Critical Risks and Opportunities of AI, Economics, and Social Network Theory

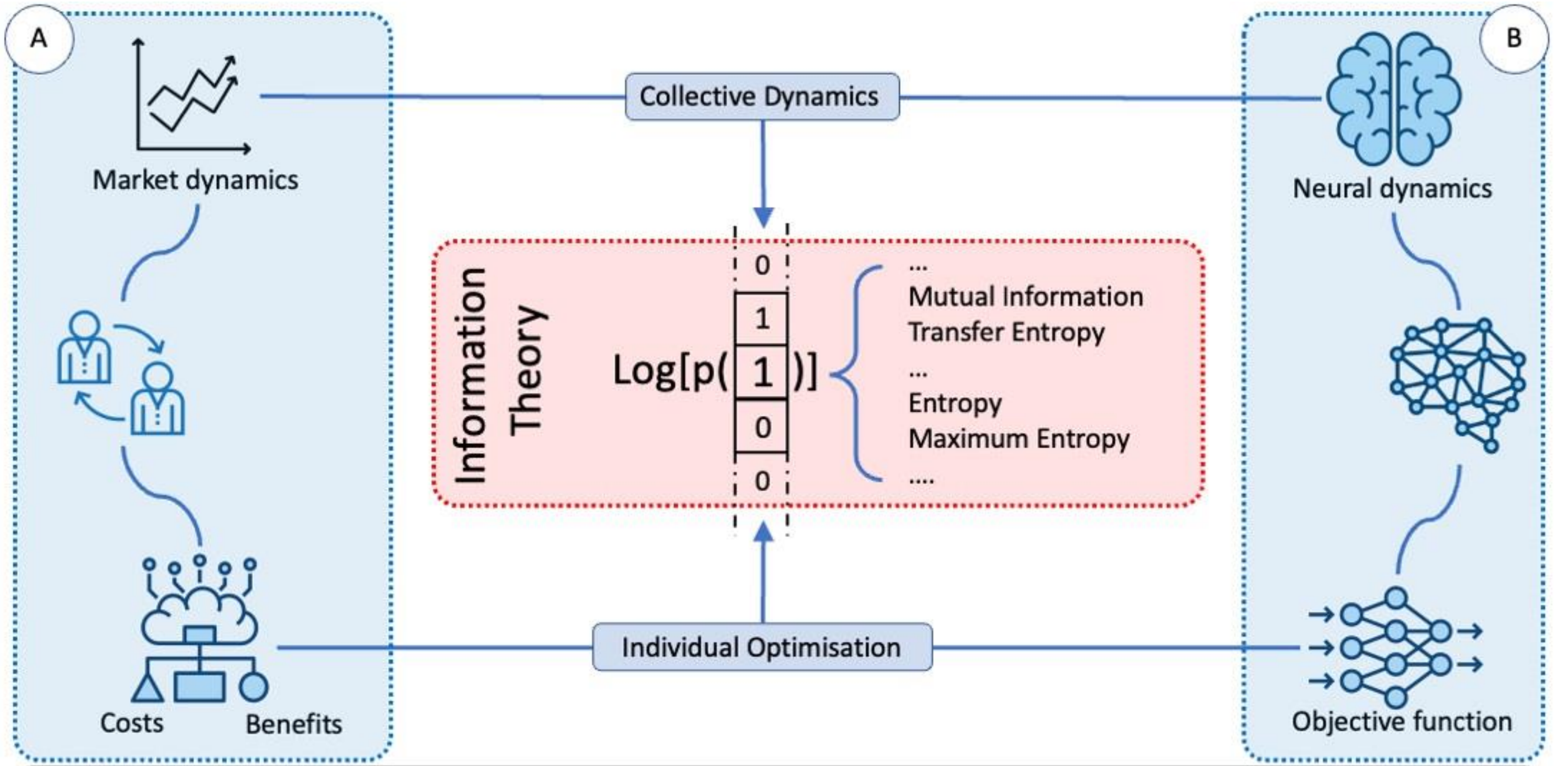
Dr. Michael Harré
Modelling and Simulation Group
School of Computer Science
University of Sydney

"Complexity, Criticality and Computation" Symposium C^3 January 2023

THE UNIVERSITY OF SYDNEY

https://enterprisetalk.com/tech-trends/ai/human-intelligence-can-fix-ai-shortcomings/

Silicon Intelligence : Economics : Psychology



Information theory and its many roles. **Box A, economics**: micro-economic optimisation and market dynamics are mediated by networks of individual exchange. **Box B, cognition**: AI and neural dynamics borrow extensively from each other. **Collective dynamics** in markets and neuroscience are measured by multivariate information theory. **Individual optimisation** and reward divergence is measured by entropy and its maximisation.

Information Theory for Agents in Artificial Intelligence, Psychology, and Economics

by Michael S. Harré

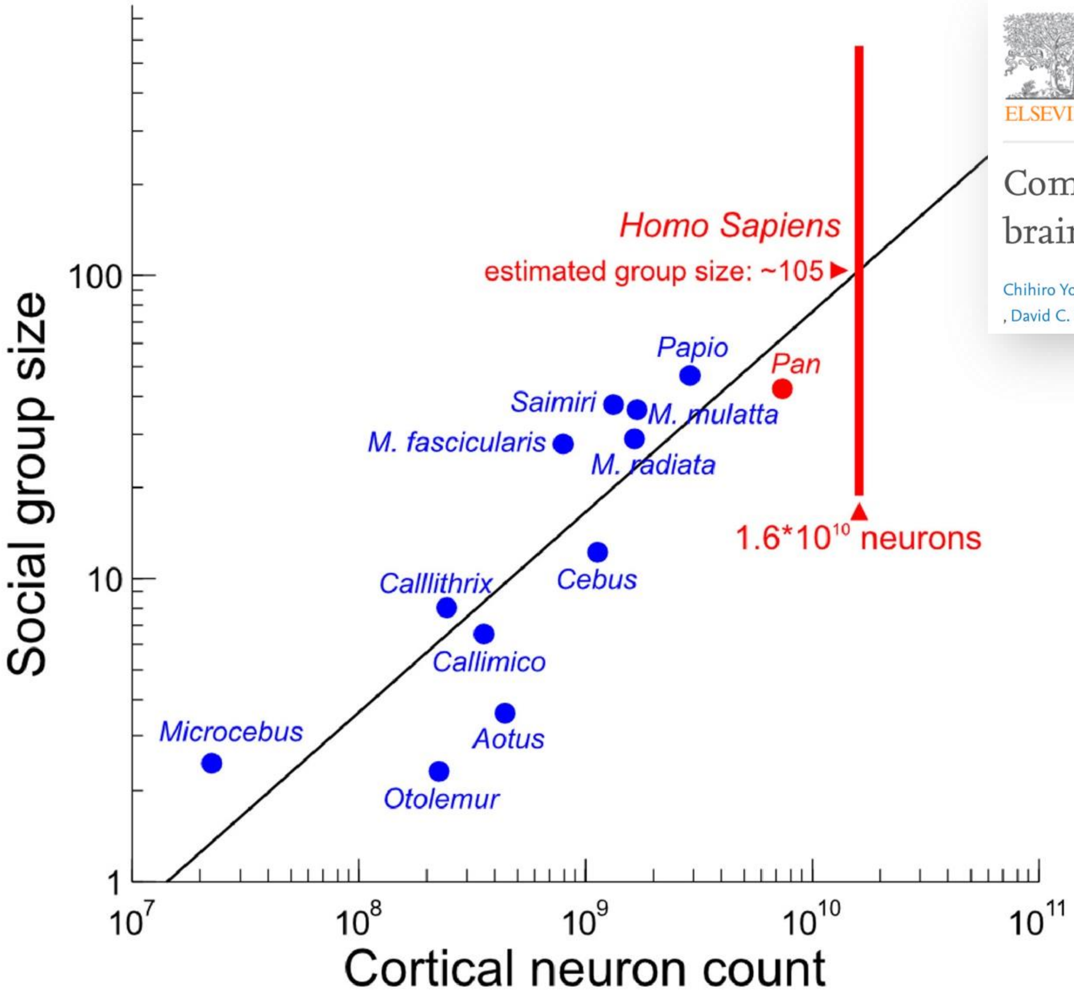
Complex Systems Research Group, Faculty of Engineering, The University of Sydney, Sydney 2006, Australia

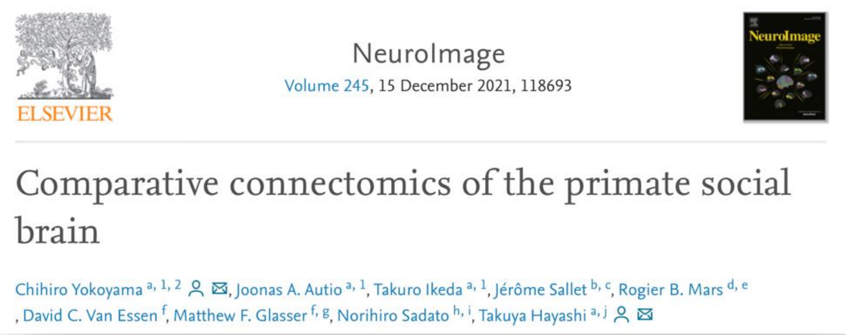
Entropy **2021**, *23*(3), 310; https://doi.org/10.3390/e23030310

Contents

THE UNIVERSITY OF SYDNEY

1. Motivating Theory of Mind for Social Networks




NeuroImage

Volume 245, 15 December 2021, 118693

Comparative connectomics of the primate social brain

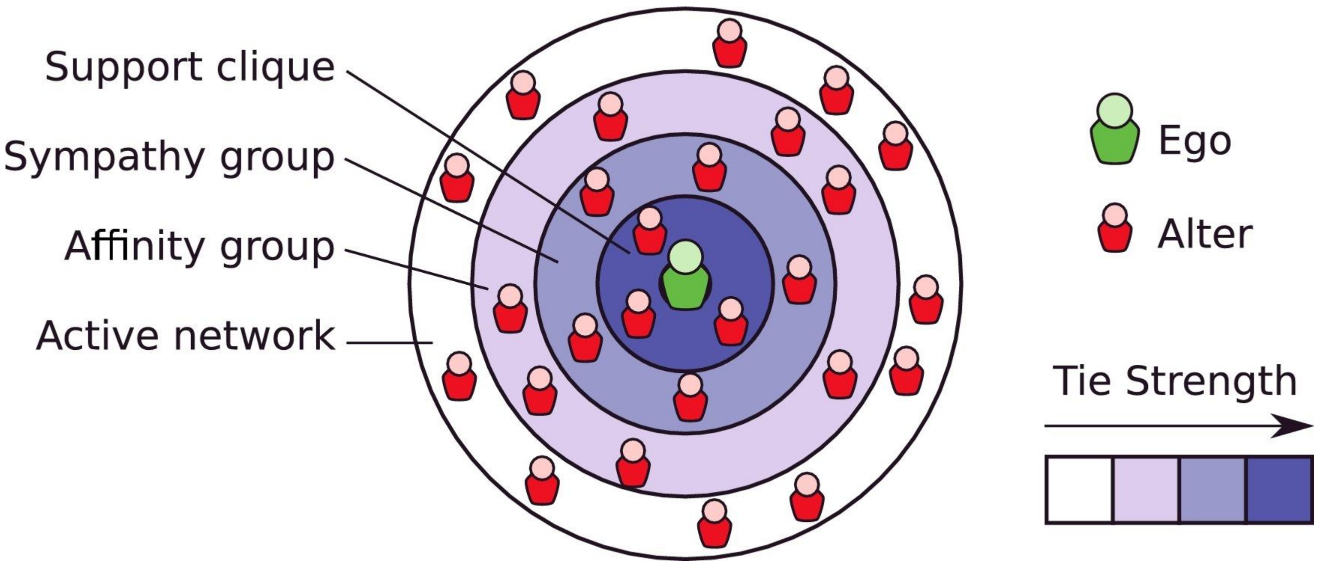
Chihiro Yokoyama [a, 1, 2], Joonas A. Autio [a, 1], Takuro Ikeda [a, 1], Jérôme Sallet [b, c], Rogier B. Mars [d, e], David C. Van Essen [f], Matthew F. Glasser [f, g], Norihiro Sadato [h, i], Takuya Hayashi [a, j]

1. Motivating Theory of Mind for Social Networks


ELSEVIER
Computer Communications
Volume 76, 15 February 2016, Pages 26-41

Ego network structure in online social networks and its impact on information diffusion

Valerio Arnaboldi, Marco Conti, Massimiliano La Gala, Andrea Passarella, Fabio Pezzoni



1. Motivating Theory of Mind for Social Networks

INTERFACE

rsif.royalsocietypublishing.org

Research

CrossMark
click for updates

The social brain: scale-invariant layering of Erdős–Rényi networks in small-scale human societies

Michael S. Harré and Mikhail Prokopenko

Complex Systems Research Group, Faculty of Engineering and IT, The University of Sydney, Sydney, Australia

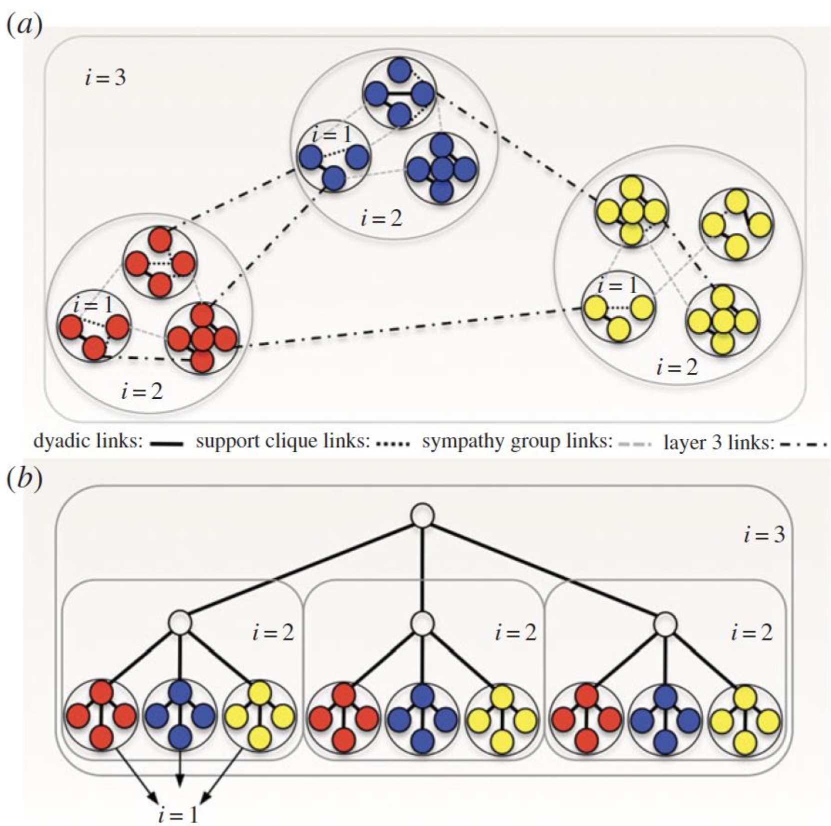


Figure 1. Two different social network models. (*a*) Random links form between sub-group members. As the average number of links per person increases in discrete steps, the network size also increases in predictable, discrete, steps. (b) A structured hierarchy similar to modern military, bureaucratic and corporate structures in which each layer is 'managed' by a coordinator. (Online version in colour.)

1. Motivating Theory of Mind for Social Networks

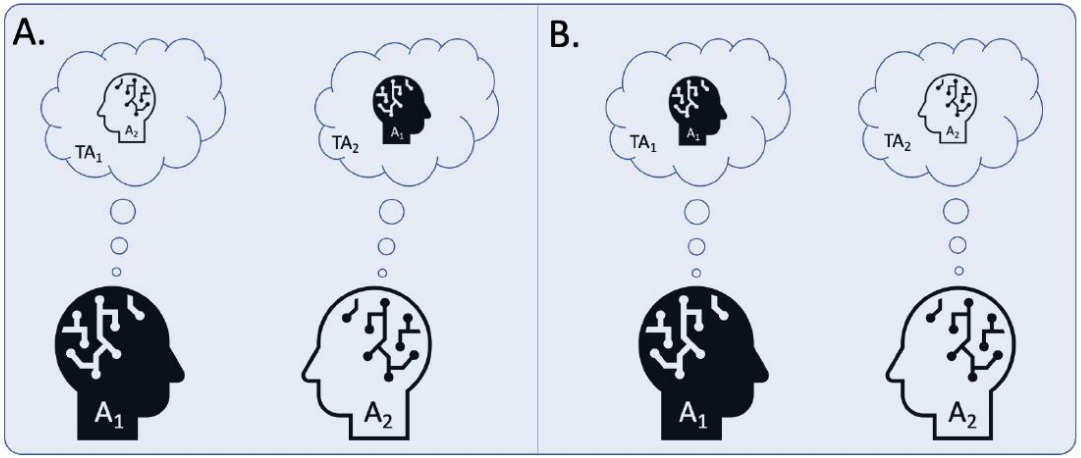


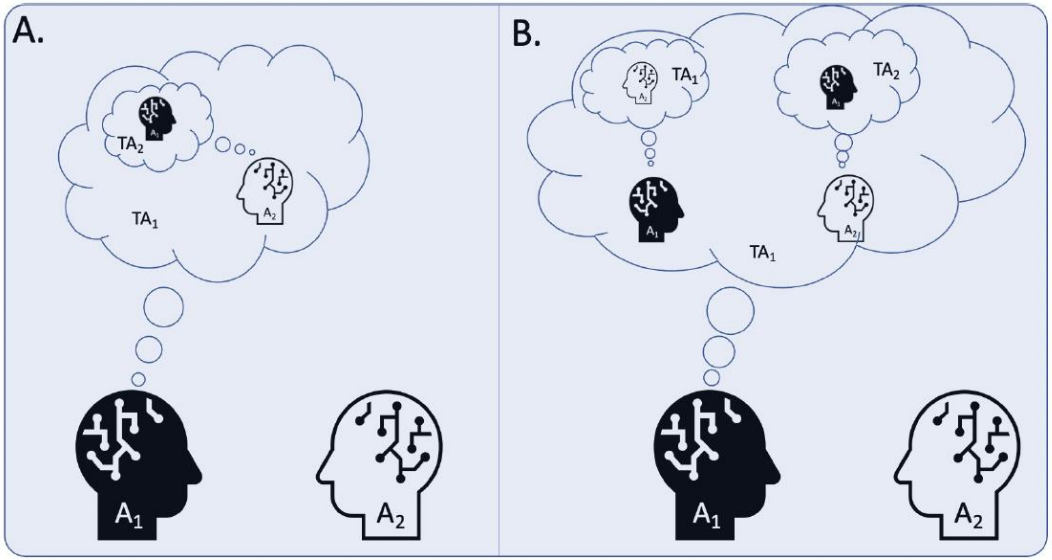
Figure 2. Left: Both agents A_1 and A_2 have a representation of the other agent's cognition (ToM). Right: Both agents have a representation of their own cognition (introspection). TA_i is a cognitive space (see for example [15,16]) in which the 'thoughts of agent i' occur.

The role of metacognition in human social interactions

Chris D. Frith

Published: **05 August 2012**
https://doi.org/10.1098/rstb.2012.0123

1. Motivating Theory of Mind for Social Networks



The role of metacognition in human social interactions

Chris D. Frith

Published: 05 August 2012

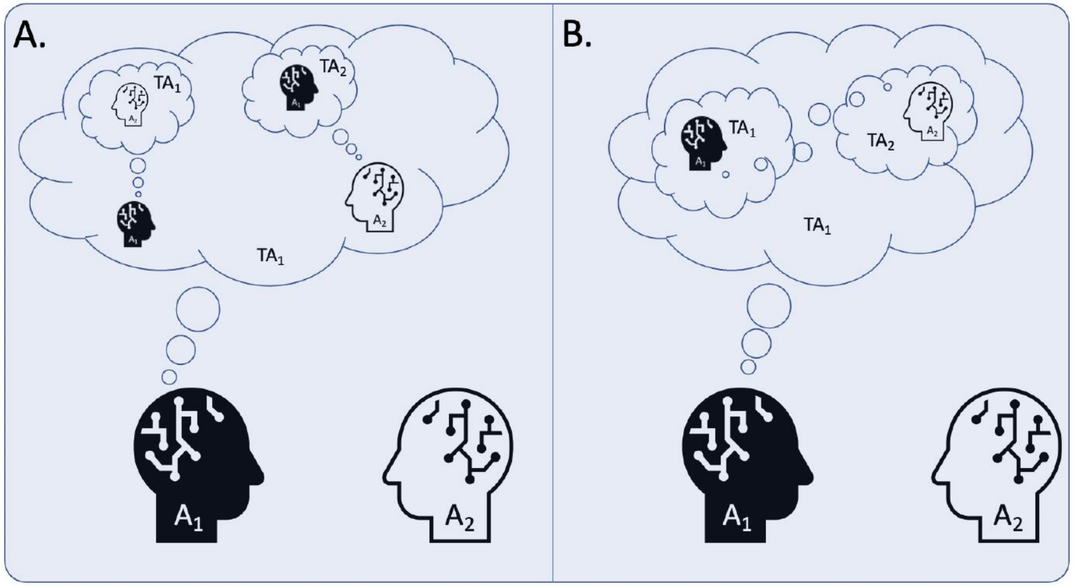
https://doi.org/10.1098/rstb.2012.0123

Game Theory of Mind

Wako Yoshida, Ray J. Dolan, Karl J. Friston

Published: December 26, 2008 • https://doi.org/10.1371/journal.pcbi.1000254

1. Motivating Theory of Mind for Social Networks



The role of metacognition in human social interactions

Chris D. Frith

Published: 05 August 2012
https://doi.org/10.1098/rstb.2012.0123

Game Theory of Mind

Wako Yoshida, Ray J. Dolan, Karl J. Friston

Published: December 26, 2008 • https://doi.org/10.1371/journal.pcbi.1000254

What Can Game Theory Tell Us about an AI 'Theory of Mind'?

by Michael S. Harré

Complex Systems Research Group, Faculty of Engineering, The University of Sydney, Sydney 2006, Australia

Games **2022**, *13*(3), 46; https://doi.org/10.3390/g13030046

Received: 30 April 2022 / Revised: 30 May 2022 / Accepted: 16 June 2022 / Published: 20 June 2022

2. Free Energy Principles of Decision Theory

2. Free Energy Principles of Decision Theory

Free Energy, Free Utility:

$$\left.\begin{aligned}
H(p) &= -\sum_x p(x)\log(p(x)) \\
V(p) &= \text{internal energy (potential function)} \\
\mathcal{F}(p) &= V(p) - T\,H(p) \quad \text{(where } T \text{ is temperature)}
\end{aligned}\right\} \text{Helmholtz Free Energy} \qquad (1)$$

2. Free Energy Principles of Decision Theory

Free Energy, Free Utility:

$$\left.\begin{aligned} H(p) &= -\sum_{x} p(x)\log(p(x)) \\ V(p) &= \text{internal energy (potential function)} \\ \mathcal{F}(p) &= V(p) - T\,H(p) \quad \text{(where } T \text{ is temperature)} \end{aligned}\right\} \text{Helmholtz Free Energy} \qquad (1)$$

$$\left.\begin{aligned} H(P) &= -\sum_{x,y} P(x,y)\log(P(x,y)) \\ U_a(P) &= E_p[\,U_a(x,y)\,] \text{ (expected utility for } a) \\ \mathcal{F}(P) &= U_a(P) - T\,H(P) \quad (T \text{ is uncertainty or error}) \end{aligned}\right\} \text{Free Utility (2 players)} \qquad (2)$$

Strategic Choice of Preferences: the Persona Model

David Wolpert , *Julian Jamison* , *David Newth* and *Michael Harre*

From the journal The B.E. Journal of Theoretical Economics

https://doi.org/10.2202/1935-1704.1593

2. Free Energy Principles of Decision Theory

$$\left.\begin{aligned} H(P) &= -\sum_{x,y} P(x,y)\log(P(x,y)) \\ U_a(P) &= E_p[\ U_a(x,y)\ ]\ \text{(expected utility for } a\text{)} \\ \mathcal{F}(P) &= U_a(P) - T\,H(P) \quad (T \text{ is uncertainty or error}) \end{aligned}\right\}\ \text{Free Utility (2 players)} \quad (2)$$

Optimising these "free functionals" leads to standard exponential solutions:

$$P(x) \propto \exp(-\beta\, V(P))$$

$$\begin{aligned} P(x \mid x = x_i) &\propto \exp(\,\beta\, U_a(\,P(y) \mid x = x_i\,)\,) \\ P(y \mid y = y_i) &\propto \exp(\,\beta\, U_b(\,P(x) \mid y = y_i\,)\,) \end{aligned}$$

Strategic Choice of Preferences: the Persona Model

David Wolpert, *Julian Jamison*, *David Newth* and *Michael Harre*

From the journal The B.E. Journal of Theoretical Economics
https://doi.org/10.2202/1935-1704.1593

2. Free Energy Principles of Decision Theory

Published: 13 January 2010

The free-energy principle: a unified brain theory?

Karl Friston

Nature Reviews Neuroscience **11**, 127–138 (2010) | Cite this article

Friston's Free Energy Principle:

One of the goals of Friston's work is to estimate the joint probability of states observed o and actual states s via Bayes Theorem:

$$P(s, o) \;=\; P(o|s)P(s)$$

2. Free Energy Principles of Decision Theory

Published: 13 January 2010

The free-energy principle: a unified brain theory?

Karl Friston

Nature Reviews Neuroscience **11**, 127–138 (2010) | Cite this article

Friston’s Free Energy Principle:

One of the goals of Friston's work is to estimate the joint probability of states observed o and actual states s via Bayes Theorem:

$$P(s, o) \;=\; P(o|s)P(s)$$

This calculation is often too difficult to compute directly so Friston's "Free Energy Principle" for the brain addresses this by estimating an alternative probability $Q(s)$ via opitmisation:

$$Q^*(s) \;=\; \underset{Q(s)}{\operatorname{argmin}}\, \mathcal{F}(\,Q\,) \tag{3}$$

$$Q^*(s) \;\simeq\; P(s|o) \tag{4}$$

2. Free Energy Principles of Decision Theory

Published: 13 January 2010

The free-energy principle: a unified brain theory?

Karl Friston

Nature Reviews Neuroscience **11**, 127–138 (2010) | Cite this article

Friston's Free Energy Principle:

One of the goals of Friston's work is to estimate the joint probability of states observed o and actual states s via Bayes Theorem:

$$P(s,o) \;=\; P(o|s)P(s)$$

This calculation is often too difficult to compute directly so Friston's "Free Energy Principle" for the brain addresses this by estimating an alternative probability $Q(s)$ via opitmisation:

$$Q^*(s) \;=\; \underset{Q(s)}{\operatorname{argmin}}\, \mathcal{F}(\,Q\,) \tag{3}$$

$$Q^*(s) \;\simeq\; P(s|o) \tag{4}$$

$$\mathcal{F}(\,Q\,) \;=\; E_Q[\,\log(Q(s)) \;-\; \log(\,P(s|o)\,)\,] \tag{5}$$

$$=\; \underbrace{E_Q(\,-\log(P(s|o))\,)}_{\substack{\text{cross entropy}\\ =\text{ expected log loss}}} \;-\; \underbrace{H(\,Q(s)\,)}_{\text{entropy}} \tag{6}$$

2. Free Energy Principles of Decision Theory

The three ideas to take away from this section:

$$\mathcal{F}(Q) = E_Q[\log(Q(s)) - \log(P(s|o))]$$

$$= \underbrace{E_Q(-\log(P(s|o)))}_{\substack{\text{cross entropy} \\ = \text{expected log loss}}} - \underbrace{H(Q(s))}_{\text{entropy}}$$

$$H(P) = -\sum_{x,y} P(x,y)\log(P(x,y))$$

$$U_a(P) = E_p[U_a(x,y)] \text{ (expected utility for } a)$$

$$\mathcal{F}(P) = U_a(P) - T\,H(P) \quad (T \text{ is uncertainty or error})$$

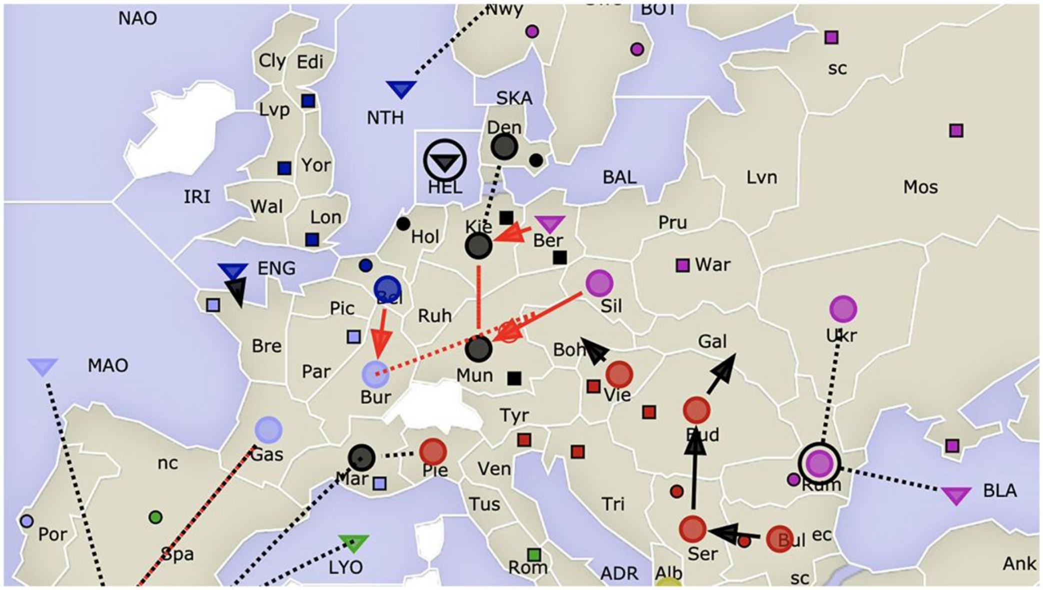
Optimising these "free functionals" leads to standard exponential solution

note: Grunwald and Dawid showed that this is a "game" between nature and decision maker (2004)

Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory
PD Grünwald, AP **Dawid** - the Annals of Statistics, 2004 - projecteuclid.org

3. Cicero: “Free Energy” for a Game Theory of Mind II

The game of “Diplomacy” and the AI Cicero that learned to play like a human



Diplomacy is a 61-year-old board game about taking over Europe. It's a true classic - highly influential, intensely beloved and widely acclaimed - and the first time I played it, I thought it was rubbish.

In Diplomacy, each player controls one of seven nations at the start of the 20th century, submitting movement orders to their armies and fleets each turn in order to gain new territories and thereby build more troops - the goal being to control 18 of the game's supply cities at the same time. Its rules are simple enough to make Risk look fiendishly difficult and yet, all the same, it is an extremely nuanced game of strategy and cunning. (JFK apparently enjoyed the game, as did Henry Kissinger.)

About Diplomacy

Released: 1959

Players: 2-7

Playing time: 360 mins

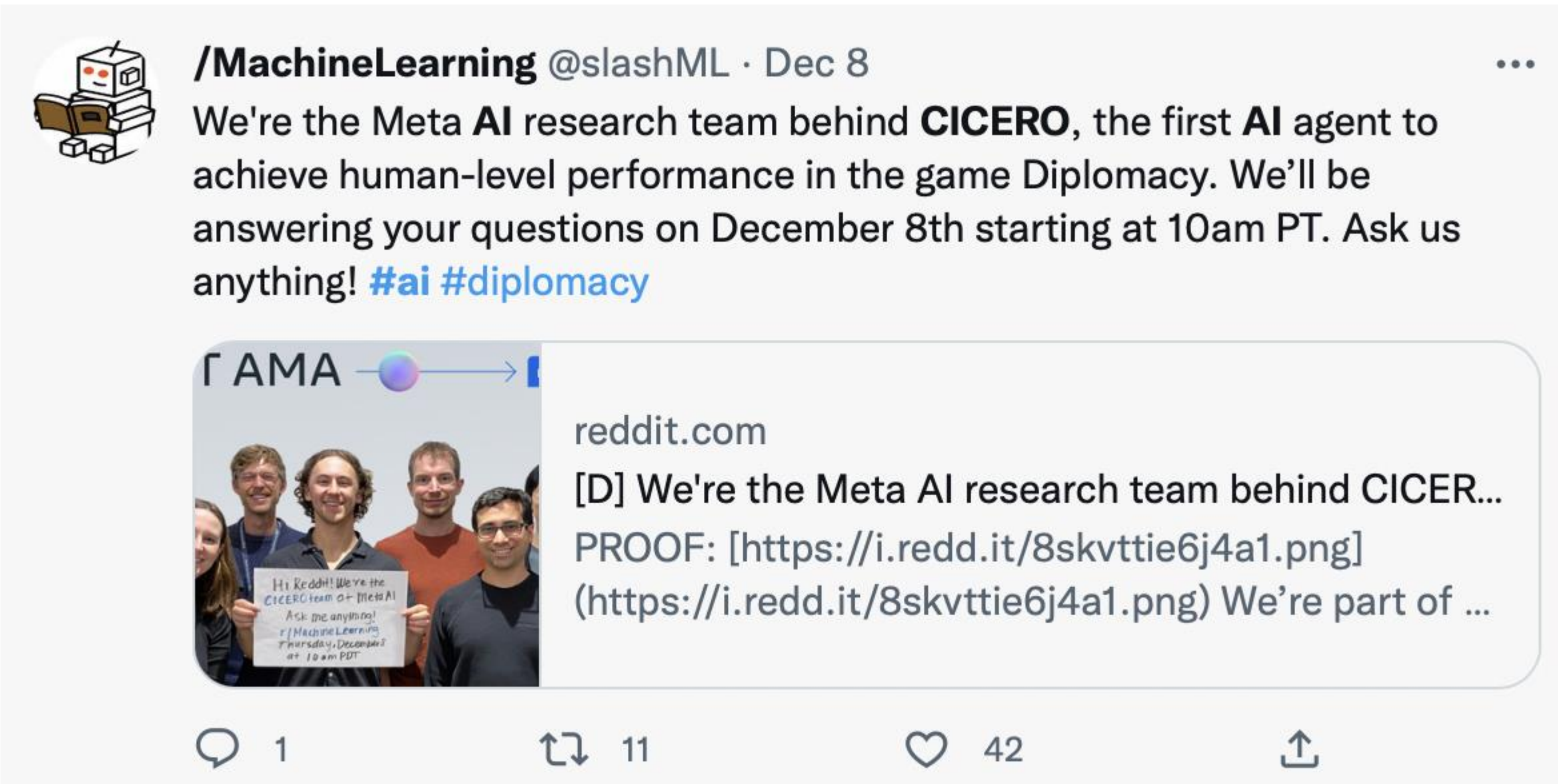
Rules complexity: Very low

Strategic depth: High

https://www.dicebreaker.com/games/diplomacy/opinion/diplomacy-online-backstabbr

3. Cicero: “Free Energy” for a Game Theory of Mind II

The game of “Diplomacy” and the AI Cicero that learned to play like a human



3. Cicero: “Free Energy” for a Game Theory of Mind II

The game of “Diplomacy” and the AI Cicero that learned to play like a human

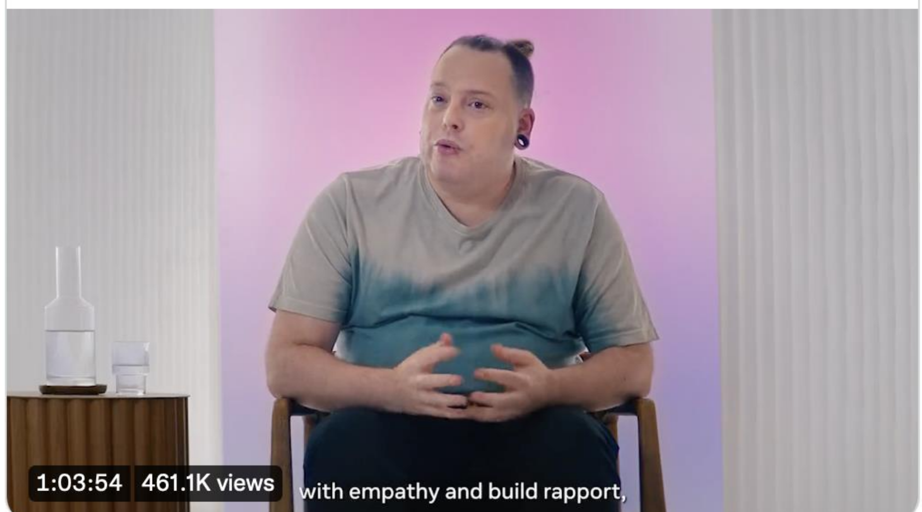
Yann LeCun @ylecun · Nov 23

Big **AI** milestone today: **CICERO**, an **AI** agent that can negotiate and cooperates with people.
It is the first **AI** system to achieve human-level performance in the popular strategy game Diplomacy.
Cicero ranked in the top 10 of participants on webDiplomacy.net

Meta AI @MetaAI · Nov 23

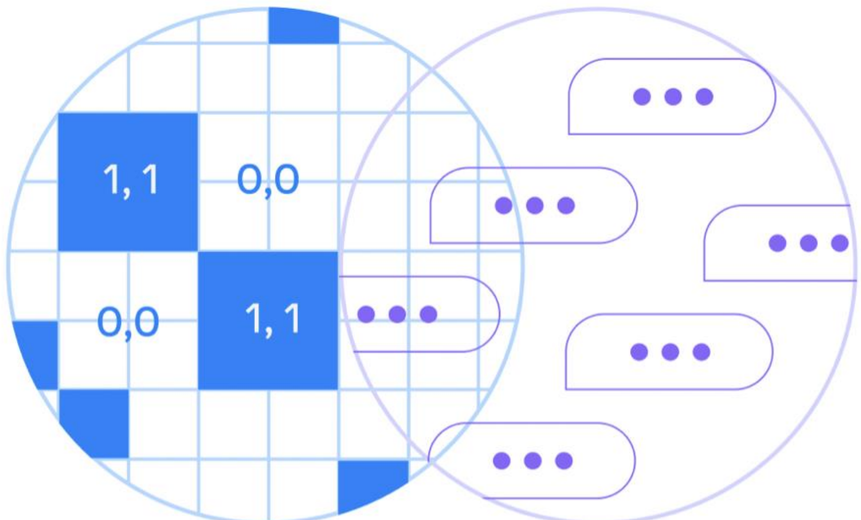
Meta AI presents CICERO — the first AI to achieve human-level performance in Diplomacy, a strategy game which requires building trust, negotiating and cooperating with multiple players.

Learn more about #CICERObyMetaAI: bit.ly/3GBwLzx



3. Cicero: “Free Energy” for a Game Theory of Mind II

The game of “Diplomacy” and the AI Cicero that learned to play like a human



Strategic Reasoning

CICERO predicts the moves other players are likely to make, as well as what moves they expect CICERO to make, and uses that information to create a strategic plan.

Natural Language Processing

CICERO grounds its conversations in a set of carefully chosen plans, so it can negotiate, offer advice, share information, and make agreements with other players.

3. Cicero: “Free Energy” for a Game Theory of Mind II

The game of “Diplomacy” and the AI Cicero that learned to play like a human

Fig. 6. **Successful dialogue examples.**

3. Cicero: “Free Energy” for a Game Theory of Mind II

The game of “Diplomacy” and the AI Cicero that learned to play like a human

What does Meta AI’s Diplomacy-winning Cicero Mean for AI?

Hint: It’s not all about scaling

Gary Marcus and Ernest Davis
Nov 26

45 21

> “The distinguishing property of humans is to search for and to follow after truth.”
>
> – Cicero

3. Cicero: “Free Energy” for a Game Theory of Mind II

The game of “Diplomacy” and the AI Cicero that learned to play like a human

What does Meta AI’s Diplomacy-winning Cicero Mean for AI?

Hint: It’s not all about scaling

Gary Marcus and Ernest Davis
Nov 26

45 21

> “The distinguishing property of humans is to search for and to follow after truth.”
>
> – Cicero

Diplomacy, a complex game that requires extensive communication, has been recognized as a challenge for AI for at least fifty years. To win, a player must not only play strategically, but form alliances, negotiate, persuade, threaten, and occasionally deceive. It therefore presents challenges for AI that are go far beyond those faced either by systems that play games like Go and chess or by chatbots that engage in dialog in less complex settings.

3. Cicero: “Free Energy” for a Game Theory of Mind II

RESEARCH ARTICLE

Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning

META FUNDAMENTAL AI RESEARCH DIPLOMACY TEAM (FAIR)†, ANTON BAKHTIN, NOAM BROWN, EMILY DINAN, GABRIELE FARINA, COLIN FLAHERTY, DANIEL FRIED, ANDREW GOFF, JONATHAN GRAY, [...], AND MARKUS ZIJLSTRA +17 authors Authors Info & Affiliations

SCIENCE • 22 Nov 2022 • First Release • DOI: 10.1126/science.ade9097

piKL: KL-regularized planning

piKL assumes player i seeks a policy π_i that maximizes the modified utility function

$$U_i(\pi_i, \pi_{-i}) = u_i(\pi_i, \pi_{-i}) - \lambda D_{KL}(\pi_i \parallel \tau_i) \tag{1}$$

where π_{-i} represents the policies of all players other than i, and $u_i(\pi_i, \pi_{-i})$ is the expected value of π_i given that other players play π_{-i}. Specifically, let $Q_i^{t-1}(a_i) = u_i(a_i, \pi_{-i}^{t-i})$ and let

$$\pi_i^{\Delta t}(a_i) \propto \tau_i(a_i) \exp\left[\frac{Q_i^{t-1}(a_i)}{\lambda}\right] \tag{2}$$

On each iteration t, piKL updates its prediction of the players’ joint policies to be

$$\pi^t = \left(\frac{t-1}{t}\right)\pi^{t-1} + \left(\frac{1}{t}\right)\pi^{\Delta t} \tag{3}$$

3. Cicero: “Free Energy” for a Game Theory of Mind II

RESEARCH ARTICLE

Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning

META FUNDAMENTAL AI RESEARCH DIPLOMACY TEAM (FAIR)†, ANTON BAKHTIN, NOAM BROWN, EMILY DINAN, GABRIELE FARINA, COLIN FLAHERTY, DANIEL FRIED, ANDREW GOFF, JONATHAN GRAY, [...], AND MARKUS ZIJLSTRA +17 authors Authors Info & Affiliations

SCIENCE • 22 Nov 2022 • First Release • DOI: 10.1126/science.ade9097

Cisero's equivalent of Free Energy is:

$$U(\pi_i, \pi_{-i}) = u(\pi_i, \pi_{-i}) + \lambda\, D_{KL}(\pi_i \mid \tau_i) \tag{12}$$

$$= u(\pi_i, \pi_{-i}) + \lambda \underbrace{(-\underbrace{H(\pi_i)}_{\text{Entropy}} + \underbrace{H(\pi_i, \tau_i)}_{\text{cross entropy}})}_{\text{Log–loss game}} \tag{13}$$

$$= \overbrace{u(\pi_i, \pi_{-i}) + \lambda(-\underbrace{H(\pi_i)}_{\text{Entropy}}}^{\text{(local) MaxEnt Game Theory}} + \underbrace{H(\pi_i, \tau_i)}_{\text{cross entropy}}) \tag{14}$$

3. Cicero: “Free Energy” for a Game Theory of Mind II

The takeaway for this section:

- Successful strategically interacting AIs are using a (simple) form of ToM
- Strategy is only one part of the solution
- Also need a language model, a filter, and human fine-tuning of opening gambit
- Requires grounding in an “objective” world
- And … only 1 player suspected they were interacting with an AI

4. Some Consequences of AI and ToM based Social Interactions

4. Some Consequences of AI and ToM based Social Interactions

Inverse Reinforcement Learning as the Algorithmic Basis for Theory of Mind: Current Methods and Open Problems

by Jaime Ruiz-Serra and Michael S. Harré *

Modelling and Simulation Research Group, School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia

* Author to whom correspondence should be addressed.

Algorithms **2023**, *16*(2), 68; https://doi.org/10.3390/a16020068 (registering DOI)

Received: 16 December 2022 / Revised: 13 January 2023 / Accepted: 16 January 2023 / Published: 19 January 2023

4. Some Consequences of AI and ToM based Social Interactions

Science
Current Issue First release papers Archive About Submit manu

Using large-scale experiments and machine learning to discover theories of human decision-making

JOSHUA C. PETERSON, DAVID D. BOURGIN, MAYANK AGRAWAL, DANIEL REICHMAN, AND THOMAS L. GRIFFITHS Authors Info & Affiliations

SCIENCE • 11 Jun 2021 • Vol 372, Issue 6547 • pp. 1209-1214 • DOI: 10.1126/science.abe2629

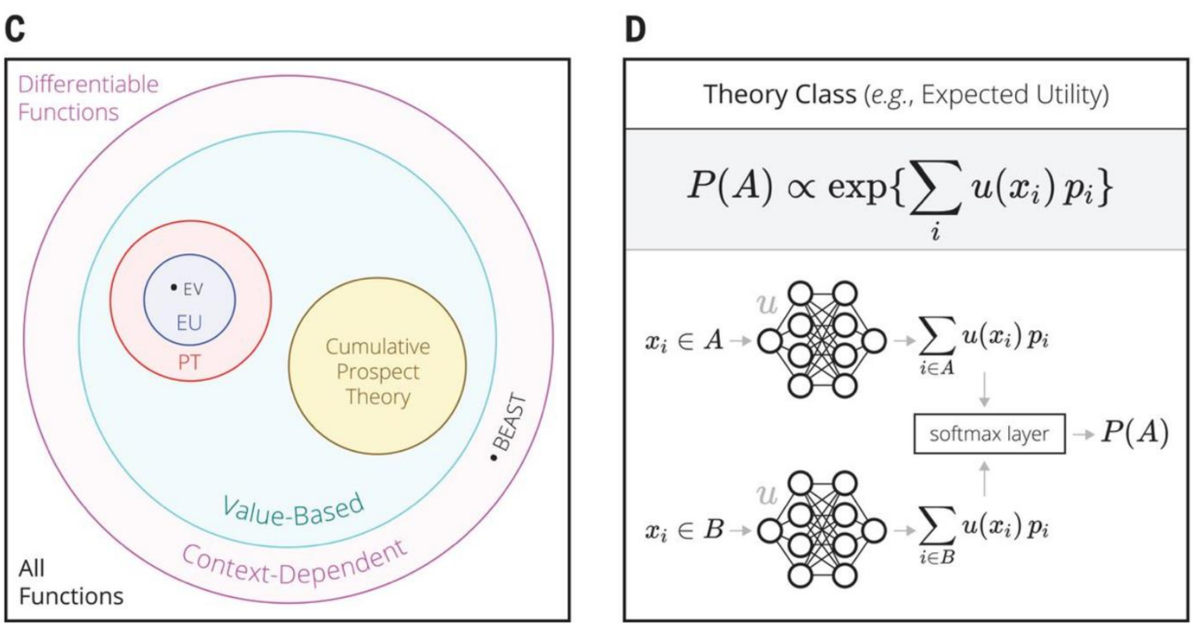


Fig. 1 Applying large-scale experimentation and theory-driven machine learning to risky choice.

4. Some Consequences of AI and ToM based Social Interactions

Theory of Mind is only one aspect

- Cicero has a complex, integrative, architecture
- Scaling up an AI doesn’t give us ToM by default: AI scaling hypothesis is probably wrong
- The human brain is modular with interacting functions

The modular and integrative functional architecture of the human brain

Maxwell A. Bertolero[a,b,1], B. T. Thomas Yeo[c,d,e,f], and Mark D’Esposito[a,b]

[a]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720; [b]Department of Psychology, University of California, Berkeley, CA 94720; [c]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077; [d]Clinical Imaging Research Centre, National University of Singapore, Singapore 117599; [e]Singapore Institute for Neurotechnology, National University of Singapore, Singapore 117456; and [f]Memory Networks Programme, National University of Singapore, Singapore 119077

Edited by Michael S. Gazzaniga, University of California, Santa Barbara, CA, and approved October 23, 2015 (received for review May 29, 2015)

" … we find a strong spatial correspondence between the cognitive functions and the network’s modules, suggesting that each module performs a discrete cognitive function. Crucially, activity at local nodes within the modules does not increase in tasks that require more cognitive functions, demonstrating the autonomy of modules’ functions. However, connector nodes do exhibit increased activity when more cognitive functions are engaged in a task."

4. Some Consequences of AI and ToM based Social Interactions

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

BIOLOGICAL SCIENCES

Open Access

Check for updates

View PDF

Tools

Share

Review articles

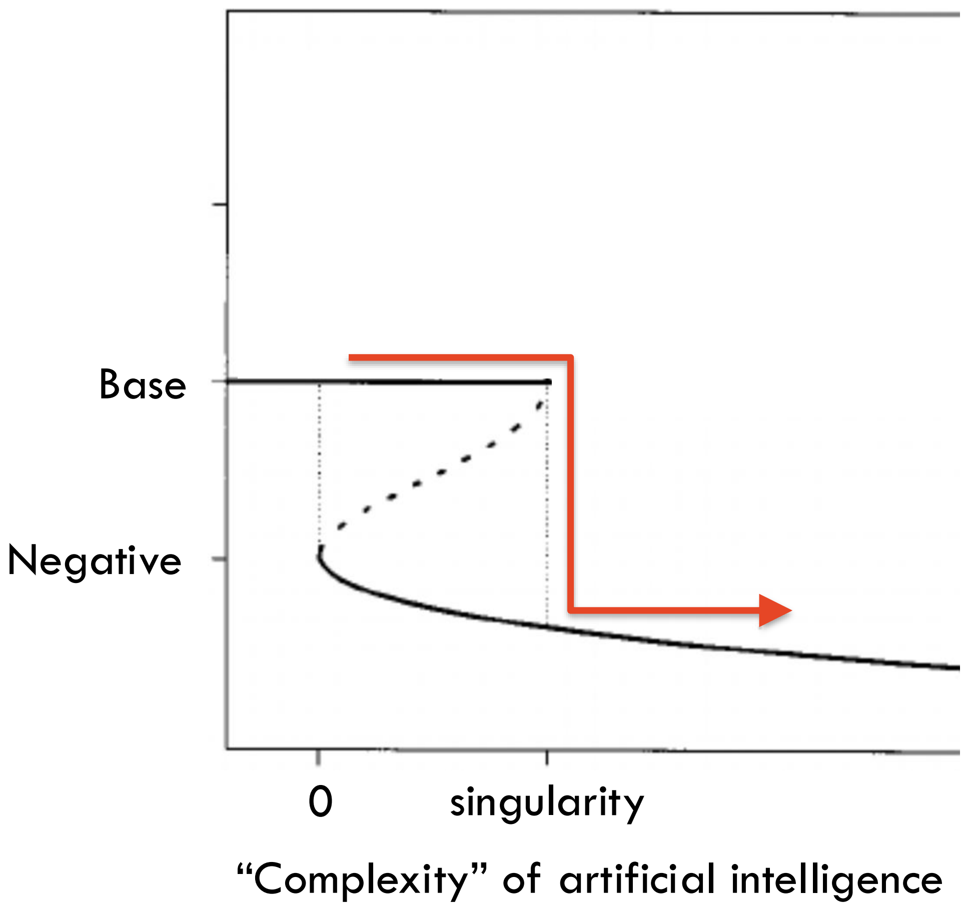
Collective minds: social network topology shapes collective cognition

Ida Momennejad

Published: 13 December 2021 | https://doi.org/10.1098/rstb.2020.0315

4. Some Consequences of AI and ToM based Social Interactions

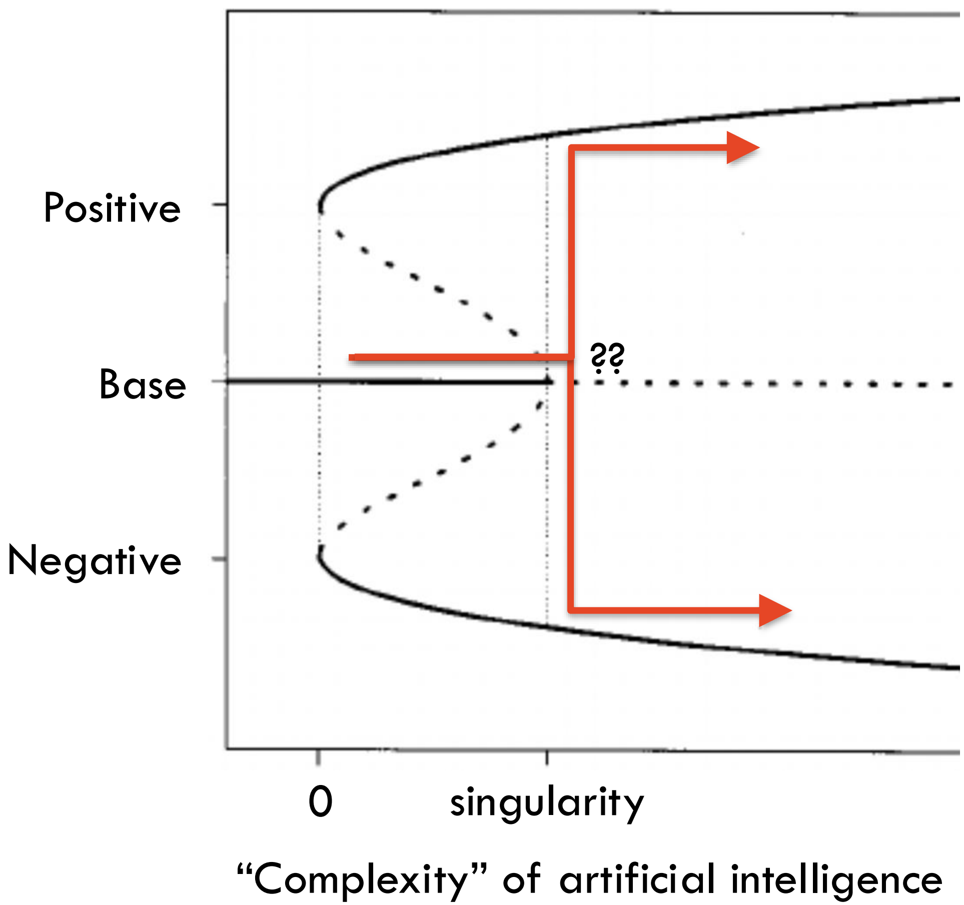
The pessimist view

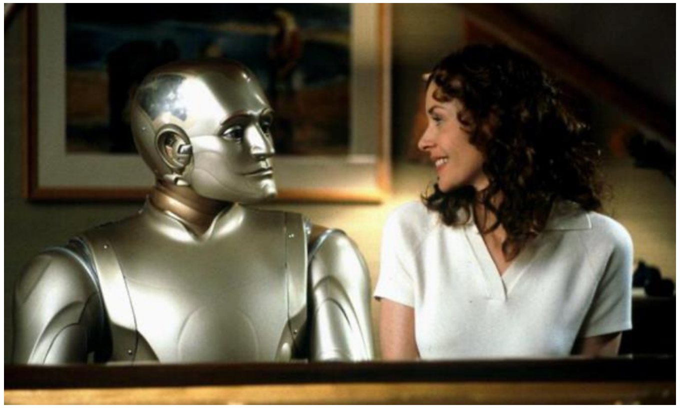


I, Robot (2004, 20th Century Fox)

4. Some Consequences of AI and ToM based Social Interactions

A bet both ways: “Sensitive intervention points” (J.D. Farmer et al, 2019)



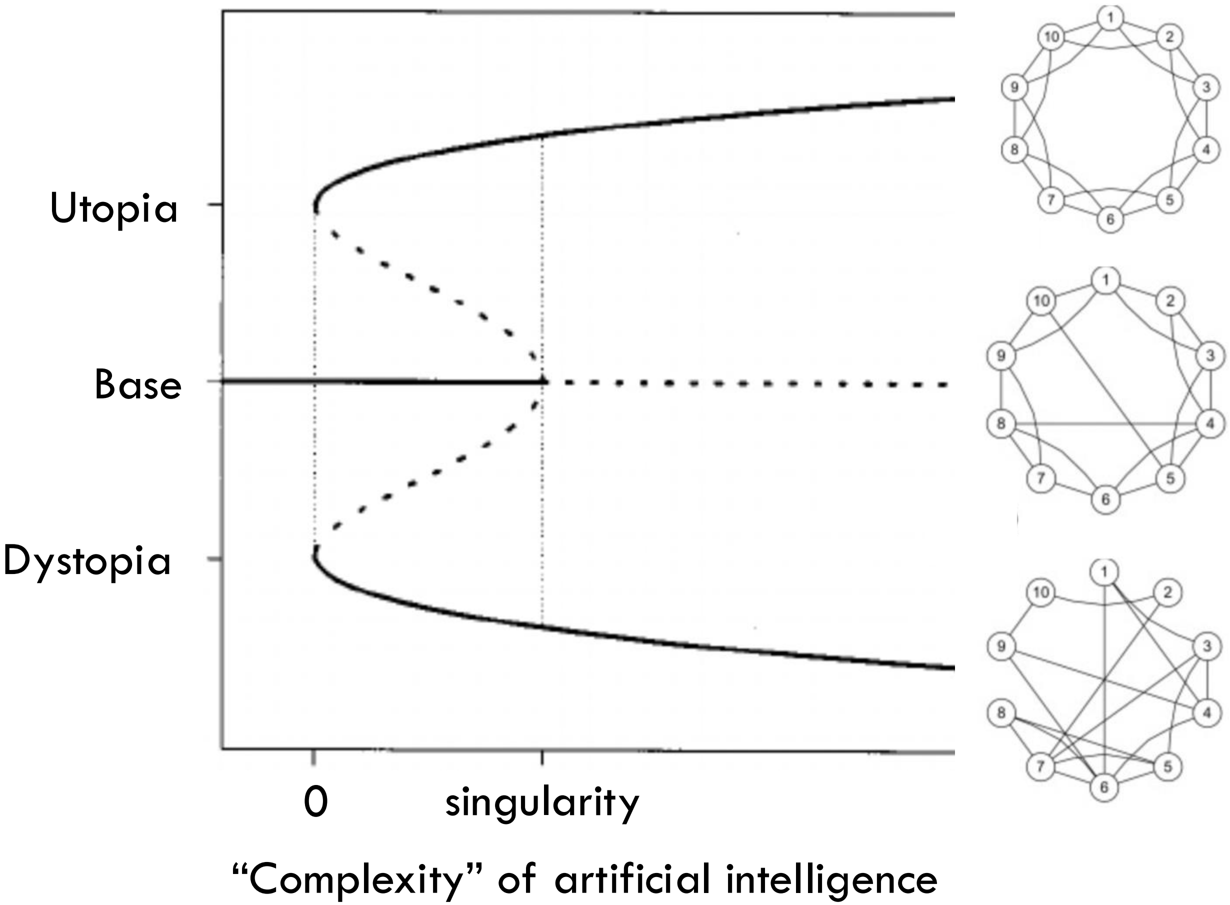


Bicentennial Man (1999 , Touchstone Pictures)

I, Robot (2004, 20th Century Fox)

4. Some Consequences of AI and ToM based Social Interactions

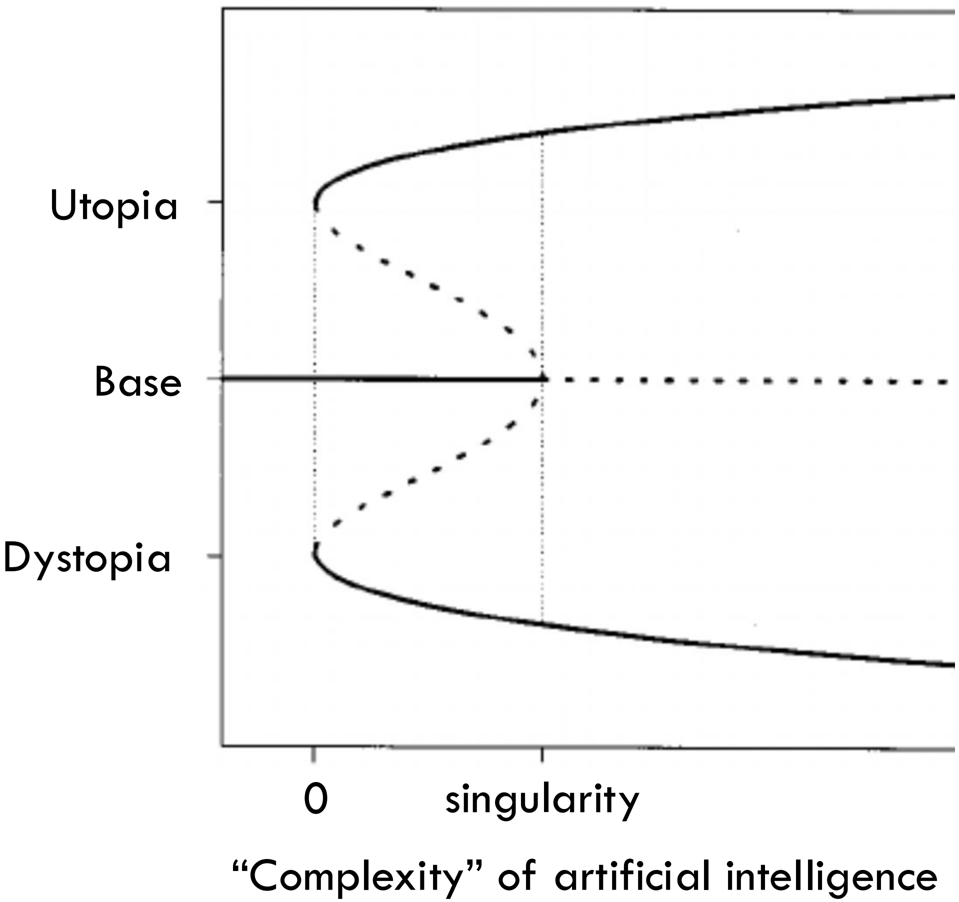
A bet both ways: "Sensitive intervention points" (J.D. Farmer et al, 2019)



Same agents,
different relationships,
different global outcomes

4. Some Consequences of AI and ToM based Social Interactions

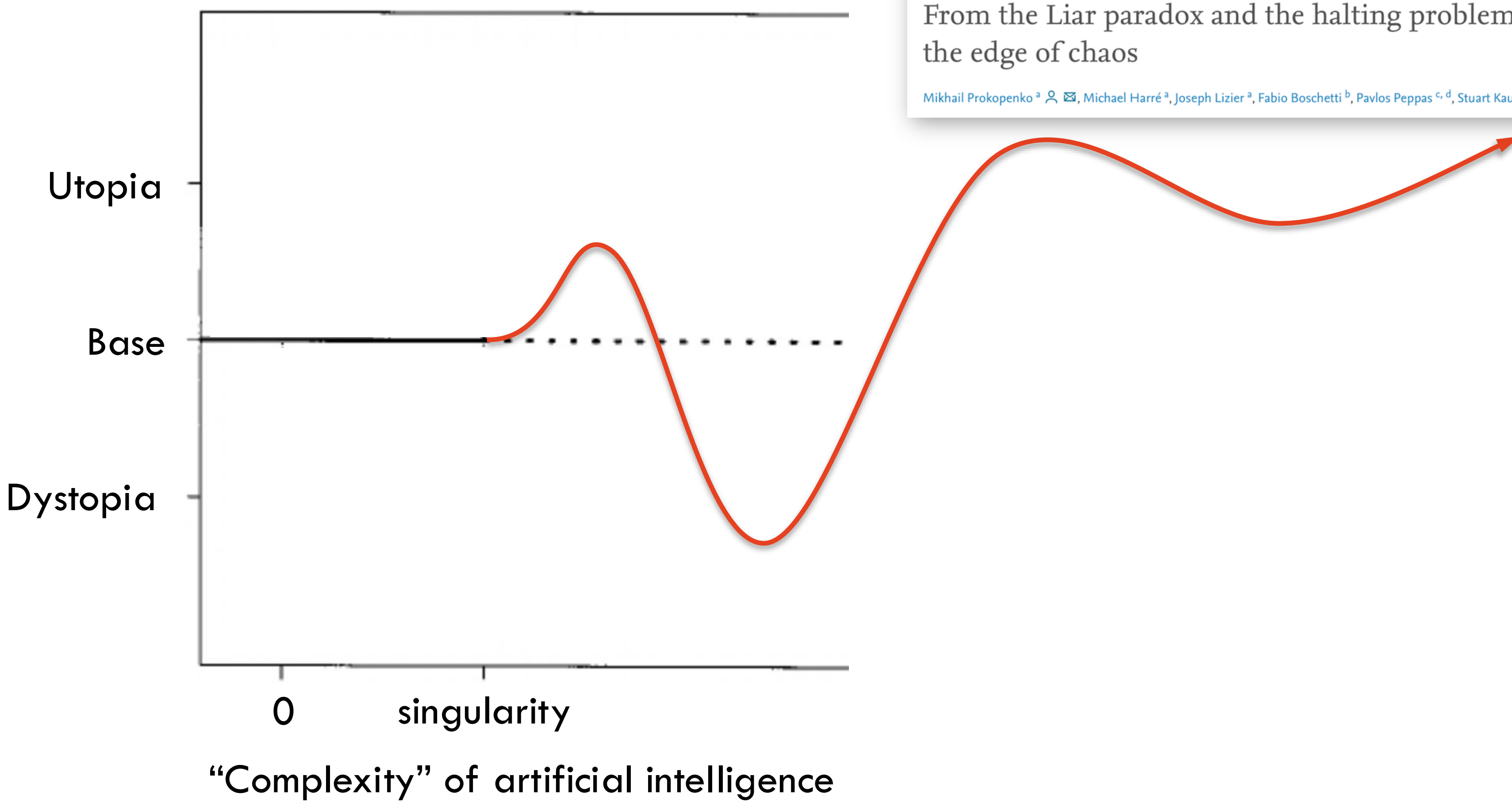
A bet both ways: "Sensitive intervention points" (J.D. Farmer et al, 2019)

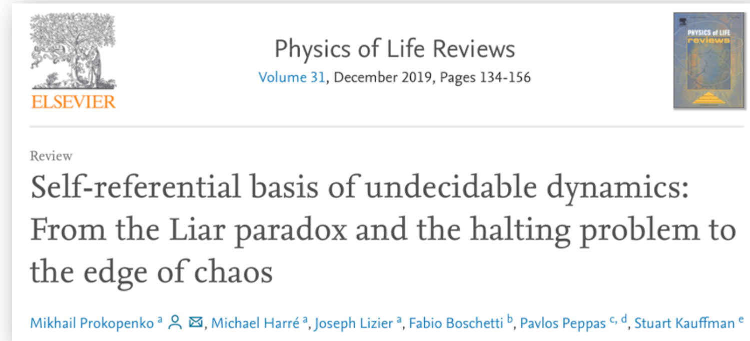


https://lithub.com/dystopia-is-realism-the-future-is-here-if-you-look-closely/
https://www.freepik.com/premium-photo/scifi-green-utopia-futuristic-city-environmentalism-concept-3d-art-illustration_32238700.htm

4. Some Consequences of AI and ToM based Social Interactions

The future is weird (not computable)




Physics of Life Reviews
Volume 31, December 2019, Pages 134-156

Review

Self-referential basis of undecidable dynamics: From the Liar paradox and the halting problem to the edge of chaos

Mikhail Prokopenko [a], Michael Harré [a], Joseph Lizier [a], Fabio Boschetti [b], Pavlos Peppas [c, d], Stuart Kauffman [e]

4. Some Consequences of AI and ToM based Social Interactions

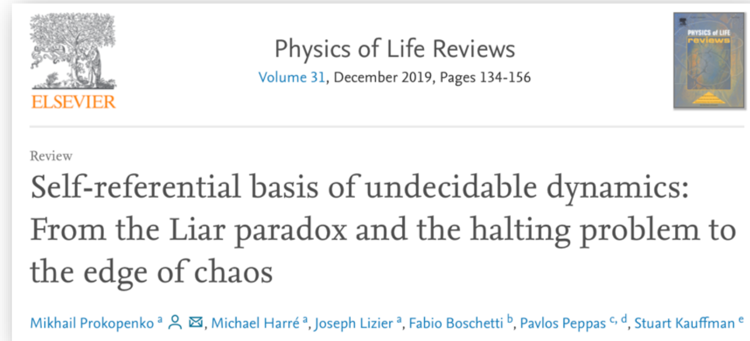

ELSEVIER
Physics of Life Reviews
Volume 31, December 2019, Pages 134-156

...ble dynamics:
...alting problem to

...ti[b], Pavlos Peppas[c, d], Stuart Kauffman[e]

4. Some Consequences of AI and ToM based Social Interactions

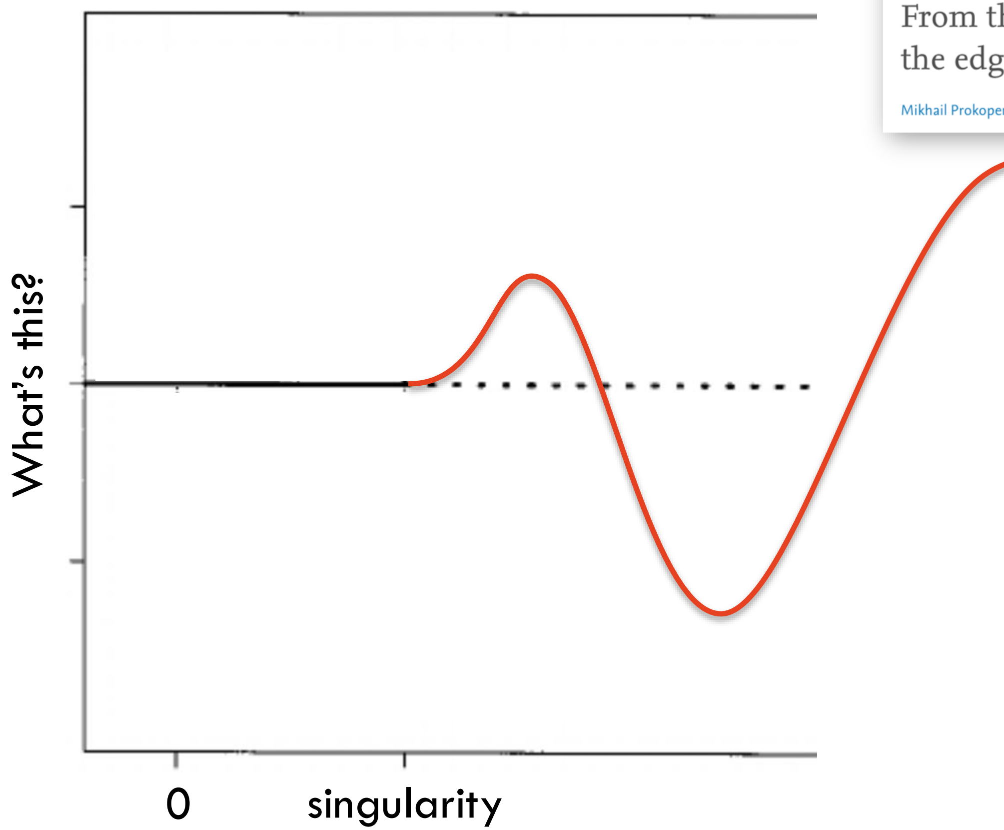
The future is weird (not computable)


Physics of Life Reviews
Volume 31, December 2019, Pages 134-156

Review

Self-referential basis of undecidable dynamics: From the Liar paradox and the halting problem to the edge of chaos

Mikhail Prokopenko [a], Michael Harré [a], Joseph Lizier [a], Fabio Boschetti [b], Pavlos Peppas [c, d], Stuart Kauffman [e]



References

1. Grünwald, P. D., & Dawid, A. P. (2004). *Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory*. Ann. Statist. 32 (4) 1367 - 1433, August 2004. https://doi.org/10.1214/009053604000000553 .
2. Friston, K. (2006). *A free energy principle for the brain*. **Journal of Physiology–Paris**, 100(1–3), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001
3. META, A Bakhtin, N Brown, E Dinan, et al. (2022). *Human-level play in the game of Diplomacy by combining language models with strategic reasoning*. **Science**, 378(6624), 1067–1074. https://doi.org/10.1126/science.ade9097
4. J. D. FARMER, C. HEPBURN, M. C. IVES, T. HALE, T. WETZER, P. MEALY, R. RAFATY, S. SRIVASTAV, AND R. WAY (2019). *Sensitive intervention points in the post-carbon transition.* **Science**, 364(6436), 132–134. https://doi.org/10.1126/science.aaw7287
5. Harré, M. S. (2025). From Firms to Computation: AI Governance and the Evolution of Institutions. arXiv preprint. https://arxiv.org/abs/2507.13616
6. Harré, M. S., Ruiz-Serra, J., & Drysdale, C. (2024). Artificial Theory of Mind and Self-Guided Social Organisation. arXiv preprint arXiv:2411.09169. https://arxiv.org/abs/2411.09169
7. Harré, M. S., Drysdale, C., & Ruiz-Serra, J. (2024). An AI Theory of Mind Will Enhance Our Collective Intelligence. arXiv preprint arXiv:2411.09168. https://arxiv.org/abs/2411.0916
8. Ruiz-Serra, J., Sweeney, P., & Harré, M. S. (2025). Factorised Active Inference for Strategic Multi-Agent Interactions. AAMAS '25 https://arxiv.org/abs/2411.07362
9. Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://www.nature.com/articles/nrn2787
10. Simon, H. A. (1997). Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations (4th ed.). Free Press. [Original work published 1947]
11. Rabinowitz, N. C., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine Theory of Mind. Proceedings of the 35th International Conference on Machine Learning (ICML 2018), 80, 4218–4227. https://proceedings.mlr.press/v80/rabinowitz18a.html
12. Kosinski, M. (2023). Theory of Mind May Have Spontaneously Emerged in Large Language Models. PNAS, 120(6), e2218523120. https://doi.org/10.1073/pnas.2218523120
13. Harré, M. S., & El-Tarifi, H. (2024). Testing Game Theory of Mind Models for Artificial Intelligence. Games, 15(1), 1. https://doi.org/10.3390/g15010001
14. Ruiz-Serra, J., & Harré, M. S. (2023). Inverse Reinforcement Learning as the Algorithmic Basis for Theory of Mind: Current Methods and Open Problems. Algorithms, 16(2), 68. https://doi.org/10.3390/a16020068
15. Harré, M. S. (2022). What Can Game Theory Tell Us About an AI 'Theory of Mind'? Games, 13(3), 46. https://doi.org/10.3390/g13030046
16. Harré, M. S., & Prokopenko, M. (2016). The social brain: Scale-invariant layering of Erdős–Rényi networks in small-scale human societies. Journal of the Royal Society Interface, 13(118), 20160044. https://doi.org/10.1098/rsif.2016.0044

Questions