Theory of Mind and Game Theory

Adam Brandenburger

J.P. Valles Professor, NYU Stern School of Business
Distinguished Professor, NYU Tandon School of Engineering
Faculty Director, NYU Shanghai Program on Creativity + Innovation
Global Network Professor
New York University

Version 04/09/23

A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

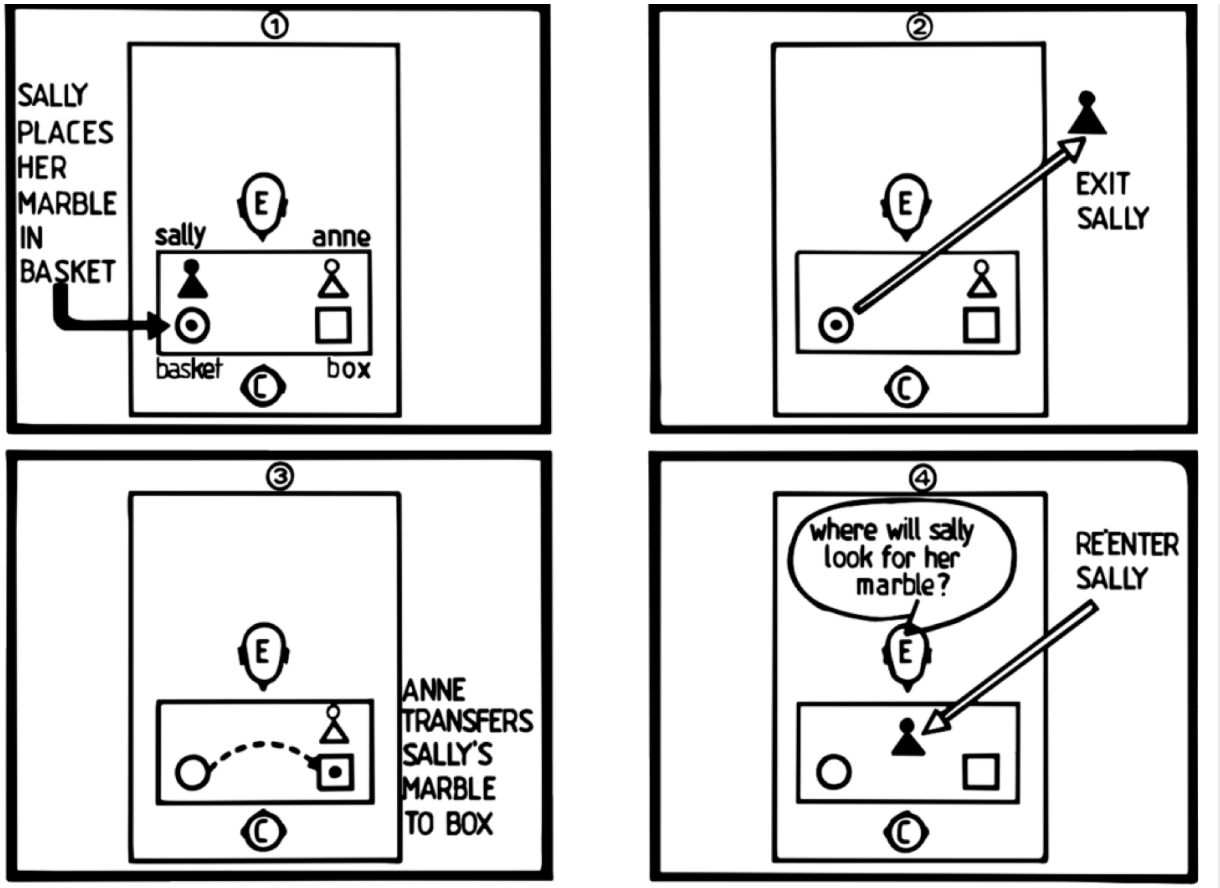
I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

Computer Science > Computation and Language

[Submitted on 4 Feb 2023 (v1), last revised 14 Mar 2023 (this version, v3)]

Theory of Mind May Have Spontaneously Emerged in Large Language Models

Michal Kosinski



Cartoon from Baron-Cohen, S., A. Leslie, and U. Firth, "Does the Autistic Child Have a 'Theory of Mind'?" *Cognition*, 21, 1985, 37-46

[Submitted on 16 Feb 2023 (v1), last revised 14 Mar 2023 (this version, v5)]

Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks

Tomer Ullman

Cartoon from Ullman op. cit.

Epistemic Game Theory in One Slide

Epistemic game theory equips each player in a strategic-form game with

a strategy set

a payoff function

a space of hierarchies of beliefs over the strategies chosen

A hierarchy of beliefs consists of a first-order belief (over strategies chosen), a second-order belief (over strategies chosen and first-order beliefs), etc.

We can think of this set-up as the multi-person analog to the "trilogy" of decision theory

Hierarchies of beliefs can (under appropriate conditions) be represented by "epistemic types," which are analogs to Harsanyi "payoff types," or by Aumann "partition structures"

Harsanyi, J., "Games with Incomplete Information Played by 'Bayesian' Players, I–III," *Management Science*, 14, 1967-68, 159-182, 320-334, 486-502; Aumann, R., "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 55, 1987, 1-18; Brandenburger, A., "The Power of Paradox: Some Recent Developments in Interactive Epistemology," *International Journal of Game Theory*, 35, 2007, 465-492

Using Game Theory to Test ToM

Start with the following basic theorem in epistemic game theory:

Fix a complete epistemic game.

(i) The set of strategies consistent with rationality and mth-order belief in rationality is the set of $(m+1)$-undominated strategies

(ii) If, in addition, the type spaces are compact and the belief maps are continuous, then the set of strategies consistent with rationality and common belief in rationality is the set of iteratively undominated strategies

Brandenburger, A., "On the Existence of a 'Complete' Possibility Structure," in Basili, M., N. Dimitri, and I. Gilboa (eds.),*Cognitive Processes and Economic Behavior*, Routledge, 2003, 30-34; Friedenberg, A., and H.J. Keisler, "Iterated Dominance Revisited," *Economic Theory*, 72, 2021, 377-421; Battigalli, P., and M. Siniscalchi, "Strong Belief and Forward Induction Reasoning," Journal of Economic Theory, 106, 2002, 356-391

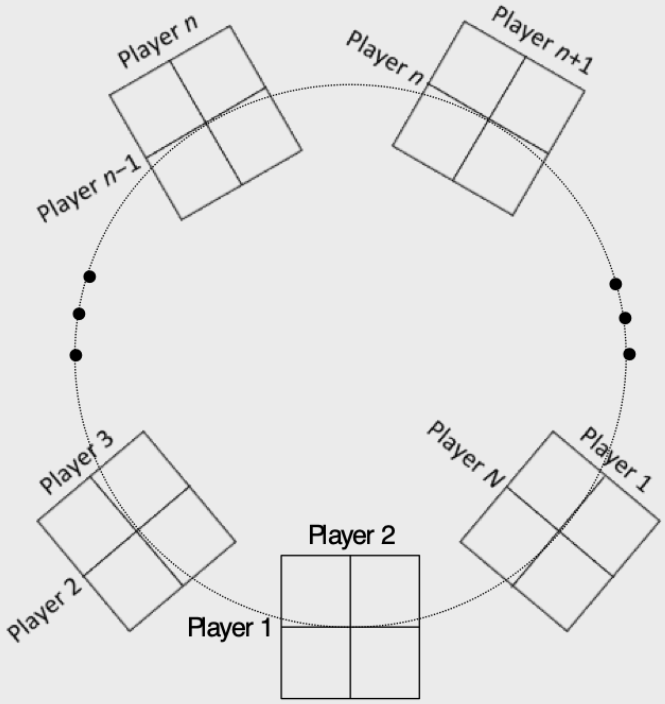
The Identification Challenge

The strategies consistent with rationality and $(m+1)$th-order belief in rationality are a subset of the strategies consistent with rationality and mth-order belief in rationality, for any $m = 0,1,2,\ldots$

This is just a re-statement of the basic argument that these epistemic conditions correspond to iterated removal of "bad" (dominated) strategies.

This simple observation implies the following identification problem: A player might choose a strategy consistent with a high number of levels of reasoning, but might do so without necessarily reasoning to this level.

Ring Games



Player 1's payoffs depend on the strategy he chooses and on the strategy player 2 chooses (and on no other player's choices)

Player 2's payoffs depend on the strategy she chooses and on the strategy player 3 chooses (and on no other player's choices) …

The experimenter now examines player 1's behavior (say) as the payoffs to other players are manipulated

Kneeland op. cit.

Beyond Iterated Dominance

	L	R
U	1,1,3	1,0,3
D	0,1,0	0,0,0

X

	L	R
U	1,1,2	0,0,0
D	0,0,0	1,1,2

Y

	L	R
U	1,1,0	1,0,0
D	0,1,3	0,0,3

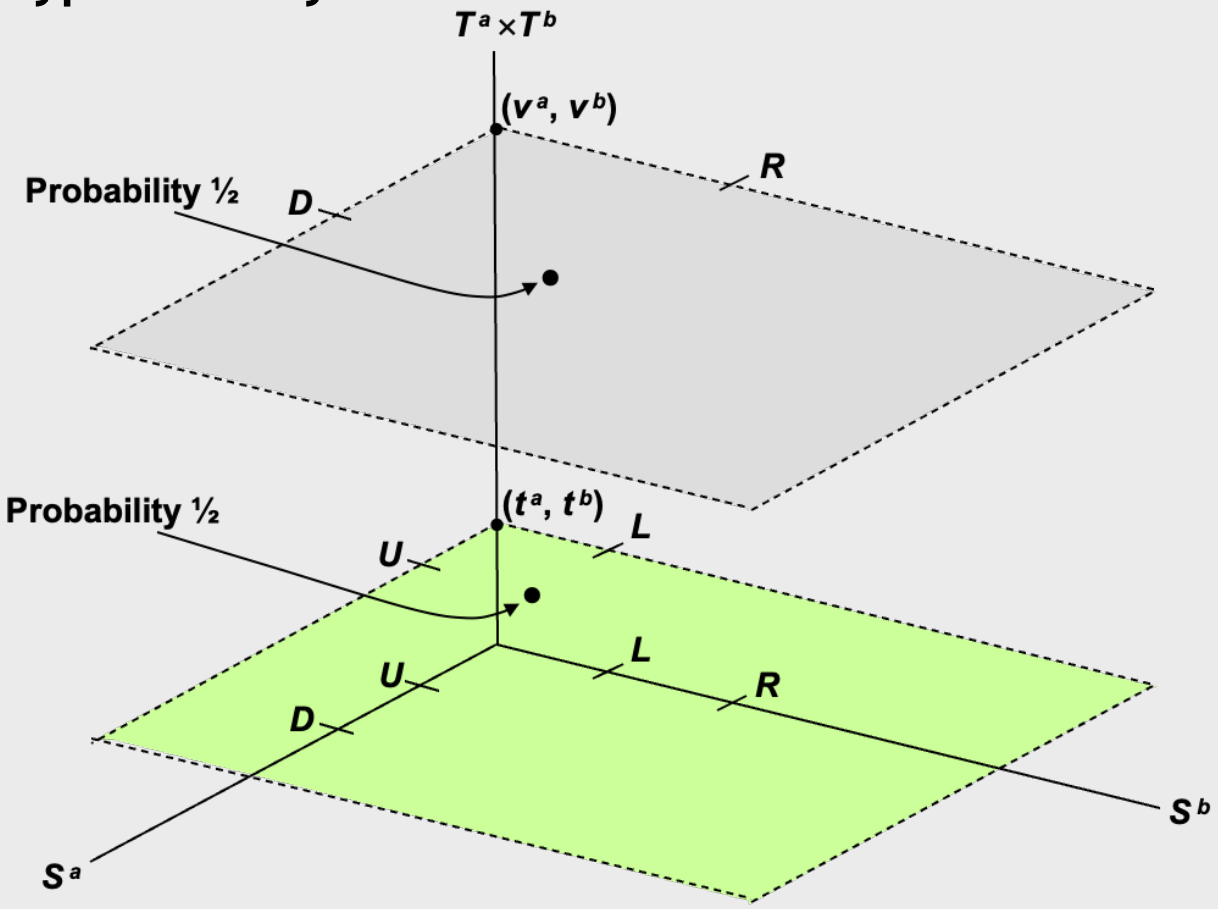
Z

The strategy Y is optimal for Charlie, under a probability measure that puts probability $1/2$ on (U, L) and probability $1/2$ on (D, R)

It is therefore undominated

But there is no product probability measure under which Y is optimal

An Epistemic Type for Player *c*



Type t^c assigns probability $1/2$ to (U, t^a, L, t^b)

and probability $1/2$ to (D, v^a, R, v^b)

Conditions on Epistemic Type Structures

Conditional Independence (CI):

Charlie's type t^c should satisfy

$p(s^a, s^b \mid t^a, t^b) = p(s^a \mid t^a, t^b) \times p(s^b \mid t^a, t^b)$ whenever $p(t^a, t^b) > 0$

(likewise for Ann and Bob)

Sufficiency (SUFF):

Charlie's type t^c should satisfy

$p(s^a \mid t^a, t^b) = p(s^a \mid t^a)$ whenever $p(t^a, t^b) > 0$

(likewise for b, and for Ann and Bob)

Under CI and SUFF:

If $p(t^a, t^b) = p(t^a) \times p(t^b)$, then $p(s^a, s^b) = p(s^a) \times p(s^b)$

In words, a correlated assessment about strategies implies a correlated assessment about types (no physical correlation)

Brandenburger and Friedenberg op.cit.

The Question

What strategies can be played in a game under the requirements of rationality and common belief in rationality (RCBR), CI, and SUFF?

In particular, can any iteratively undominated strategy be played under these conditions?

Note: Brandenburger and Dekel (1987) show that any iteratively undominated strategy can be played under an a posteriori equilibrium (Aumann, 1974), and vice versa

Brandenburger, A., and E. Dekel, "Rationalizability and Correlated Equilibria," *Econometrica*, 55, 1987, 1391-1402;
Aumann, R., "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, 1, 1974, 76-96

An Impossibility Theorem

There is a game G and an iteratively undominated strategy s^i of G, such that the following holds: For any epistemic type structure, there does not exist a state at which each type satisfies CI, RCBR holds, and s^i is played

There is an analogous theorem for SUFF

(There is also a finite-levels analog to this result)

Brandenburger and Friedenberg op.cit.

Conclusions

It is 110 years since the first formal theorem in game theory (Zermelo, 1913)

But it is still early days in the development of a game theory based on Theory of Mind!

Under the epistemic view, a full Theory of Mind in games can be argued to require a characterization of the strategies that can be played under the preceding theory of "intrinsic correlation"

This appears to be an open (hard?) question

Speculations:

- Will epistemic game theory be useful in further testing for ToM in AI's (such as large language models)?
- Would an AI's reasoning to a higher number of levels than a human be a kind of "super-intelligence"?

Zermelo, E., Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels," in *Proceedings of the Fifth International Congress of Mathematicians* (Cambridge, 1912), Cambridge University Press, 1913, 501-504