Thinking About Thinking and Its Cognitive Limits

Adam Brandenburger  
Stern School of Business  
Polytechnic School of Engineering  
NYU Shanghai  
New York University

Xiaomin Li  
Behavioral & Social Neuroscience Program  
California Institute of Technology
In minimax theory --- whether for the two-player zerosum case or the \( n \)-player non-zerosum case --- players avoid a predictive approach.
John von Neumann: “Each player must choose his strategy in ‘complete ignorance’”

John Nash: “[A] rational prediction should be unique”
We proceed by investigating the question: what would be a "rational" prediction of the behavior to be expected of rational playing the game in question? By using the principles that a rational prediction should be unique, that the players should be able to deduce and make use of it, and that such knowledge on the part of each player of what to expect the others to do should not lead him to act out of conformity with the prediction, one is led to the concept of a solution defined before.

In equilibrium theory, players are assumed to have access to the actual strategies chosen by the other players.

Theory of Mind in tasks:

Theory of Mind in games:

Theory of Mind ability:

<table>
<thead>
<tr>
<th>Study</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stiller and Dunbar (2007)</td>
<td>Modal level of failure is 5 (−1?)</td>
</tr>
<tr>
<td>Arad and Rubinstein (2012)</td>
<td>Maximum level is 3</td>
</tr>
<tr>
<td>Kneeland (2015)</td>
<td>Maximum level is 4</td>
</tr>
</tbody>
</table>

A **cognitive limit** on thinking about thinking comes into effect at a small finite number of levels.
Models of thinking about thinking:

In game theory, uncertainty about the structure of the game (Harsanyi 1967-68) or about the strategies in the game (epistemic game theory) has been modeled via sequences of spaces.

\[ S \] 

0th-order space

\[ \mathcal{P}(S) \] 

1st-order space

\[ \mathcal{P}(\mathcal{P}(S)) \] 

2nd-order space

Under standard regularity assumptions, the sizes of the spaces are:

\[ |S| \quad 2^{\aleph_0} \quad 2^{\aleph_0} \quad \ldots \]

Neural evidence in Bhatt and Camerer (2005): Players were asked to make choices and to state first- and second-order beliefs in games

\[ \hat{S}_a \] (1)
\[ \square_a \hat{s}_b \] (2)
\[ \square_a \square_b \hat{r}_a \] (3)

Brain activity in event (3) was found to have more similarities with activity in event (1) than with activity in event (2)

Greater activity was found in the anterior insula in event (3) than in event (2)
An anchoring and adjusting process:

Ann selects a candidate strategy choice (she anchors)

She then examines her view as to whether or not Bob thinks she intends to make this choice (she adjusts)

There is precedent for anchoring and adjusting processes in the Theory of Mind literature, e.g., in gauging another individual’s preferences (Epley, Keysar, Van Boven, and Gilovich 2004; Tamir and Mitchell 2010)

In some sense, this process can be thought of as an (internal) equilibrium vs. disequilibrium process
Say Ann is **rational** if \( \hat{s}_a \) and \( \Box_a \hat{r}_b \) imply that \( s_a \) maximizes Ann’s (expected) payoff when she assigns probability 1 to Bob’s choosing \( r_b \). Define rationality for Bob similarly (with Ann and Bob interchanged).

**Epistemic conditions for Nash equilibrium:**

- \( \hat{s}_a \)
- \( \hat{s}_b \)
- \( \Box_a \hat{s}_b \)
- \( \Box_b \hat{s}_a \)
- Ann is rational
- Bob is rational

- \( \Box_a \hat{s}_b \)
- \( \Box_b \hat{s}_a \)
- \( \Box_a \Box_b \hat{s}_a \)
- \( \Box_b \Box_a \hat{s}_b \)
- \( \Box_a \text{ [Bob is rational]} \)
- \( \Box_b \text{ [Ann is rational]} \)
Axioms:

\[ \hat{s}_a \rightarrow \Box_a \hat{s}_a, \Box_a \hat{s}_b \rightarrow \Box_a \Box_a \hat{s}_b, \text{ and } [\text{Ann is rational}] \rightarrow \Box_a [\text{Ann is rational}] \]

Epistemic conditions subjectivized:

\begin{align*}
\hat{s}_a \\
\hat{s}_b \\
\Box_a \hat{s}_b \\
\Box_b \hat{s}_a \\
\text{Ann is rational} \\
\text{Bob is rational}
\end{align*} \quad \rightarrow \quad
\begin{align*}
\hat{s}_a \\
\Box_a \hat{s}_b \\
\Box_a \Box_b \hat{s}_a \\
\text{Ann is rational} \\
\Box_a [\text{Bob is rational}]
\end{align*}
Disequilibrium:

Change Ann’s second-order belief relative to her choice via negations, and then push negation in
Disequilibrium cont’d:

Disqualify conditions consistent with equilibrium
Following the analogous process with the second set of epistemic conditions for Nash equilibrium:
Complexity of thinking about thinking:

Ann (tentatively) fixes a strategy-pair \((s_a, s_b)\)

In forming an \(m\)th-order belief, she considers \(2^m + 1\) cases

The process could be repeated for different candidate strategy pairs

The **exponential** increase in the number of cases at each level fits qualitatively with a small finite cognitive bound (we would not expect quantitative agreement)

To-do’s:

- Extend experiments to collect data on higher-order beliefs
- Distinguish more carefully internal epistemic equilibrium from reasoning about levels of rationality
- Study games with more than two players
- Examine belief modalities in-between probabilistic and point-belief