Probability and the People

Edward Elliott, University of Notre Dame

8 October 2026

Slides

Introduction

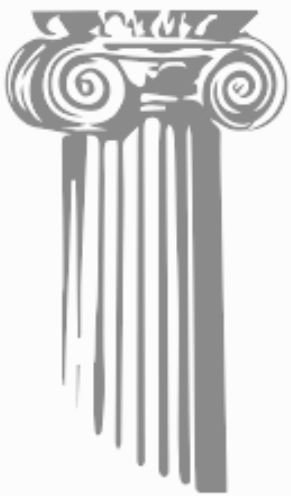
Decision theory (at least if we omit the frills) is not esoteric science… Rather, it is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference and choice. It is the very core of our commonsense theory of persons, dissected out and elegantly systematized.

(Lewis 1974: 337–8)

Introduction

Analytic Methodology

Folk-theoretic terms (e.g., ‘belief’, ‘desire’) are implicitly defined by the roles they occupy within folk theory



Bayesianism ≈ Folk Psychology

Bayesian models of belief and desire are close enough approximations to the folk psychological theory

Introduction

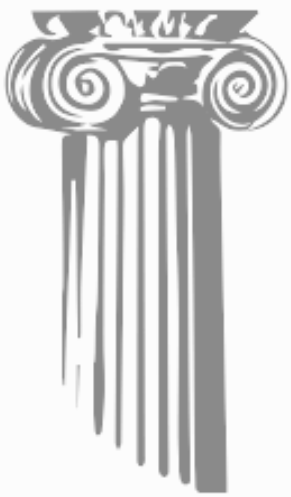
Solving the **location problem**...

1. Belief is whatever best satisfies the belief-role (*analytic content*)
2. **X** is what best satisfies the belief-role (*aposteriori discovery*)
3. Belief is **X** (*theoretical identification*)

Introduction

Analytic Methodology

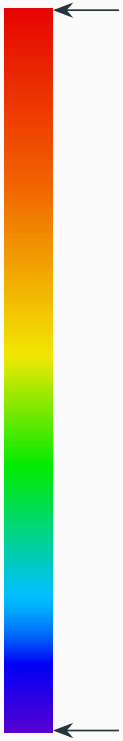
Folk-theoretic terms (e.g., 'belief', 'desire') are implicitly defined by the roles they occupy within folk theory



Bayesianism $\approx$ Folk Psychology

Bayesian models of belief and desire are close enough approximations to the folk psychological theory

Introduction



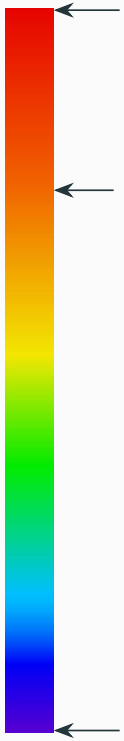
Incredulous stare

Folk psychology treats everyone as if an ideal (orthodox) Bayesian agent with perfectly precise and probabilisitcally coherent credences, who always conditionalizes correctly and always maximizes expected utility

Uncontroversial weaksauce

Bayesianism and folk psychology are broadly similar in overall structure, insofar as both have beliefy things and desirey things that together somehow determine behavior in a rationalish way

Introduction



Incredulous stare

Folk psychology treats everyone as if an ideal (orthodox) Bayesian agent with perfectly precise and probabilisitcally coherent credences, who always conditionalizes correctly and always maximizes expected utility

The good place

Much more similar than not, and similar enough in specific respects to be fruitfully employed for analytic purposes

Uncontroversial weaksauce

Bayesianism and folk psychology are broadly similar in overall structure, insofar as both have beliefy things and desirey things that together somehow determine behavior in a rationalish way

Introduction

Conceptual Mismatch
There seem to be differences between folk concepts and the technical concepts of Bayesianism — e.g., all-or-nothing belief and desire versus subjective probabilities and utility distributions

Applicability Mismatch
There seem to be differences between the imperfect agents to which folk psychology is normally applied versus the rationally idealized agents of standard Bayesian models

Introduction

Conceptual Mismatch
There seem to be differences between folk concepts and the technical concepts of Bayesianism — e.g., all-or-nothing belief and desire versus subjective probabilities and utility distributions

Applicability Mismatch
There seem to be differences between the imperfect agents to which folk psychology is normally applied versus the rationally idealized agents of standard Bayesian models

Introduction

Main claims...

1. There is a way to interpret what FP is — as what the folk typically believe, in a certain sense of 'belief', about intentional psychology — such that $\mathbf{B} \approx \mathbf{FP}$ is empirically plausible
2. This interpretation of FP is not *ad hoc*, but what advocates of $\mathbf{B} \approx \mathbf{FP}$ (such as Lewis) had in mind all along
3. Consequently, *some* of the skepticism towards $\mathbf{B} \approx \mathbf{FP}$ arises simply from a misunderstanding as to what's being said

Introduction

Main claims...

1. There is a way to interpret what FP is — as what the folk typically believe, in a certain sense of 'belief', about intentional psychology — such that $\mathbf{B} \approx \mathbf{FP}$ is empirically plausible
2. This interpretation of FP is not *ad hoc*, but what advocates of $\mathbf{B} \approx \mathbf{FP}$ (such as Lewis) had in mind all along
3. Consequently, *some* of the skepticism towards $\mathbf{B} \approx \mathbf{FP}$ arises simply from a misunderstanding as to what's being said

Introduction

Main claims...

1. There is a way to interpret what FP is — as what the folk typically believe, in a certain sense of 'belief', about intentional psychology — such that $\mathbf{B} \approx \mathbf{FP}$ is empirically plausible
2. This interpretation of FP is not *ad hoc*, but what advocates of $\mathbf{B} \approx \mathbf{FP}$ (such as Lewis) had in mind all along
3. Consequently, *some* of the skepticism towards $\mathbf{B} \approx \mathbf{FP}$ arises simply from a misunderstanding as to what's being said

Introduction

"YOU MUST UNLEARN WHAT YOU HAVE LEARNED"

1. Initial Clarifications

Initial Clarifications

Folk Psychology ≠ the psychology of the folk

- FP = the folk's *theory* of psychology, especially as it relates to intentional action, which might be inaccurate, incomplete, etc.
- **B ≈ FP** is *not* the thesis that Bayesian models closely approximate actual human psychology

Initial Clarifications

Folk Psychology ≠ conjunction of folk platitudes

> *Think of common-sense psychology as a term-introducing scientific theory… Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses… Add also all the platitudes to the effect that one mental state falls under another —* ***'toothache is a kind of pain'*** *and the like… I am going to claim that names of mental states derive their meaning from these platitudes.*
>
> (Lewis 1972: 256)

Initial Clarifications

Perhaps you and I disagree on what it is for someone to know the platitudes. He doesn't have to be able to state them, or even to know the technical vocabulary in which they might best be stated… It's enough if he could recognize them as being something he believed all along, when and if someone else formulated them and explained to him what they meant.*

(Lewis 2020 [1973]: 35–6)

Think of it as a theory: folk psychology. *It is common knowledge* among us; but it is tacit, as our grammatical knowledge is. We can tell which particular predictions and explanations conform to its principles, but we cannot expound those principles systematically. (Pace Lewis, 1972, eliciting the general principles of folk psychology is no mere matter of gathering platitudes.)*

(Lewis 1994: 416)

Initial Clarifications

Perhaps you and I disagree on what it is for someone to know the platitudes. He doesn't have to be able to state them, or even to know the technical vocabulary in which they might best be stated… It's enough if he could recognize them as being something he believed all along, when and if someone else formulated them and explained to him what they meant.*

(Lewis 2020 [1973]: 35–6)

Think of it as a theory: folk psychology*. It is common knowledge* among us; but it is tacit, as our grammatical knowledge is. We can tell which particular predictions and explanations conform to its principles, but we cannot expound those principles systematically. (Pace Lewis, 1972, eliciting the general principles of folk psychology is no mere matter of gathering platitudes.)*

(Lewis 1994: 416)

Initial Clarifications

> *A common theme among philosophers who give importance to conceptual analysis, including those associated with the Canberra Plan, is precisely that the claims that analysis seeks to isolate may not be readily accessible to ordinary speakers…*
>
> *[These claims] extend beyond the immediately intuitive and the banal. They constitute working assumptions to what competent speakers are committed by their shared argumentative and related habits and may go well beyond anything that is accessible without much thought.*
>
> (Pettit 2020: 336-8)

See also: Lewis 1974, Jackson & Pettit 1990, Jackson, Mason & Smith 2008, Pettit 1991, 1993, Braddon-Mitchell & Nola 2009

2. FP as what the folk (ethocentrically) believe

FP as what the folk believe

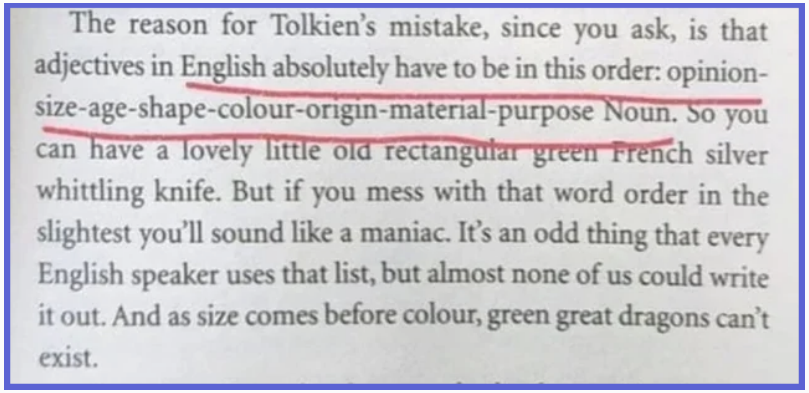
Key features of the grammar analogy:

1. Principles may be difficult to recognize as things already believed
2. Principles may conflict with what the folk explicitly and sincerely endorse
3. Principles may be expressed with unfamiliar technical vocabulary

FP as what the folk believe

Key features of the grammar analogy:

1. Principles may be difficult to recognize as things already believed
2. Principles may conflict with what the folk explicitly and sincerely endorse
3. Principles may be expressed with unfamiliar technical vocabulary

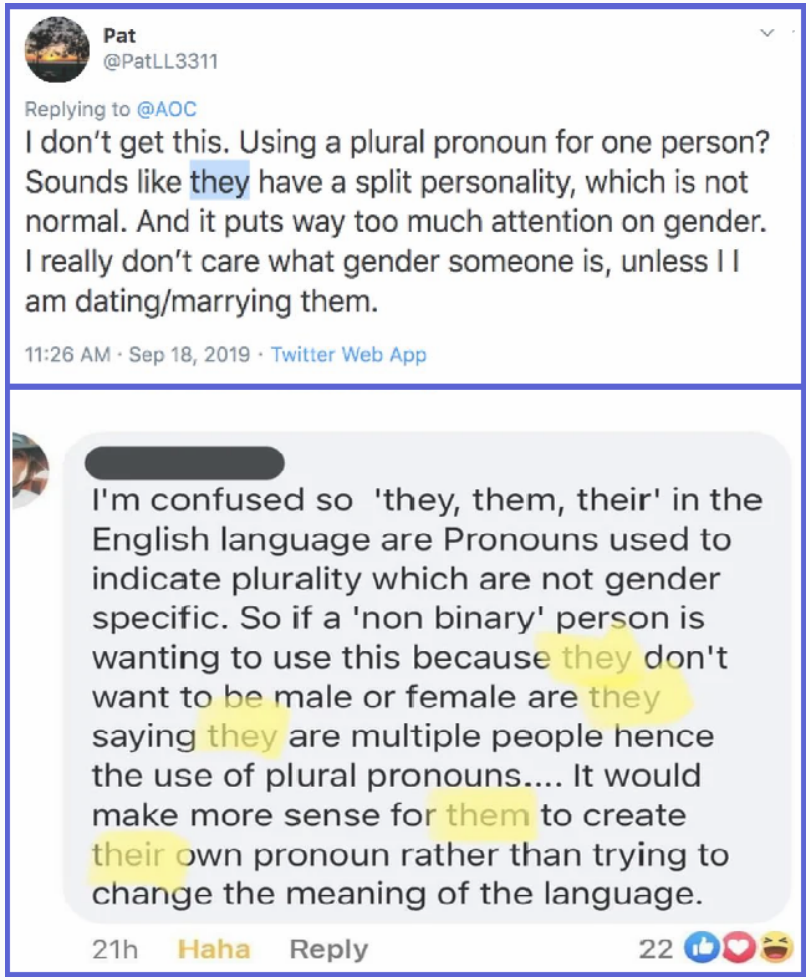


The reason for Tolkien's mistake, since you ask, is that adjectives in English absolutely have to be in this order: opinion-size-age-shape-colour-origin-material-purpose Noun. So you can have a lovely little old rectangular green French silver whittling knife. But if you mess with that word order in the slightest you'll sound like a maniac. It's an odd thing that every English speaker uses that list, but almost none of us could write it out. And as size comes before colour, green great dragons can't exist.

FP as what the folk believe

Key features of the grammar analogy:

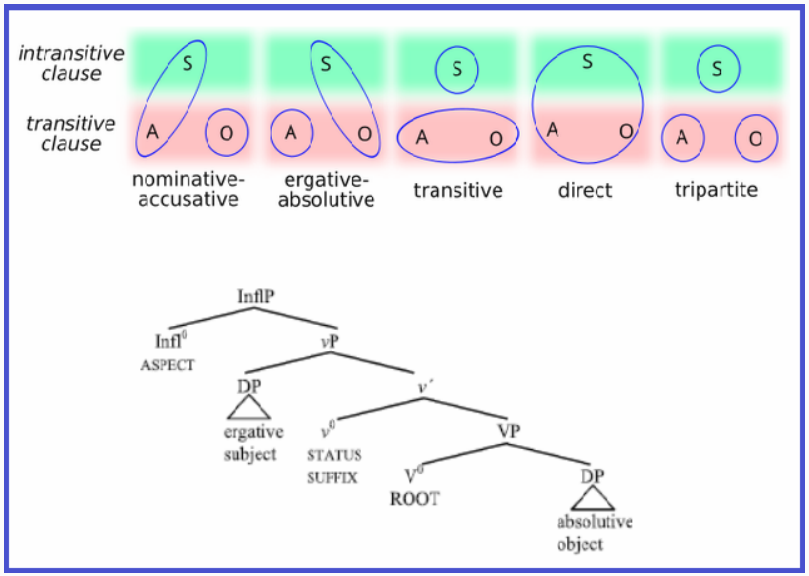
1. Principles may be difficult to recognize as things already believed
2. Principles may conflict with what the folk explicitly and sincerely endorse
3. Principles may be expressed with unfamiliar technical vocabulary



FP as what the folk believe

Key features of the grammar analogy:

1. Principles may be difficult to recognize as things already believed
2. Principles may conflict with what the folk explicitly and sincerely endorse
3. Principles may be expressed with unfamiliar technical vocabulary



FP as what the folk believe

Intellectualized belief

- Characterized by role in explicit reasoning and theorizing
- Particulate, beliefs as a relation to a sentence
- Readily available to conscious introspection, verbalizable
- Primarily evidenced by sincere assertion, or use as a premise

Ethocentric belief

- Characterized by role in relation to rational behavior
- Holistic, belief as "a map by which we steer"
- Often hidden from immediate awareness, non-verbalizable
- Primarily evidenced by unreflective practices and habits

FP as what the folk believe

Lewis's view...

- Folk psychology employs *two* conceptions of 'belief'
- The fundamental conception is holistic, probabilistic, explains rational behavior
- The derivative conception is intellectualized — where S is a sentence, an agent "believes S" iff S is appropriately related to their ethocentric beliefs
- The $\mathbf{B} \approx \mathbf{FP}$ hypothesis relates to the tacit theory of rational behavior

3. What do the folk (ethocentrically) believe?

What do the folk (ethocentrically) believe?

Bayesian Inverse Planning (BIP) paradigm,
especially **Bayesian Theory of Mind** (BToM) models of interpretation...

- Interpreters employ a *generative model* of the interpretee, which...
  i. predicts behavior at a time given beliefs, desires and options
  ii. predicts change in beliefs over time given perceptual evidence
- Interpreter starts with initial probability distribution **p** regarding what the interpretee's beliefs and desires might be like
- Interpreter observes interpretee's actions and perceptual evidence, updates **p** until observations fit with expectations

Intuitively: the interpreter "inverts" their generative model of the interpretee

What do the folk (ethocentrically) believe?

Wide range of BToM models...

Simple models...

- decision-making is ordinary expected utility maximization
- belief update is ordinary conditionalization ("Bayes' rule")

Complex models...

- decision-making is Boltzmann softmax, satisficing rules, etc.
- belief update is imperfect (e.g., Monte Carlo approximations)

What do the folk (ethocentrically) believe?

The Bayesian Theory of Mind

- **No:** The Bayesian generative model $\mathcal{M}$ is a scientifically accurate representation of ordinary human psychology
- **Yes:** At a suitable level of abstraction, the cognitive processes underlying interpretation can be represented as the "inversion" of a Bayesian generative model $\mathcal{M}$

What do the folk (ethocentrically) believe?

Applied to...

- attribution of preferences, goals, policies
- attribution of intelligence
- attribution of emotion
- inference of norms from punishment
- children's performance at the false belief task
- interpretation of epistemic language
- interpretation of machine behavior
- developing human-like socially intelligent AI

4. The challenges revisited

The challenges revisited

Conceptual Mismatch
There seem to be differences between folk concepts and the technical concepts of Bayesianism — e.g., all-or-nothing belief and desire versus subjective probabilities and utility distributions

Response:

- Presupposes an 'intellectualized' conception of FP — e.g., in terms of platitudes, how the folk explicitly talk and think

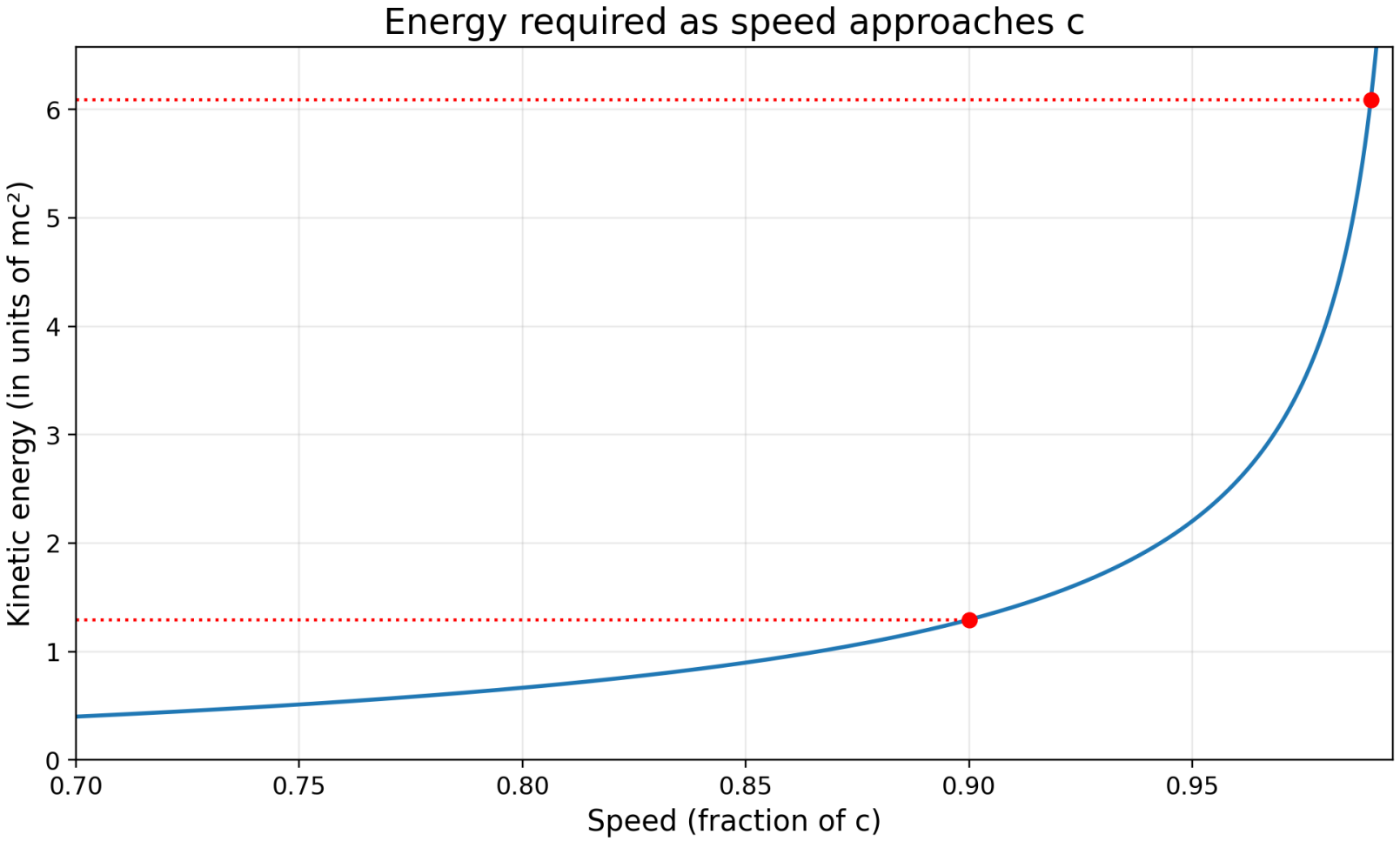
The challenges revisited

Applicability Mismatch
There seem to be differences between the imperfect agents to which folk psychology is normally applied versus the rationally idealized agents of standard Bayesian models

Responses:

1. BToM models are typically non-normative, but in all cases recognizably *Bayesian* in structure
2. Ordinary agents aren't so far from the normative ideal

The challenges revisited



The challenges revisited

Ordinary agents aren't so far from the normative ideal

1. Confusion about what the literature on human rationality establishes:
   Lots of evidence of irrationality ≠ evidence of lots of irrationality
2. Most examples of ordinary human rationality fly under the radar
   - perceptual learning
   - successful communication
   - social navigation
   - ordinary decision-making

Bibliography

Philosophy

- Braddon-Mitchell, D. and R. Nola (2009b). Introducing the Canberra Plan. In D. Braddon-Mitchell and R. Nola (Eds.), *Conceptual Analysis and Philosophical Naturalism*, pp. 1–20. MIT Press.
- Jackson, F., K. Mason, and S. Stich (2008). Folk Psychology and Tacit Theories: A Correspondence between Frank Jackson, and Steve Stich and Kelby Mason. In D. Braddon-Mitchell and R. Nola (Eds.), *Conceptual Analysis and Philosophical Naturalism*, pp. 45–98. Oxford University Press.
- Jackson, F. and P. Pettit (1990). In Defence of Folk Psychology. *Philosophical Studies* 59,31–54.
- Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* 50 (3), 249–258.
- Lewis, D. (1974). Radical interpretation. *Synthese* 27 (3), 331–344.
- Lewis, D. (1994). Reduction of Mind. In S. Guttenplan (Ed.), *Companion to the Philosophy of Mind*, pp. 412–431. Blackwell.
- Lewis, D. (2020). *Philosophical Letters of David K. Lewis, Volume 2: Mind, Language, Epistemology*. Oxford University Press.
- Pettit, P. (1991). Decision theory and folk psychology. In M. Bacharach and S. Hurley (Eds.), *Foundations of Decision Theory: Issues and Advances*, pp. 147–175. Oxford: Basil Blackwater.
- Pettit, P. (1993). *The Common Mind: an Essay on Psychology, Society and Politics*. Oxford University Press.
- Pettit, P. (2020). Analyzing Concepts and Allocating Referents. In A. Burgess, H. Cappelan, and D. Plunkett (Eds.), *Conceptual Engineering and Conceptual Ethics*, pp. 333–357. Oxford University Press.
- Yoda, M. (1980). *The Empire Strikes Back*. Dagobah: Lucasfilm.

Bibliography

Bayesian Inverse Planning, BToM

- Baker, C. L., J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behavior* 1, 1–10.
- Baker, C. L., R. Saxe, and J. B. Tenenbaum (2009). Action understanding as inverse planning. *Cognition* 113 (3), 329–349.
- Baker, C. L., R. R. Saxe, and J. B. Tenenbaum (2011). Bayesian Theory of Mind: Modeling Joint Belief–Desire Attribution. In L. Carlson, C. Hoelscher, and T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pp. 2469–2474. Austin, TX: Cognitive Science Society.
- Bergen, L., O. R. Evans, and J. B. Tenenbaum (2010). Learning Structured Preferences. *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society*, 853–858.
- Chuey, A. and H. Gweon (2025). Theory of Minds: Early Understanding of Interacting Minds. *Annual Review of Developmental Psychology* 7, 91–115.
- Diaconescu, A. O., C. Mathys, L. A.Weber, J. Daunizeau, L. Kasper, E. I. Lomakina, E. Fehr, and K. E. Stephan (2014). Inferring on the Intentions of Others by Heirarchical Bayesian Learning. *PLoS Computational Biology* 10 (9), 1–19.
- Goodman, N., C. L. Baker, E. Bonawitz, V. Mansinghka, A. Gopnik, H. Wellman, L. Schulz, and J. B. Tenenbaum (2006). Intuitive theories of mind: A rational approach to false belief. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*, 1382–1390.
- Goodman, N., C. L. Baker, and J. B. Tenenbaum (2009). Cause and intent: Social reasoning in causal learning. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*, 2759–2764.
- Hellou, M., S. Vinanzi, and A. Cangelosi (2023). Bayesian Theory of Mind for False Belief Understanding in Human-Robot Interaction. In *32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

Bibliography

Bayesian Inverse Planning, BToM (cont.)

- Houlihan, S., M. Kleiman-Weiner, L. Hewitt, J. B. Tenenbaum, and R. Saxe (2023). Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A* 381, 1–31.
- Jara-Ettinger, J., H. Gweon, L. E. Schulz, and J. B. Tenenbaum (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences* 20 (8), 589–604.
- Jern, A., C. G. Lucas, and C. Kemp (2017). People learn other people's preferences through inverse decision-making. *Cognition* 168, 46–64.
- Jin, C., Y. Wu, J. Cao, J. Xiang, Y. Kuo, Z. Hu, T. Ullman, A. Torralba, J. B. Tenenbaum, and T. Shu (2024). Mmtom-qa: Multimodal theory of mind question answering. In *62nd Annual Meeting of the Association for Computational Linguistics* (ACL).
- Kryven, M., T. Ullman, W. Cowan, and J. B. Tenenbaum (2016). Outcome or Strategy? A Bayesian Model of Intelligence Attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society* 38, 1–6.
- Lucas, C., T. Griffiths, F. Xu, and C. Fawcett (2009). A rational model of preference learning and choice prediction by children. *Advances in Neural Information Processing Systems* 21, 985–992.
- Radkani, S., J. B. Tenenbaum, and R. Saxe (2025). What people learn from punishment: A cognitive model. *Proceedings of the National Academy of Sciences* 122 (32), e2500730122.
- Rafferty, A. N., R. A. Jansen, and T. L. Griffiths (2020). Assessing Mathematics Misunderstandings via Bayesian Inverse Planning. *Cognitive Science* 44, 1–47.
- Ullman, T. D., J. B. Tenenbaum, C. L. Baker, O. Macindoe, O. R. Evans, and N. D. Goodman (2009). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems* 22, 1–9.
- Ying, L., T. Zhi-Xuan, L. Wong, V. Mansinghka, and J. B. Tenenbaum (2025). Understanding Epistemic Language with a Language-augmented Bayesian Theory of Mind. *Transactions of the Association for Computational Linguistics* 13, 613–637.