here node 1 of the target is susceptible to change, as we can see by looking at key rule \(\{0,0\}0\). The result of this interaction will be that the target changes the status of node 1 from 1 to 0, that is it changes the truth value for proposition \(P_1\) from True to False. If the attack node and the target node have the same status, nothing changes. Note that attacks on nodes without children are always successful because nodes without children are unconstrained by key rules.

\[ \text{targets n1} \]

\[ \text{key}[1] = \begin{pmatrix} 1, 1 \\ 0, 0 \end{pmatrix} \]

\[ \text{key}[1] = \begin{pmatrix} 1, 1 \\ 0, 0 \end{pmatrix} \]

Figure 2: The left brain targets the right brain on node 1.

Now we can describe the key rule more formally. Let \(z\) be the target brain and \(a\) be the attacker brain. In general, the key rule is a function that maps the current beliefs of \(a\) and \(z\) at time \(t\) to their new beliefs at time \(t + 1\). Formally, \(\text{key function} : \{b_{ad}, b_{za}, \{\forall b_{zd}^{\gamma} \in ch(d)\}\} \rightarrow b_{zd}^{t+1}\). The specific key functions we are using take as arguments the disputed node \(d\), the beliefs the attacker and target have over the disputed note \(b_{ad}^{t}\) and \(b_{za}^{t}\) at time \(t\), and the beliefs the target hold over the children of the disputed node \(ch(d)\), which are \(\{\forall b_{zd}^{\gamma} | \gamma \in ch(d)\}\). The output of our key rule is the new belief the target brain has over the disputed node \(b_{zd}^{t+1}\) at time \(t + 1\).

In practice, this means that each node with children has its own key rule. We assume that all agents have the same key rule for the same node positions in the brains. The key rule is a look-up table, giving instructions how the belief

\(^2\)This assumption is made for the sake of simplicity. More complex “discussions” between brains are conceivable and should be addressed in future models.
connected to the node reacts when attacked. The full look-up table for the example we discussed above is in Table 1.

<table>
<thead>
<tr>
<th>children beliefs</th>
<th>susceptible to be</th>
<th>action</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>1</td>
<td>if belief of node is 0, change to 1 if attacked by 1</td>
</tr>
<tr>
<td>(0,0)</td>
<td>0</td>
<td>if belief of node is 1, change to 0 if attacked by 0</td>
</tr>
<tr>
<td>(1,0)</td>
<td>-</td>
<td>do not change</td>
</tr>
<tr>
<td>(0,1)</td>
<td>-</td>
<td>do not change</td>
</tr>
</tbody>
</table>

Table 1: Example of a key rule for a node with 2 children.

To simplify, we usually state key rules by writing the first two columns and omit the lines where no change is prescribed. Again, it is important to emphasize that the key rules describe susceptibilities. The belief of the node can stay in a state where the key rule is not satisfied when no attack on the node is made.

The deliberation process is modeled in the following way. The simulation works with randomly mixed agents, i.e., for each iteration two agents are randomly chosen from the population to play against each other. One of them takes the attacker role, the other one is the target. One node for interaction is chosen randomly (signifying that the agents argue about one and only one proposition at a time). The attacker tries to convince the target to copy the attacker’s status of the interaction node. This leads to the following process:

1. Randomly select attacker and target brain with target ≠ attacker.
2. Randomly select interaction node.
3. Attacker tries to convince target to adapt his belief over interaction node.
   (a) if beliefs are identical nothing happens.
   (b) if beliefs are different, target changes belief if key rule demands change or interaction node has no children, otherwise no change.
4. Go to step 1.

Here is the basic framework in pseudocode:

create N brains with defined world view
initialize node status randomly
While iterations ≤ max iterations
   Select attacker brain and target brain randomly
   Choose interaction node iNode randomly
   IF status[target, iNode] ≠ status[attacker, iNode]
      AND (key rule demands change given
           status[children[iNode]]
           OR target node has no children)
      THEN status[target, iNode] = status[attacker, iNode]
Loop
2 A Simple Example

To develop an idea how the model behaves, we ran some simulations with fairly simple world views. When setting up a simulation, we aim to motivate our assumptions by drawing parallels to real belief systems and decision problems. We begin by exploring the dynamics of the simple world view.

Consider the following example. A group of students at the SFI Complex Systems Summer School have to decide whether they want to go to the El Farol bar or to the Cowgirl bar. There are two criteria: Which bar has better drinks and which bar has better music. The students agree on a joint world view, i.e. they accept the same logical dependencies: If one thinks that one of the bars has both better drinks and better music, then one will update one's belief and prefer this bar as soon as one is challenged on that belief. However, if one thinks that one bar has better music while the other has better drinks, it is consistent to prefer either of the two bars.

Figure 3 shows how this decision situation can be represented in our framework. Each agent assumes the same logical dependencies as represented by the graph and the key rule for node 1. If an agent believes $P_2$ and $P_3$ (El Farol is better for music and drinks), then she is susceptible to believe $P_1$ (Go to El Farol). Similarly, if an agent believes $\neg P_2$ and $\neg P_3$ then she is susceptible to believe $\neg P_1$.

We simulate the dynamics of the deliberative process. We start with 11 agents. In the beginning, all agents have random beliefs over all propositions. Then they start interacting according to their world view and the interaction rules described above. We have developed a number of measures to describe the outcomes:

- Percent of nodes with status 1.
- Persistence: percent of nodes that have never changed status.
- Resilience: percent of nodes that have their initial status.
• Average consistency over nodes. The consistency of each node is counted as
  
  - +1 if the key rules imply susceptibility to adopt the value that is already currently held by the node (i.e., the node is in a consistent state)
  - 0 if the key rules imply nothing (i.e., the node is in a neutrally consistent state)
  - -1 if the key rules imply susceptibility to adopt a value that is not currently held by the node (i.e., the node is in an inconsistent state)

• Majority consistency: same measure for the majority position for each single proposition.

With these measures we can look at the dynamic of the “Bar Example”. We ran a simulation with 11 agents and 1000 interactions. We can see in figure 4 that percentage of nodes stabilizes after around 900 interactions. This indicates that the brains have all reached a stable state. The persistence approaches 0, suggesting that virtually all nodes have been changed at least once. The resilience stabilizes at around 50%.

![Graph showing 100.0% of nodes with 1, Persistence, and Resilience over time.]

Figure 4: Results for Bar Example I.

Figure 5 reports the development of the two consistency measures and it shows the majority beliefs over the three propositions. In this run, the majority thinks that El Farol has better music and cheaper drinks. Consistently, the majority wants to go to El Farol. We can see that the average agent and the (fictive) majority agent are consistent in their beliefs after deliberation.

We can see the effect of deliberation when looking at the average consistency measure before and after deliberation, as shown in figure 6. Clearly, deliberation “washes out” inconsistent belief sets. After prolonged deliberation, only very few freak cases remain inconsistent, most outcomes are either neutral (in the sense that there is no key rule applying) or positively consistent. The freak cases occur when all the brains have the same inconsistent set of beliefs. Since all beliefs are similar, these beliefs are not changed, even though they are inconsistent. These freak cases would disappear if we introduced trembling (agents occasionally change their beliefs at random).

Finally, the results show that a vote in the Bar Example often leads to a very clear majority. Figure 7 shows the distribution of vote percentages obtained
after running 1000 simulations. In more than 60% of the cases did the simulation result in an unanimous vote for one of the bars.

These results are interesting because they support the claim that deliberation has a positive effect to “launder preferences” (Goodin, 1986) and to avoid majority inconsistencies and produce clear results (Dryzek and List, 2003). One
can also see that deliberation often leads to high levels of homogeneity. This is interesting because these results underpin the theory put forward by Knight and Johnson (1994), by Miller (2003), and then treated more systematically by Dryzek and List (2003), that deliberation can help to solve social choice problems by removing inconsistent arguments from the political sphere. As a consequence, collective inconsistencies such as the discursive dilemma (List, 2007) are less likely.

3 Closed Belief Systems

One advantage of modeling belief systems as graphs is that one can easily explore the effect of different world views. The tree graphs for the Bar Example are quite an “open” belief system because two of the nodes are without children, which means there are no logical constraints for the corresponding propositions. In this section, by contrast, we look at agents with a more closed world view. Our example is simple, and maybe not very realistic, but it gives us a good first intuition how the structure of the logical dependencies matters for the outcome.

![Figure 8: A closed belief system. Key rule for all nodes: \{1\} 1 \{0\} 0](image)

Figure 8 shows a very simple highly interdependent belief system. The key rule states that a node is susceptible to have status 1 if its child has status 1, and susceptible to be 0 if its child has status 0. One does not need to run a simulation to grasp the belief change dynamics when agents have this shared world view. Each brain will “lock” into a state where all three nodes have the same status. Once locked, the brain is immune to any outside influence. Since each node satisfies its key rule, the nodes will never change status again. This is what we should expect: More closed belief systems are less flexible in changing and may even be completely isolated from any new information.

A second interesting aspect for the dynamics of this world view is that it allows for heterogeneity between the agents. It is possible that some agents lock into an “all 1” state, while others lock into an “all 0” state. This is a stable state, but a stable state where different agents have completely opposite views on the whole agenda. Deliberation does not help to ameliorate this disagreement because the brains are too closed to change.
4 Extensions

4.1 Diverse Graph Structures

What happens when we relax the assumption that all agents share the same world view? We have to work on this question more, but early simulations and theoretical considerations suggest that oftentimes stable states will not be reached. Consider a situation where we mix agents with the world view of the Bar Example ("tree brains") and agents with the "closed" world view discussed in the last section ("closed brains"). We know that the closed brains will lock in quickly, and we know that they might lock into different states (some with "all 1", some with "all 0"). If that happens, the system cannot settle into a stable state: Since the tree brains meet the different locked closed brains in the interactions, they change their beliefs all the time and the system does not reach a stable state.

4.2 The Discursive Dilemma

Apart from the question of whether the dynamics of belief change lead to homogeneity or heterogeneity, there are at least two other questions that can be addressed with our model. The first question relates to research regarding "judgement aggregation", the second to question of social epistemology and truth tracking. We discuss judgement aggregation in this section and issues of social epistemology in the next section.

Judgement aggregation has recently emerged as an important question within the field of social choice theory (List, 2007; List and Pettit, 2002). Roughly speaking, theories of judgement aggregation analyze how and to what extent a group of agents can aggregate their individual judgements over a set of propositions to a collective judgement on these propositions, given certain requirements of rationality, fairness, etc.. It turns out that there are often no satisfying aggregation rules when the propositions are logically connected. In this paper we limit ourselves to a specific problem of judgement aggregation, the so-called "discursive dilemma". Table 2 gives an example of the standard form the dilemma as described by List (2007). There are three agents (1, 2, 3). Each agent has a belief regarding the propositions \( P, Q, R \), and the compound proposition \( (P \land Q) \leftrightarrow R \). All agents accept the compound proposition. One could interpret this as a logical framework shared by all agents. However, the agents assign different truth values to \( P, Q \), and by applying the rule \( (P \land Q) \leftrightarrow R \) also different truth values to \( R \). The dilemma consists of the inconsistent majority positions on each of the propositions: The majority thinks that \( P, Q \), and \( (P \land Q) \leftrightarrow R \) is true. This implies that \( R \) is also true. Inconsistently, the majority believes that \( R \) is false.

The discursive dilemma as shown in table 2 conveys into our model as a graph with three nodes and key rules that are equivalent to \( (P \land Q) \rightsquigarrow R \). Our model does not exactly model the discursive dilemma because we allow for temporary inconsistency, as indicated by the curly equivalence symbol. This
<table>
<thead>
<tr>
<th></th>
<th>$P$</th>
<th>$Q$</th>
<th>$(P \land Q) \leftrightarrow R$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent 1</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>Agent 2</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Agent 3</td>
<td>True</td>
<td>False</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Majority</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>False</td>
</tr>
</tbody>
</table>

Table 2: The discursive dilemma.

translates into a simple tree graph and a key rule as shown in figure 9.

![Figure 9: Model of the discursive dilemma. The key rule is](image)

It would be interesting to run simulations of this and determine how often the discursive dilemma occurs, or whether deliberation is able to prevent collective inconsistencies. First results suggest that this might be the case, but further research is needed.\(^3\)

### 4.3 Epistemic Considerations

Another interesting modeling question arises when we let agents “interact with the world” and see whether they approach the truth of the world. One could do this by creating one or more “truth agents” whose beliefs are fixed. When agents interact with a world agent they obtain evidence about the real truth values. One could then determine how well the agents “track the truth” contained in the world.

\(^3\)One problem we currently have is that our coding does not allow changes of a node status when no key rule applies. This seems to be unrealistic. In the discursive dilemma, nodes 2 and 3 should be susceptible to change to 1 (true) when node 1 has status 1. In our current coding this also implies that nodes 2 and 3 do not change in an interaction when node 1 has status 0. This needs to be changed to create a more realistic deliberation process. A more general implementation of the key rules would allow for greater modeling flexibility.
5 Conclusion

Modeling processes of dynamic belief change and deliberation requires a model that takes logical interdependencies into account. Unlike earlier models by Axelrod (1997) and Bednar et al. (2006), our approach incorporates these interdependencies. Using graphs to display the structure of logical connections enables us to develop an intuitive grasp on different logical belief structures. Agents can have hierarchical, tree-like structures when they have some core beliefs which are “protected” by many lower-level beliefs. For instance, ideologies or religious convictions could be understood as high-level beliefs, whereas less central beliefs are on a lower level. Oftentimes, belief structures are not hierarchical, but rather “holistic” or circular. These belief systems have a tendency to be self-supporting, but also more isolated from external information.

We have sketched some possible extensions. We also aim to make our model more general, in particular regarding the key rules and the interaction process and to develop a precise notation for the different world views. It would also be sensible to add a process of internal reflection, where agents iron out internal inconsistencies.

Our model can be used add a more systematic, analytical take on several issues in political theory and sociology, and social psychology. Processes of opinion dynamics, social epistemology, and collective intelligence have attracted attention in recent years, but the complexity of the processes involved makes it difficult to obtain analytical results. Computer simulations may be a way forward to understand these complex systems better. Our model is a first step in this direction, but there is a lot of work left to be done.

References


**URL:** [http://personal.lse.ac.uk/list/doctrinalparadox.htm](http://personal.lse.ac.uk/list/doctrinalparadox.htm)

