

Drumul spre Inteligența artificială de încredere trece (și) prin Teoria Jocurilor ?

(și despre echilibre Kantiene)

Gabriel Istrate

Universitatea din București

gabrielistrate@acm.org



Ce caută un teoretician la un Workshop de Inteligență Artificială, la Diaspora Științifică ?

- Am fost "diaspora" 13 ani (Rochester, Los Alamos).
- Cariera științifică "de bază": **algoritmi și complexitate** (dar și sisteme complexe, simulări sociale multiagent).
- Recent: **teoria algoritmică a jocurilor și sisteme multiagent: lucrări (long/BlueSky) la AAMAS în fiecare an 2019-2022**, TARK, SAT, etc.
- Lucrări în pregătire în curs de trimitere JAIR, Artificial Intelligence Journal, o anumită conferință cu deadline-ul în 17 mai 😊
- **Sunt interesat de colaborare pe teme de teoria algoritmică a jocurilor, agenți, metode (neuro)simbolice, etc.**

Ce caută un teoretician la un Workshop de Inteligență Artificială ?

(Submitted on 21 Jan 2023)

Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change)

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, Subbarao Kambhampati

The recent advances in large language models (LLMs) have transformed the field of natural language processing (NLP), from GPT-3 to PaLM, the state-of-the-art performance on natural language tasks is being pushed forward with every new large language model. Along with natural language abilities, there has been a significant interest in understanding whether such models, trained on enormous amounts of data, exhibit reasoning capabilities. Hence there has been interest in developing benchmarks for various reasoning tasks and the preliminary results from testing LLMs over such benchmarks seem mostly positive. However, the current benchmarks are relatively simplistic and the performance over these benchmarks cannot be used as an evidence to support many a times outlandish claims being made about LLMs' reasoning capabilities. As of right now, these benchmarks only represent a very limited set of simple reasoning tasks and we need to look at more sophisticated reasoning problems if we are to measure the true limits of such LLM-based systems. With this motivation, we propose an extensible assessment framework to test the abilities of LLMs on a central aspect of human intelligence, which is reasoning about actions and change. We provide multiple test cases that are more involved than any of the previously established reasoning benchmarks and each test case evaluates a certain aspect of reasoning about actions and change. Initial evaluation results on the base version of GPT-3 (Davinci), showcase subpar performance on these benchmarks.

Subjects: [Computation and Language \(cs.CL\)](#); [Artificial Intelligence \(cs.AI\)](#)

Chat at: [arXiv:2301.10483v1 \[cs.CL\]](#)

(arXiv:2209.10483v1 [cs.CL] for this version)

- Agenții inteligenți **trebuie să interacționeze cu oamenii și să înțeleagă raționalitatea lor, atâta câtă e (sau lipsa ei)**
- Proiect (pe termen lung): să incorporăm **tot ce ne spun Economia Comportamentală/teoria deciziei despre comportamentul uman într-o bibliotecă software.**

Game Theory



- My payoff may depend not only on what I am doing but also on what others are doing.
- Search for "equilibrium" points.

In this talk: shorthand for Decision Theory + Game Theory.

Game Theory in one minute

	C	D
C	2, 2	0, 3
D	3, 0	1, 1

- agents: want to maximize their utility
- **Nash equilibrium**: action profile so that **everyone best-responds to what others are doing**.
- may not exist, **mixed strategies** guaranteed to exist (Nash) by non-constructive methods (Borsuk-Ulam or similar fixed-point theorems).
- Prisoners' Dilemma Paradox: agents will find rational to defect even though **if they cooperated they would both be better off**

What's wrong with Game Theory?

In Theory:

- Omniscient.
- Perfectly rational
- Sociopath: only thing that matters - its utility.

In Reality:

- Bounded (if at all) rational.
- Dual process reasoning: fast/heuristic System I, slow/expensive Syst. II.
- Emotional.
- Social.

Richard Thaler called these extremes "Econs" and "Humans".

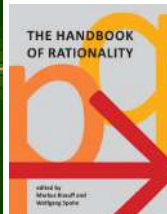
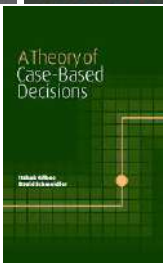
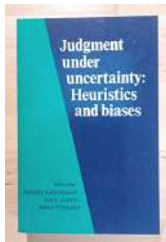
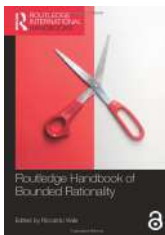
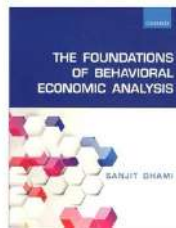
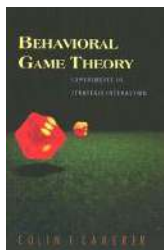
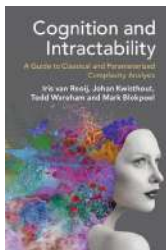
Normative rather than descriptive: How the Econ should play, rather than how the Human actually plays.

Why is the Classical Game-Theoretic Person a Caricature ?

- **We're (too much) in love with proving theorems.**
- Mathematically tractable models of behavior: simplified.
Real life: complex/messy.
- Insistence on mathematical (rather than computational) models: **we may not even have the "right" framework !**

Disclaimer: **I will show you some theorems later.** In my defense, they all try to point towards **what (and how) to implement (or not) game theoretic concepts.**

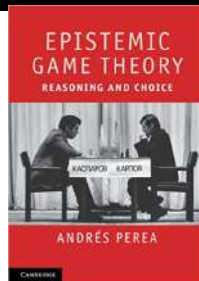
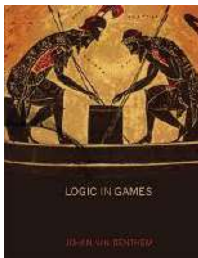
(Some) Good News



Bad News

- All this knowledge: far from being implemented/available for building intelligent agents.
- Game theory software: **GAMBIT**, **nashpy** track classical game theory.
- Not even clear we have the "right" models to implement !

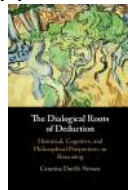
OK, so let's use logic to formalize games ...



- ... Except that **logic may have many of the same problems.**
- Barto, Smets & Solaki (Erkenntnis, 2021): "Now just as mainstream economics has forgotten Humans to focus on Econs, **so has mainstream logic forgotten them to focus on Logons.** We name this way the ideal agents studied in "static" epistemic logic with possible worlds semantics (Hintikka 1962) and in AGM belief revision theory (Alchourron et al. 1985). **These agents are logically omniscient: perfectly consistent, closed under classical logical consequence in their beliefs, and free from framing effects in their belief revision policies [...].** In fact, **Econs may just be Logons engaged in rational choice.** The focus on Logons has opened a rift between logic and cognition, similar to the one between the latter and economics"

Good News: The Cognitive Turn in Logic & KR

Mercier & Sperber: reasoning **not** primarily useful for infer new facts, but to **win arguments** (convince opponents).



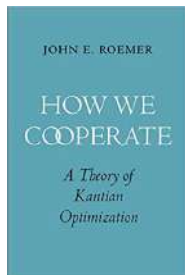
Also:

- "natural logic" (Moss): **small, natural, tractable** fragments.
 - "cognitive logics" (Kern-Isberner, see e.g. KR'2021 tutorial)
 - cost of reasoning (Solaki & Smets, *WOLLIC 2015*)
 - dual-process logics (Barto, Solaki & Smets, *Erkenntnis*'19).
- **Bad News:** Not that many implementations.
(Perhaps) reason DL won against Symbolic AI: **lots of computational tools, easy to tinker with**

Now for sampler of concrete (theoretical) results (TARK 2021)

Kantian equilibria:

- Non-Nashian equilibrium notion, clear(est) definition: **symmetric coordination games**.



	s_1	s_2	\dots	s_n
s_1	a_1, a_1	$0, 0$	$0, 0$	$0, 0$
s_2	$0, 0$	a_2, a_2	$0, 0$	$0, 0$
\dots	\dots, \dots	\dots	\dots	\dots
s_n	$0, 0$	$0, 0$	$0, 0$	a_n, a_n

- (loosely) based on **Kantian categorical imperative**: "act only according to that maxim whereby you can, at the same time, will that it should become a universal law"
- all agents play **the same action** x_{OPT} , chosen to maximize $\pi(x, x, \dots, x)$.

Example: Kantian Equilibria in Prisoners' Dilemma

	C	D
C	2, 2	0, 3
D	3, 0	1, 1

- Nash equilibrium: both agents defect.
- Kantian equilibrium: both cooperate and are both better off

Requirements

Interested in **minimal rationality equilibrium notions**, easily implemented in agents that reason "like people".

Therefore, want notions that are:

- **expressive**: can model plausible scenarios.
- **computationally tractable**: computing equilibria should be easy (cf van Rooij, *The tractable cognition thesis*, Cognitive science, 2008).
- **cognitively tractable, easy to formally specify**: no costly epistemic assumptions (common knowledge, many rounds of iterated elimination of dominated strategies; see also complexity notions of Solaki & Smets)

Main research problem: **Are Kantian equilibria (suitably generalized) such a notion ? Are there more suitable related/competing notions ?**

Questions & Intuitive Conclusions

- What is the **complexity of (mixed) Kantian equilibria** ?

NP-hard even for two-player symmetric games. Such equilibria problematic if multiple equilibria exist.

- Can one **define Kantian equilibria in more general games** ?

Yes, in games with certain "symmetry".

- Can one "interpolate" between Kantian and Nash behavior ? **yes.**
- Kantian equilibria or other related notions ?

Theoretically interesting, but probably team reasoning more useful for implementations.

Mixed Kantian equilibria: example

probabilistic combinations of pure strategies. Want combination that maximizes expected payoff when played by everyone

"Platonia Dilemma" (Hofstadter): each agent one of two strategies S, N . Payoff: 1 if the only agent to play S , 0 otherwise.

- Pure Kantian equilibrium: everyone plays N , payoff 0, or everyone plays S , payoff 0.
- mixed Kantian equilibrium: everyone independently plays S w.p. $\frac{1}{n}$, 0 otherwise.

Symmetric Coordination Games: Mixed Kantian equilibria are easy (but useless)

THEOREM : In symmetric coordination games (more generally in diagonally dominant games:) all mixed Kantian equilibria are pure (hence easy to compute)

Proof idea: $E[p] = \sum_{i,j} a_{i,j} p_i p_j \leq \max(a_{i,j}) (\sum p_k)^2 = \max(a_{k,k})$.

Any equilibrium makes this equality. If there were two different actions i, j in the support of p then we would contradict the diagonally dominant hypothesis.

Mixed Kantian equilibria: "hard" beyond symmetric coordination games

THEOREM : The following problem, **MIXED KANTIAN EQUILIBRIUM**, is NP-hard:

INPUT: Two-player symmetric game G , and an aspiration level $r \in \mathbb{Q}$.

TO DECIDE: Is there a mixed strategy profile $x = (x_1, \dots, x_m)$ such that the utility of every player under common mixed action $x_1 a_1 + x_2 a_2 + \dots + x_m a_m$ is $\geq r$?

Proof Idea: Follows implicitly from results in the literature (for problem QUADRATIC OPTIMIZATION, which turns out to be equivalent): Motzkin & Strauss (1965), computing X_{OPT} in $O(1)$ two-player symmetric games equivalent to computing MAX-CLIQUE in equivalent graph.

Cognitive Science 32 (2008) 939–984
Copyright © 2008 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210801897856

The Tractable Cognition Thesis

Iris van Rooij

Nijmegen Institute for Cognition and Information, Radboud University Nijmegen

Received 7 May 2007; received in revised form 26 November 2007; accepted 12 December 2007

Abstract

The recognition that human minds/brains are finite systems with limited resources for computation has led some researchers to advance the *Tractable Cognition thesis*: Human cognitive capacities are constrained by computational tractability. This thesis, if true, serves cognitive psychology by constraining the space of computational-level theories of cognition. To utilize this constraint, a precise and workable definition of “computational tractability” is needed. Following computer science tradition, many cognitive scientists and psychologists define computational tractability as polynomial-time computability, leading to the *P-Cognition thesis*. This article explains how and why the P-Cognition thesis may be overly restrictive, risking the exclusion of veridical computational-level theories from scientific investigation. An argument is made to replace the P-Cognition thesis by the *FPT-Cognition thesis* as an alternative formalization of the Tractable Cognition thesis (here, FPT stands for fixed-parameter tractable). Possible objections to the Tractable Cognition thesis, and its proposed formalization, are

Problems with mixed Kantian Equilibria

	C	D	E
C	5, 5	3, 6	1, 2
D	6, 3	4, 4	6, 3
E	2, 1	3, 6	5, 5

	C	S
C	10, 10	100, 200
S	200, 100	6, 6

- **First game:** two Kantian equilibria, (C, C) , (E, E) playing mixture of them bad.
- Theorem: "price of miscoordination" for Kantian equilibria.
- **Second game:** as defined, Kantian equilibria "bad".
Players would like to **anti**coordinate.

Program Equilibria

- Tennenholtz (*Games Econ. Behavior*, 2004): to any game one associate extended game, whose actions are **programs**.
- agents know the text of other's programs, **can act on it**.
- **program equilibrium** = Nash equilibrium of extended game.

IF (your-program == my-program) THEN Cooperate ELSE Defect

Kantian Program Equilibria

- want: same idea (Kantian equilibrium of extended game).
- only possible when **players have identical action sets.**
- also: Not clear **what a program being "best for all" means in general.**
- Tennenholtz's formalization of programs \Rightarrow paradoxical results.

Kantian Program Equilibria: Platonian Dilemma

- Platonian Dilemma: protocol "best for all": (collectively) choose a random participant, it plays S, others N.
- can be implemented by agents playing the same program:

Choose a random $x_i \in \mathbb{Z}_n$. Broadcast it. Collectively compute $x = \bigoplus_{k=1}^n x_k$. If $x == i$ send S, otherwise send N.

Towards Kantian Program Equilibria

	C	S
C	10, 10	100, 200
S	200, 100	6, 6

Anticoord(i::ID)

Randomly choose bit
 $\text{myb} \in \{0, 1\}$
send myb to the other
player as its otherb.
if $[\text{myb} \oplus \text{otherb} \equiv$
 $i \pmod{2}]$
 then play C
 else play S

Kantian (program) equilibria: our notion

- Only for **games with a certain symmetry**, quantified by group actions.
- expected player utility across an orbit of the action: **the same**.

Definition

A game Γ is called *Pareto symmetric* if there exists a group H acting on the **set of Pareto-optimal action profiles** such that

- For every Pareto optimal profile $a = (a_1, a_2, \dots, a_n)$ and $u \in H$ there exists a permutation $\sigma \in S_n$ such that $u \cdot a = (a_{\sigma(1)}, \dots, a_{\sigma(n)})$.
- For every two players $i \neq j$ and value λ

$$|\{u \in H : (u \cdot a)_i = \lambda\}| = |\{u \in H : (u \cdot a)_j = \lambda\}|$$

- notion of program: **technical**.

Kantian equilibria in Pareto symmetric games

	C	S
C	10, 10	100, 200
S	200, 100	6, 6

	C	D
C	2, 2	0, 3
D	3, 0	1, 1

Platonian Dilemma: Kantian equilibrium orbit
(S, N, ..., N), (N, S, ..., N), ..., (N, N, ..., S).

- Theorem (extended version): Kantian equilibrium can be characterized as convex combinations of orbit(s) that maximize player expected payoffs.

Conclusions/Work in Progress

- Mixed Kantian equilibria: not cognitively plausible.
- generalization: goes through **program equilibria**.

- **How do agents recognize symmetry in (our notion of) Kantian games ?**

- **Hopeless if game represented by game matrix**
- (In progress) Connect it to **general game playing**. Version of *Game Description Language* (GDL, Genesereth et al., AI Magazine 2005, Tielscher IJCAI 2019) with **symmetries embedded in the description**

- (in progress) **Implementing Kantian optimization in (our version of) GDL.**