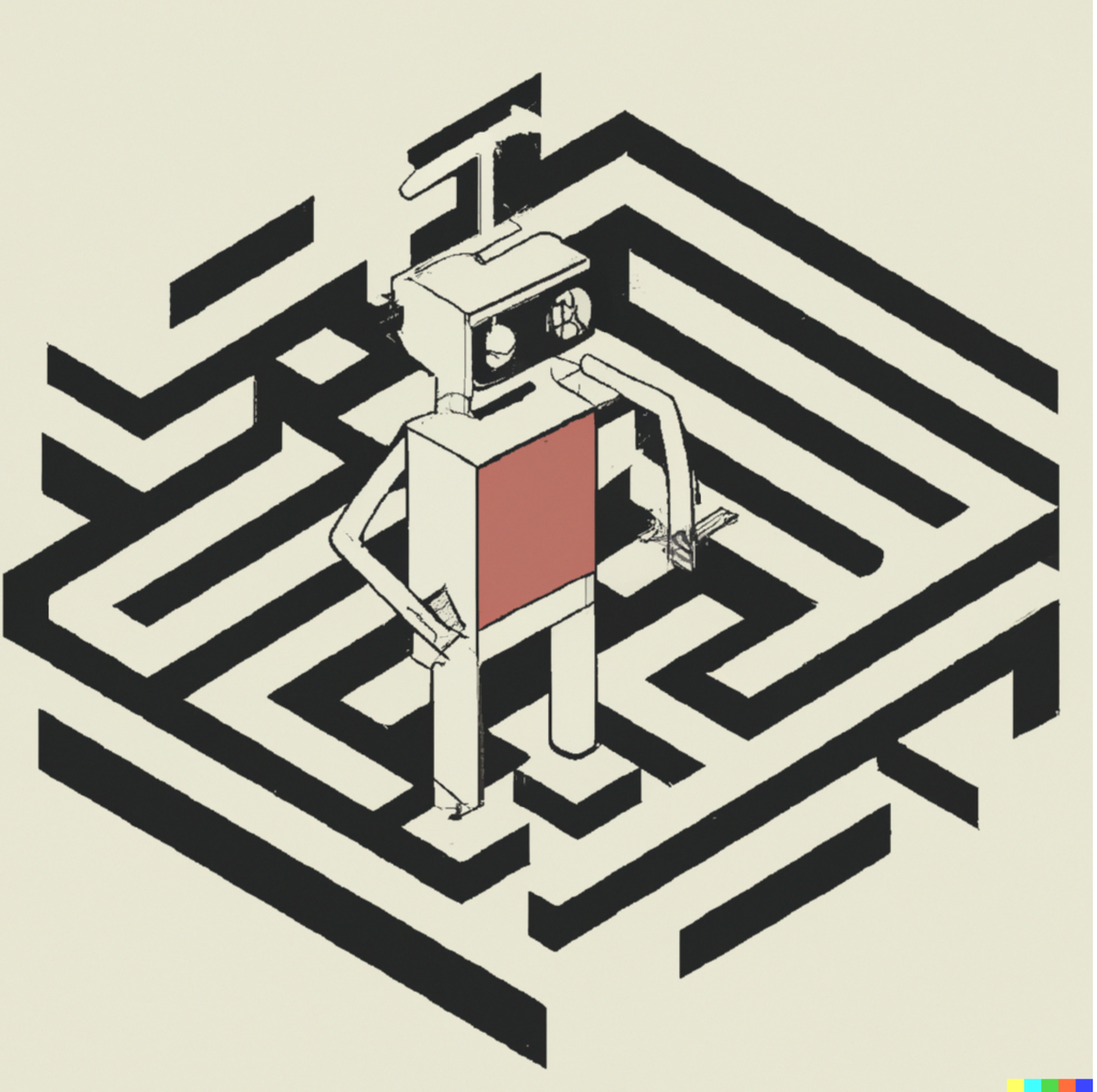


Seminar in Cognitive Modelling

Lecture 5 - Model Evaluation

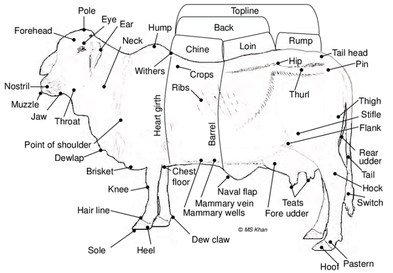
Neil Bramley

- **Semester 1:** Tour of cognitive modelling topics approaches themes and perspectives

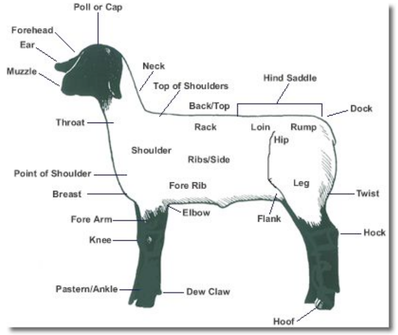
- **Semester 2:** Your turn…but also…

- …increasing emphasis on **critiquing the role of formal modeling in advancing science**

- "*All models are wrong, but some are useful*"
  — George Box



i.e. they are inductive, abstractions
bound to predict/explain their target less than perfectly (else just a clone/replica)



- More expansively: *"Since all models are wrong, the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad"* (Box, 1976)

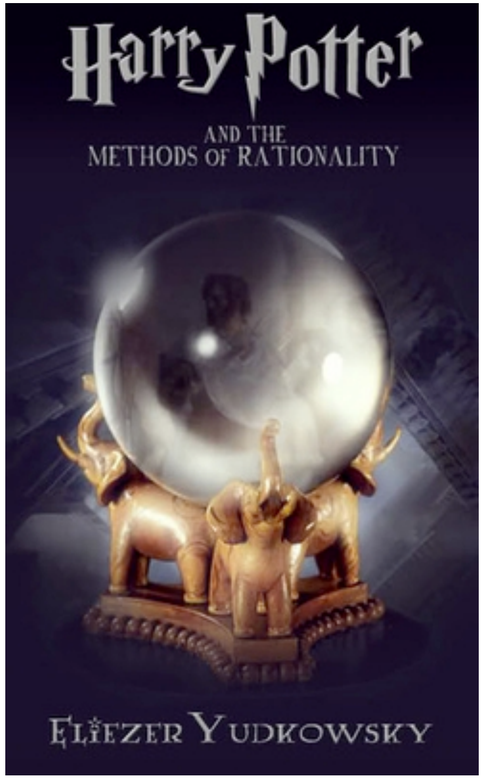
- **How *can we*, *should we*, *do we* use models in our scientific endeavours?**

Replication Crisis

- Since ~2010, Psychology research (& other areas inc. medicine) in “crisis” — many ostensibly “established” results are proving non-reproducible
    - e.g. 36/100 Psychology classic studies replicated, effect sizes 40±0.18% of original (Open Science Collaboration; 2015)
- **Are we doing science right?**

Effects in search of theory

- Non-replicable effects tend to lack a formalizable theory:
  - **“Power posing will make you act bolder”** (Carney et al, 2010) — Why? How much? In what circumstances?
  - **“Exposure to words pertaining to ageing will make you walk more slowly”** (Bargh et al, 1996) **-** What is the mechanism? What other predictions would this mechanism make if it were true?
  - **Pre-cognition “predicting stimuli from the future”** (Bem, 2011) **& Extra-Sensory Perception** (“morphic resonance”, Sheldrake) - Supernatural mechanisms, precluding systematic theorising



https://hpmor.com/

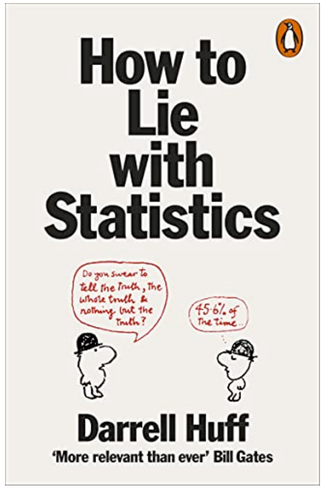
Open Science Movement 1.0

- Push toward making Open Science practices the default/norm e.g.:
  - Pre-registration — publishing plan for experiments and analyses ahead of running them (Hardwicke & Wagenmakers, 2023)
  - Including data & code with journal submissions (e.g. Hardwicke et al, 2018)
  - Using Bayesian statistics (Kruschke, 2010)
- Goal: guard against Questionable Research Practices e.g.:
  - HARK-ing: Hypothesising After Results are Known
  - P-hacking: making multiple comparisons or stopping decisions to force significance from statistical tests. Selective reporting/file drawer effect etc

Open Science Movement 2.0

But is this enough?

- How much does pre-registation & open-sourcing fix? (Szollosi et al, 2020)
  - Arguably, even easier to be bad-faith-Bayesian than frequentist…
  - & what is wrong with theorising being inspired by data?
- Isn't science *"inherently post hoc"*? (Shiffrin, Börner & Stigler, 2018)
- Perhaps issue is not just only at the "statistical analysis of data" level!
- **Need to improve the quality & rigour of our theorising** (cf Yarkoni, 2022)



How computational modeling can force theory building in psychological science

Olivia Guest
Research Centre on Interactive Media, Smart Systems and Emerging Technologies — RISE,
Nicosia, Cyprus &
Department of Experimental Psychology,
UCL, UK

Andrea E. Martin
Max Planck Institute for Psycholinguistics,
Nijmegen, The Netherlands &
Donders Centre for Cognitive Neuroimaging, Radboud University,
Nijmegen, The Netherlands

Psychology endeavors to develop theories of human capacities and behaviors based on a variety of methodologies and dependent measures. We argue that one of the most divisive factors in our field is whether researchers choose to employ computational modeling of theories (over and above data) during the scientific inference process. Modeling is undervalued, yet holds promise for advancing psychological science. The inherent demands of computational modeling guide us towards better science by forcing us to conceptually analyze, specify, and formalise intuitions which otherwise remain unexamined — what we dub "open theory". Constraining our inference process through modeling enables us to build explanatory and predictive theories. Herein, we present scientific inference in psychology as a path function, where each step shapes the next. Computational modeling can constrain these steps, thus advancing scientific inference over and above stewardship of experimental practice (e.g., preregistration). If psychology continues to eschew computational modeling, we predict more replicability "crises" and persistent failure at coherent theory-building. This is because without formal modeling we lack open and transparent theorising. We also explain how to formalise, specify, and imple-

Computational modelling as Open Theorising

1. Formal modelling makes explicit the pathway linking theories to evidence

2. Path constrains succession of conceptual moves from theory to experiment (guides what hypotheses to test, experiments to run, what to do with the results), facilitating progressive alignment of theories with reality

3. Lots of important science goes as we *articulate, run & refine* computational models

4. Skipping these steps risks burying inconsistencies, mistaking impact of evidence, motivated reasoning, leads to unfalsifiable theories & undermines progress

5. Therefore we should formalise our theories with models, to allow for explicit **Open Theorising**

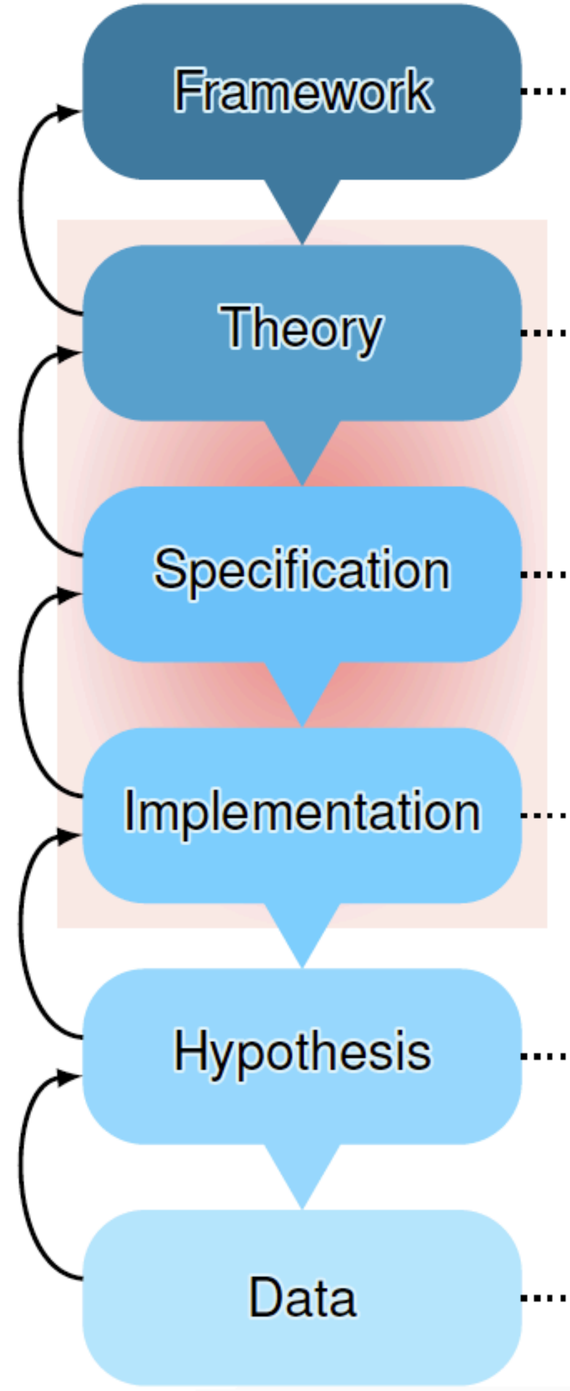
fwiw, I think this argument applies as much to theory development in AI as in Psychology

For us

- Guest & Martin's analysis a helpful framework for analysing role of models in papers (i.e. in essays & presentations, own research)
    - i.e. What scientific purpose does their formal modelling exercise serve?
    - What do they learned along the way?
    - What mistaken inferences might have occurred without the formalizing?
    - What sins occur nonetheless?

The account:

- Scientific enquiry analysed as a **path function*** connecting framework/theory to data
- Theory must pass through several states to gain explanatory force wrt Data
- Data must pass through same states in reverse to yield confirmatory/falsificatory force wrt a Theory
- Function expresses series of constraints on mappings (working downward), which then guides data-driven adjustment to content at each level (working upward)
- **Formalisability of a scientific research programme in this way determines its coherence hence its explanatory force**



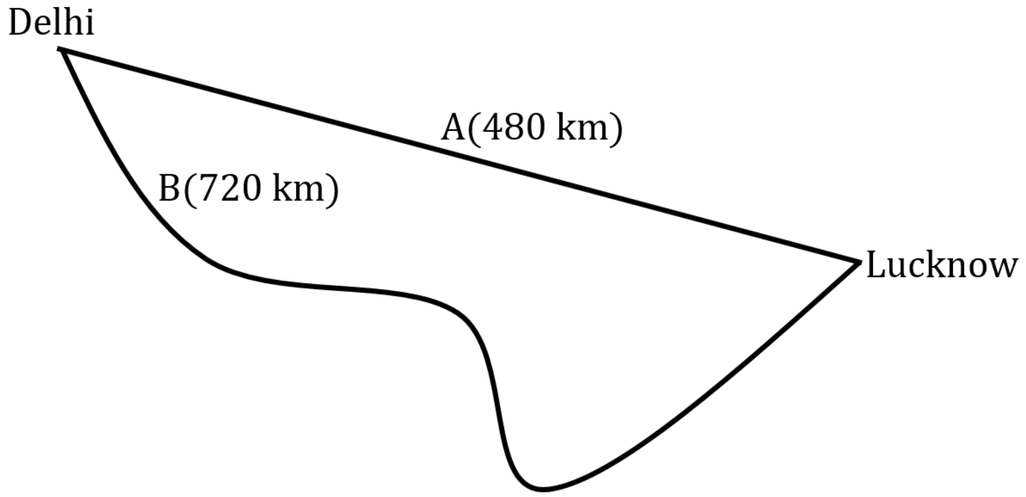
The account:

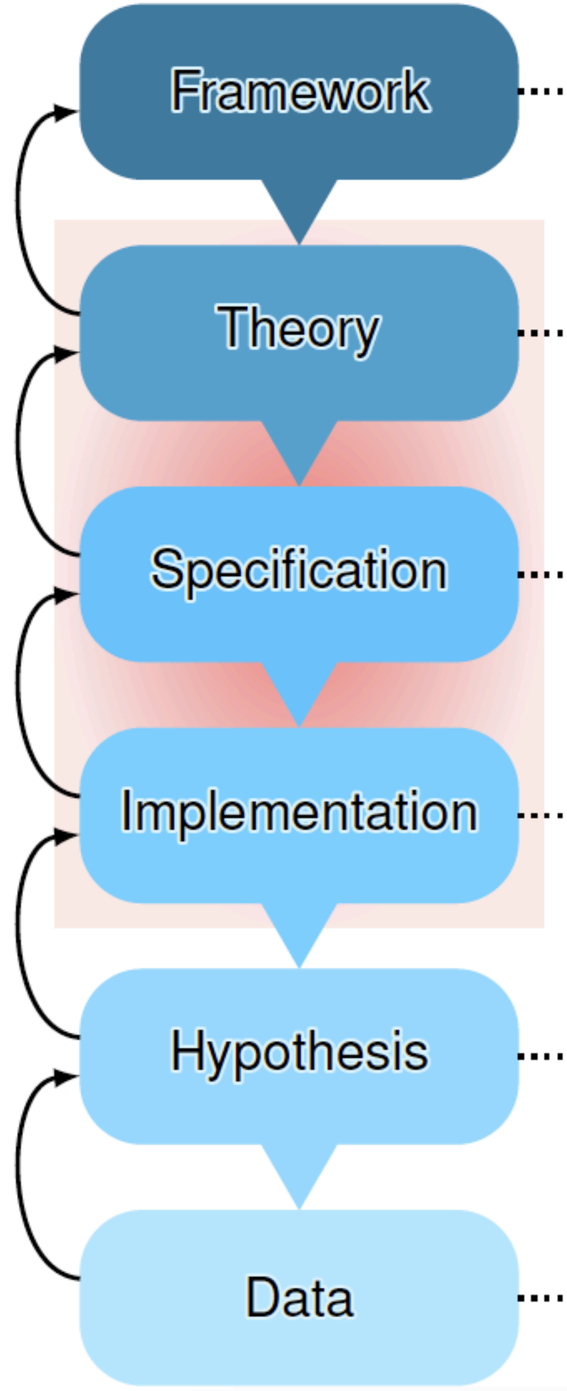
- *__path function -__ Output depends on path of transformations the input undergoes

- Distinct from more familiar *state* function, e.g.

  State function: distance(Delhi, Lucknow) -> 480km

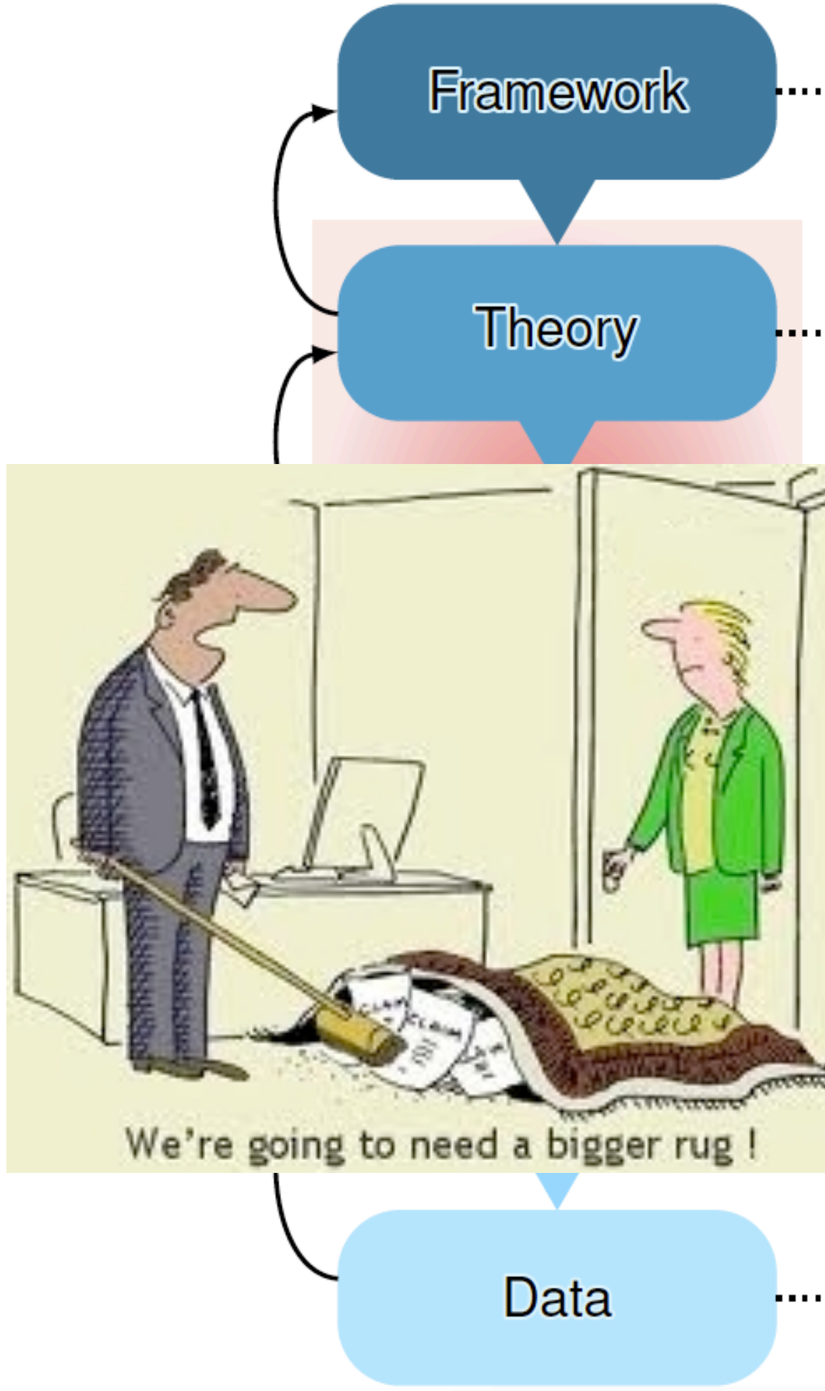
  Path function: travel_time(Delhi, Lucknow) -> Pick route, pick mode of transport, derive time





The account:

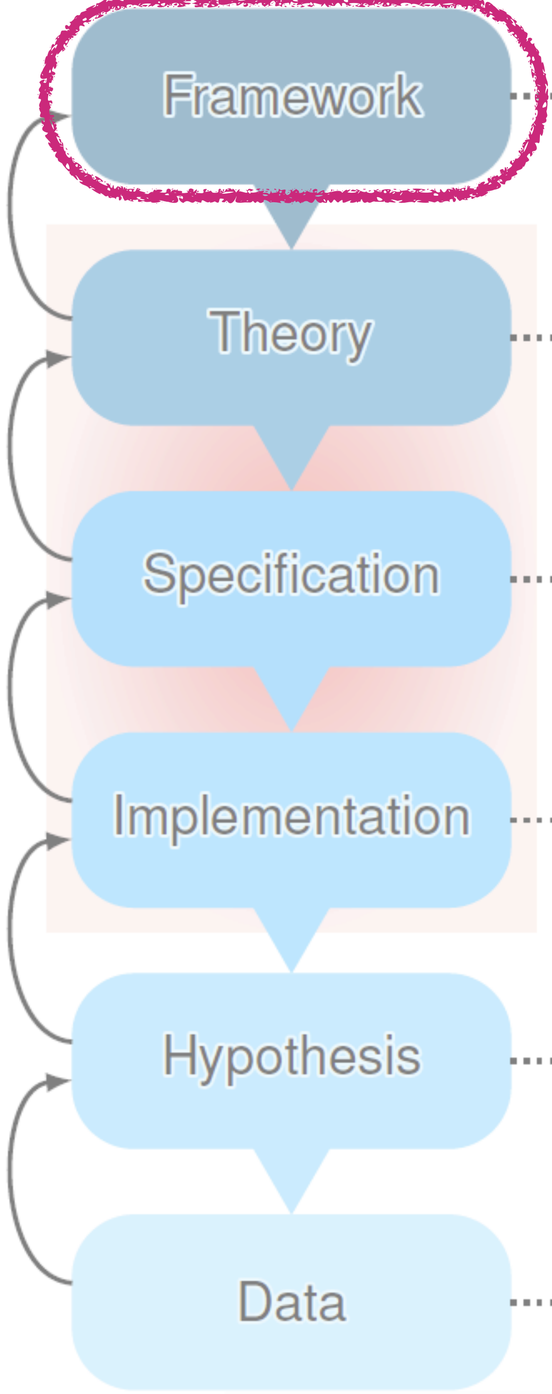
- In practice this depends on:
  1. How Theory is Specified
  2. How Specification is Implemented
  3. How Implementation is mapped to a Hypothesis
  4. And how Hypothesis is tested against Data



Framework

- Our meta-theoretical commitments e.g.:
- Connectionism:
  - Parallel processing
  - Distributed representations
- Bayesian cognitive models (/FEP if you prefer):
  - Hypothesis space
  - Prior beliefs
  - Bayesian updating
- Determines what theories we generate / entertain

Only testable extremely indirectly, beyond scope of a single paper

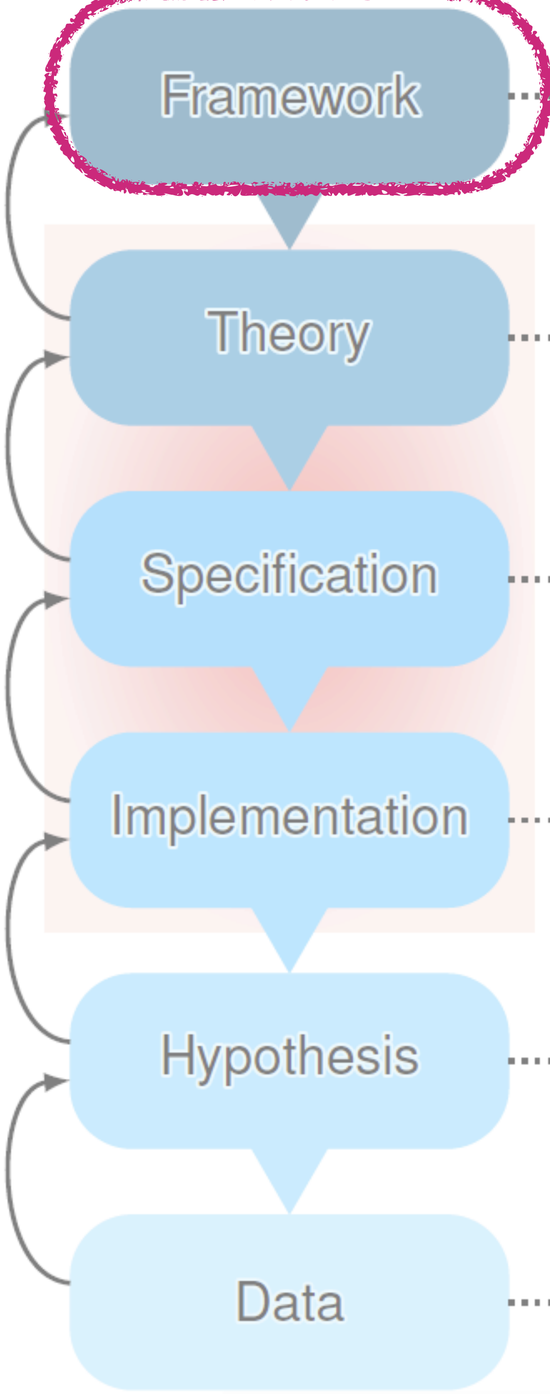


Theory

- *“a scientific proposition — described by a collection of natural language sentences, mathematics, logic, and figures”*
  - Prospect Theory (Kahneman, Tversky)
  - Rescorla Wagner Theory (of conditioning) (R&W)
  - Dual Process Theory (Evans, K & T)
  - Causal Model Theory (Rehder, Holyoak, Waldmann)
  - SUSTAIN Theory (of concepts) (Love, Gureckis)
  - (linguistic) Optimality Theory (Prince, Smolensky)
  - Evolutionary Theory (Darwin, Wallace)
  - Control Theory (Wiener, Kalman)
  - Behaviourism (Skinner, Watson)
  - Stages Theory (of development) (Piaget)
  - Theory Theory (of development) (Carey, Gopnik)
  - The Hierarchy of Needs (Maslow)
  - Attachment Theory (Bowlby, Ainsworth)

Live within some framework but this often left implicit

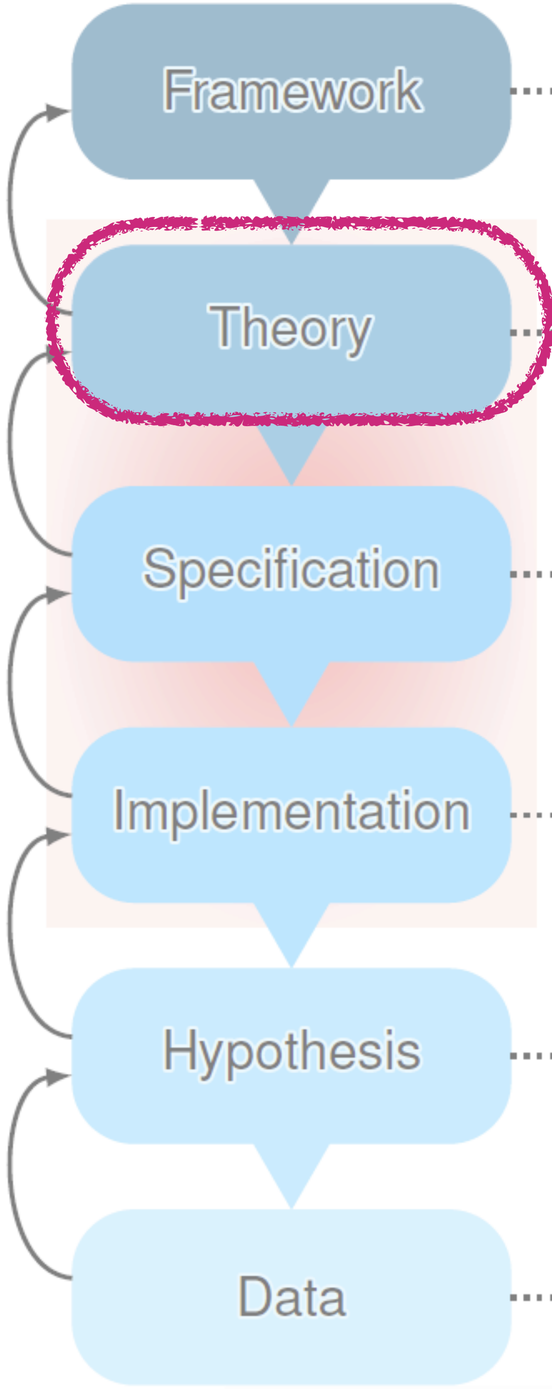
Bad scientific discourse level skips, leaving relationship between Data & Theories implicit at best, but more likely opaque or vague or false



Specification

- Must capture the core assumptions of the model
  - If implementation does not meet these requirements then it cannot be considered a valid implementation of the theory

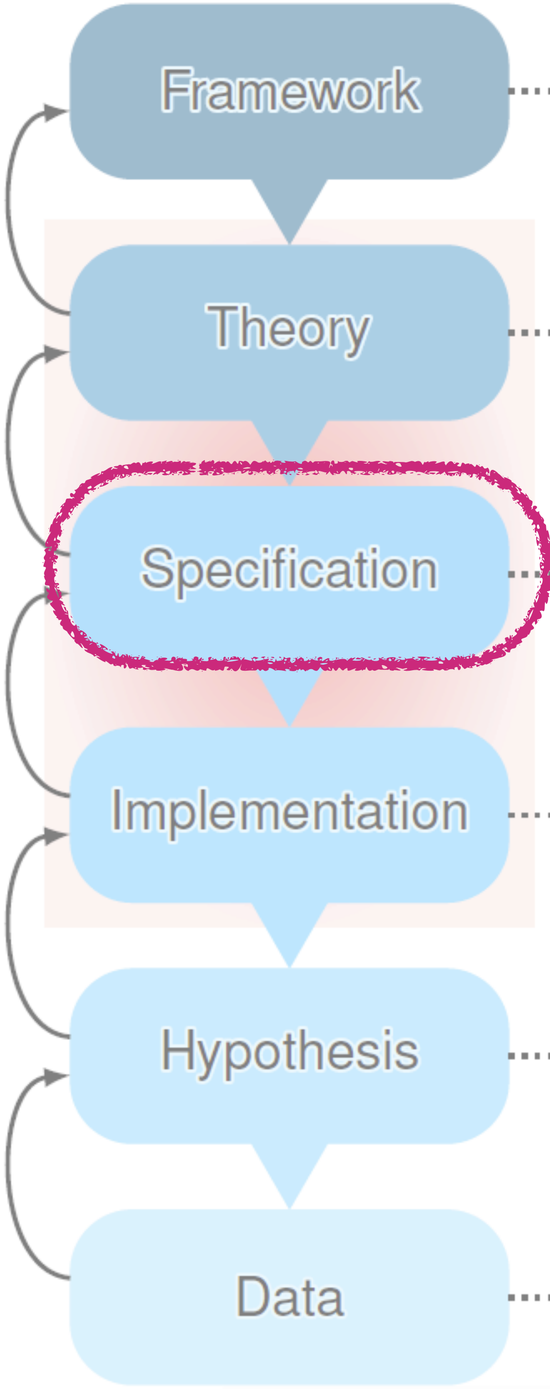
- Multiple potentially equally valid formal languages (equations, diagram, psuedocode, [perfectly unambiguous] natural language)



Implementation

- *“In psychology, creating an implementation typically involves taking the specification implicitly embedded in a journal article and writing code that is faithful to it.”*
- Auxiliary Assumptions
  - Insignificant commitments (e.g. written in Python)
  - Mutable commitments (e.g. noise is Gaussian)

- *“Without specifications we cannot debug our implementations, and we cannot properly test our theories”*
- If an implementation detail proves pivotal to what a model predicts, it must be upgraded to a specification detail



Hypothesis

“A narrow, testable statement… in psychology focus on properties of the world that can be measured and evaluated by collecting data and running inferential statistics”

Prevalence

- X occurs greater than chance

Relationship

- X is associated with Y

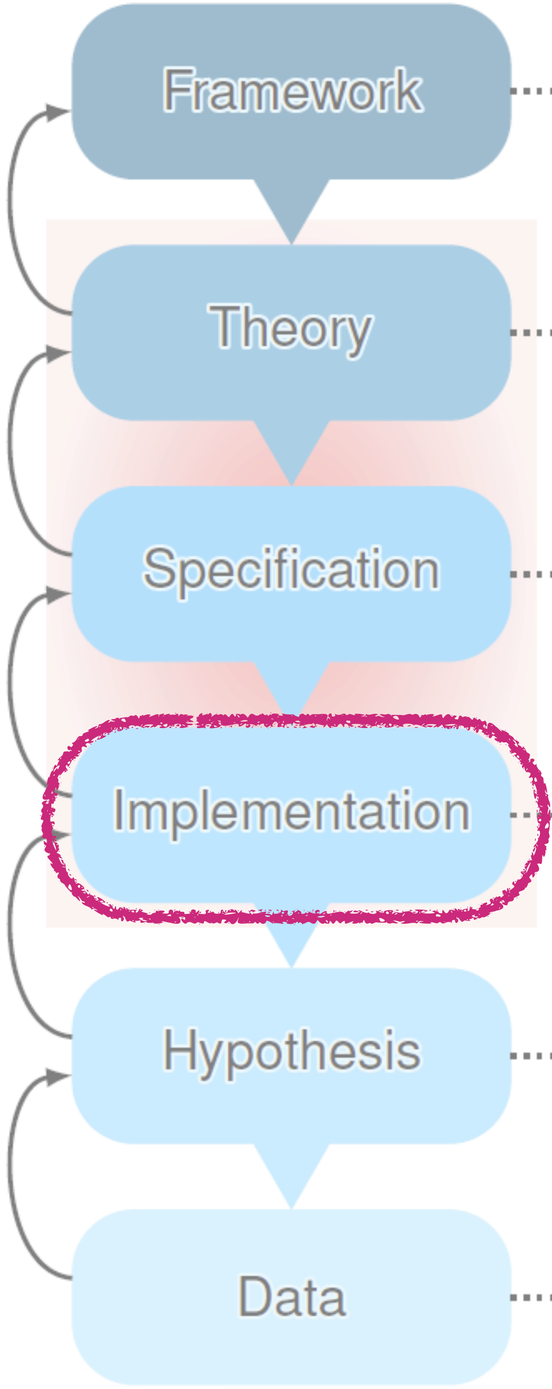
Causality

- X precedes Y
- X is sufficient to cause Y
- X is necessary to cause Y

Or perhaps:
“Model A is more likely than Model B”*

* e.g. Has higher complexity-penalised likelihood of producing data; lower cross validation error, smaller BIC etc

“Running our computational model’s code, allows us to generate hypotheses. For example, if our model behaves in a certain way in a given task, e.g., it has trouble categorising some types of visual stimuli more than others, we can formulate a hypothesis to test this.”

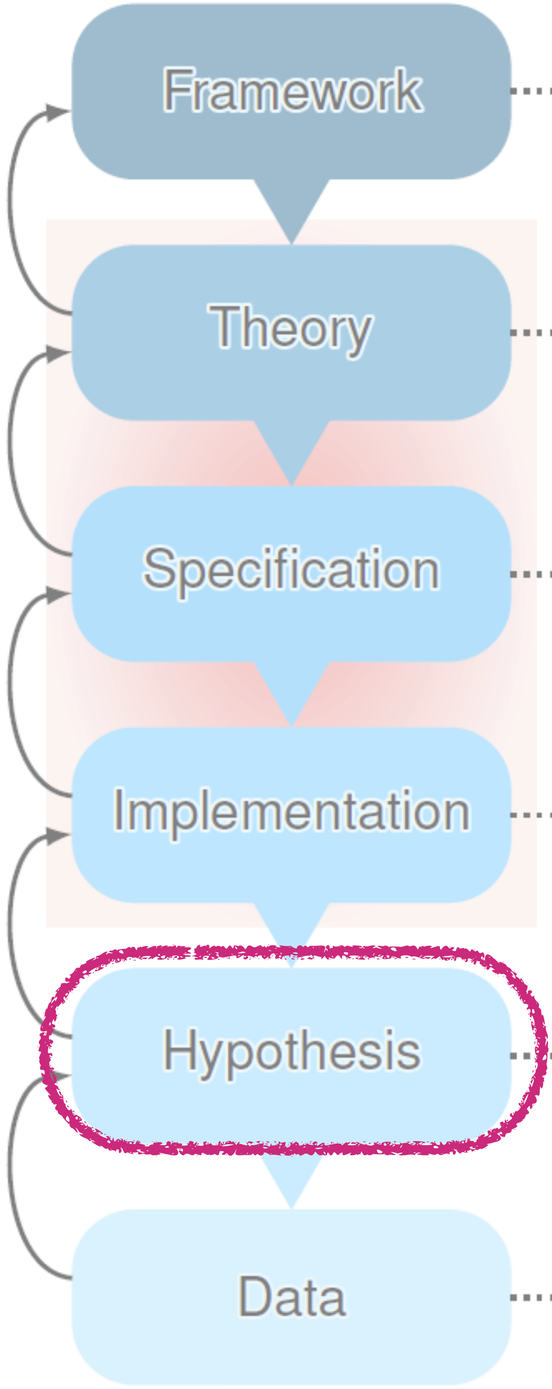


Data

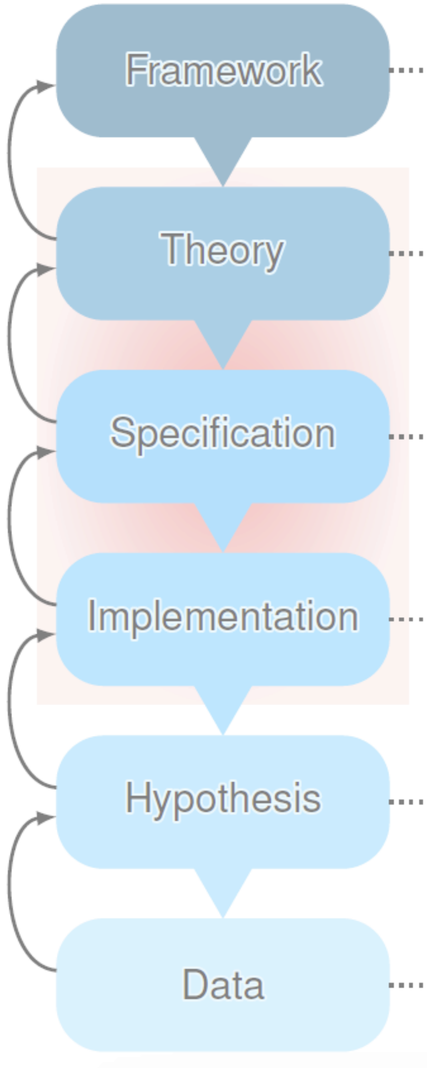
- Observations
- Simulations
- Experiments

- Regardless, data are not theory-neutral

- Measured/represented/encoded for some purpose, couched in some theoretical commitments, i.e. in order to test a hypotheses
- Their semantics are dependent on supporting theory
  - fMRI assumes link between blood-oxygenation & activation
  - behavioural responses depend on assumptions about participant's perceptions, motor control, motivations, task understanding, correctly functioning software etc (Szollosi et al, 2023)



Is *'Two 12 inch pizzas for the price of one 18 inch pizza'* a good deal?



Concepts of 'pizza', 'food', 'order'

T_0: 'number of pizzas corresponds to amount of pizza' **initial/naive**

T_1: 'the surface areas of the pizzas per order correspond to the amount of pizza' **posthoc/corrected**

$$\phi_i = \sum_{j=1}^{N} \pi R_j^2$$

```python
import numpy as np
import math

def food(ds):
    '''
    Amount of food in an order as a function
    of the diameters per pizza (eq. 3).
    '''
    return (math.pi * (ds/2)**2).sum()


# Order option a in fig. 1, two 12'' pizzas:
two_pizzas = np.array([12, 12])
```

H_0: 'two pizzas is more pizza than one pizza'
H_1: 'an 18 inch pizza is more pizza than two 12 inch pizzas'

Weigh pizzas **expectation violation!**

Summing up Guest & Martin

- By specifying, implementing and deriving hypotheses from theory, a research program becomes robust to the inevitable expectation violations
  - path provides multiple locations to adjust
  - Without, can only discard theory or ignore result
- "*Mathematically specifying and/or computationally implementing models, for example, can demonstrate that accounts are identical or overlap even when their verbal descriptions (i.e., informal specifications) are seemingly divergent.*"

Examples of revelations from formalisations

Psychological Review
2014, Vol. 121, No. 1, 1–32

0033-295X/14/$12.00 DOI: 10.1037/a0034190

Unfalsifiability and Mutual Translatability of Major Modeling Schemes for Choice Reaction Time

Matt Jones
University of Colorado, Boulder

Ehtibar N. Dzhafarov
Purdue University

e.g. modelling reveals several diffusion-based theories of choice are identical or unfalsifiable

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

& that various well known cognitive models are formally intractable

Successful modelling facilitates “open theorising”

- Complementary to other forms of “Open Science”
- Makes the commitments of theories explicit & distinguishes them from incidental implementational choices made while testing them
- *“More data—however open, will never solve the issue of a lack of formal theorising. Data cannot tell a scientific story, that role falls to theory and theory needs formalisation to be evaluated”*

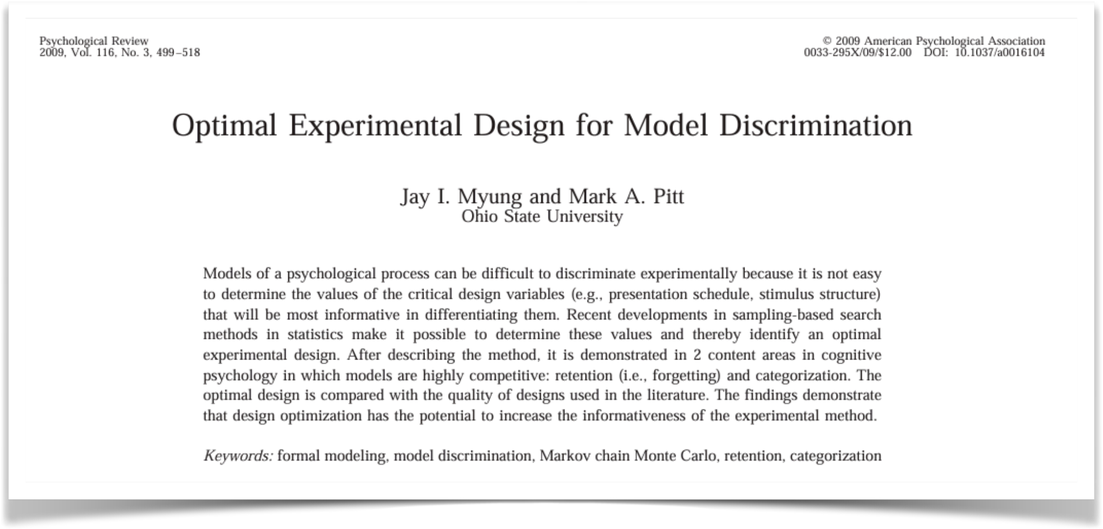
Relationship with Marr's levels of analysis?

- Similar in that it captures role of bidirectional constraints in driving scientific workflow:
  - Downward: Articulating [Computational problem/Theory] to be solved *constrains* [Algorithms/Experiments] to those that can [solve it/test it]
  - Upward: Observing [Brain/Data] constrains what [algorithms are in play/ hypotheses are borne out], these in turn help reveal what [problem is being tackled by system/theory is true of system]

Optimising scientific experimentation?

- Theory of Optimal Experimental Design (Peirce, 1898/1992), Active Learning (cf, Settles, 2012)

- Having formalised a set of models (or model with unknown parameters) can also formally derive the most efficient way to resolve uncertainty wrt them



Psychological Review
2009, Vol. 116, No. 3, 499–518

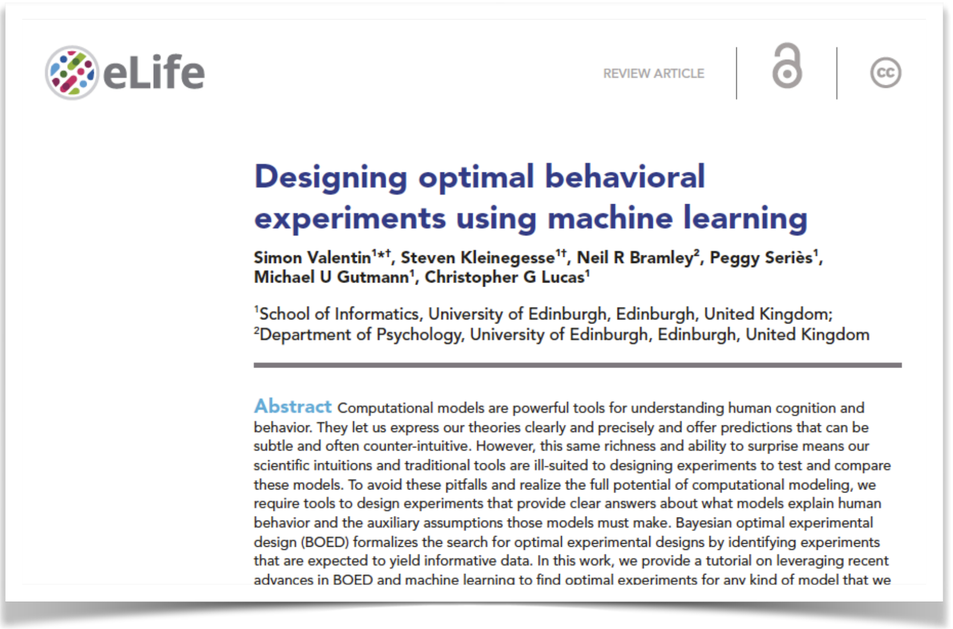
0033-295X/09/$12.00 DOI: 10.1037/a0016104

Optimal Experimental Design for Model Discrimination

Jay I. Myung and Mark A. Pitt
Ohio State University

Models of a psychological process can be difficult to discriminate experimentally because it is not easy to determine the values of the critical design variables (e.g., presentation schedule, stimulus structure) that will be most informative in differentiating them. Recent developments in sampling-based search methods in statistics make it possible to determine these values and thereby identify an optimal experimental design. After describing the method, it is demonstrated in 2 content areas in cognitive psychology in which models are highly competitive: retention (i.e., forgetting) and categorization. The optimal design is compared with the quality of designs used in the literature. The findings demonstrate that design optimization has the potential to increase the informativeness of the experimental method.

Keywords: formal modeling, model discrimination, Markov chain Monte Carlo, retention, categorization



eLife

REVIEW ARTICLE

Designing optimal behavioral experiments using machine learning

Simon Valentin[1]*†, Steven Kleinegesse[1†], Neil R Bramley[2], Peggy Seriès[1], Michael U Gutmann[1], Christopher G Lucas[1]

[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom; [2]Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom

Abstract Computational models are powerful tools for understanding human cognition and behavior. They let us express our theories clearly and precisely and offer predictions that can be subtle and often counter-intuitive. However, this same richness and ability to surprise means our scientific intuitions and traditional tools are ill-suited to designing experiments to test and compare these models. To avoid these pitfalls and realize the full potential of computational modeling, we require tools to design experiments that provide clear answers about what models explain human behavior and the auxiliary assumptions those models must make. Bayesian optimal experimental design (BOED) formalizes the search for optimal experimental designs by identifying experiments that are expected to yield informative data. In this work, we provide a tutorial on leveraging recent advances in BOED and machine learning to find optimal experiments for any kind of model that we

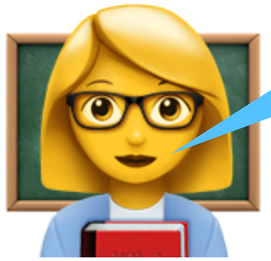
Break here?

Questions etc?

High level model evaluation

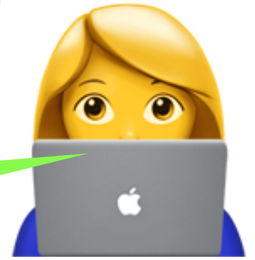
- van Rooij & Blokpoel (2020) characterises models development as Socratic dialog between **Verbal** and **Formal**

- Highlights how the act of formalising reveals various pressure points



I'd like to explain how a host decides whom to invite to a party

Why would the host not invite everybody?



They may like some people but dislike others.

Then the host invites everybody they like?

Not all people get along. If people get into an argument that can spoil a party.

I see. So a host may choose to invite people they like and that all get along.

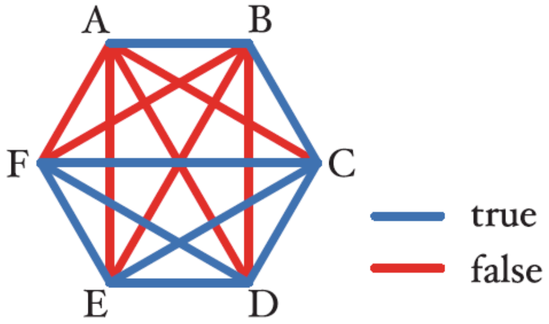
Yes, that sounds right. I think that's what a host will tend to do. Can we formalise this idea?

SELECTING INVITEES (VERSION 1)

Input: A set of people P, some of whom the host likes ($L \subseteq$ P) and some of whom the host dislikes ($D \subseteq P$), with $L \cap D = \emptyset$ and $L \cup D =$ P, and a function *like*: $P \times P \rightarrow$ {true, false} specifying for each pair of persons $(p_i, p_j) \in P$ whether or not they like each other.

Output: A set of liked guests $G \subseteq L$ that all like each other (i.e., $like(p_i, p_j) =$ true for each $p_i, p_j \in G$).

High level model evaluation



Of course in that situation the host would invite {C, D, E, F}

Or they would invite {A, B}

I would not think so

But according to Version 1 of the model, subset {A, B} is as likely to be the selected invitees as {C, D, E, F}, or at least there is no reason why the host would select the one and not the other.

But a party with only two guests is not much of a party!

So there are more constraints on the subset of guests that you have in mind but did not tell me yet. The host wants to have at least 3 guests?

As many as possible, the more the merrier.

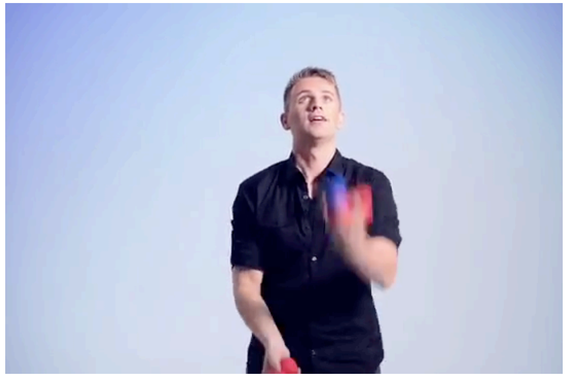
OK. Here's an adjusted version of the model:

SELECTING INVITEES (VERSION 2)

Input: A set P, subsets $L \subseteq P$ and $D \subseteq P$ with $L \cap D = \emptyset$ and $L \cup D = P$, and a function *like*: $P \times P \rightarrow$ {true, false}.

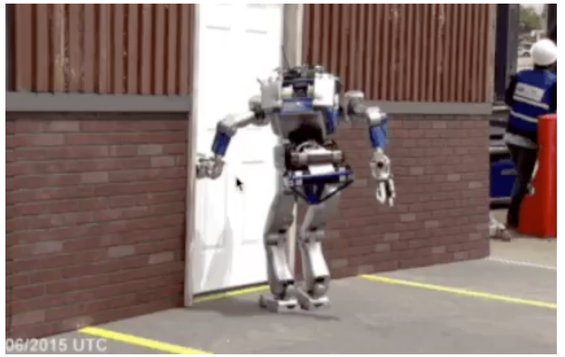
Output: A subset $G \subseteq L$ such that $\forall_{p_i,p_j \in G} like(p_i, p_j) =$ true and the size of G is maximized (i.e., there exists no G' such that $\forall_{p_i,p_j \in G'} like(p_i, p_j) =$ true and $|G'| > |G|$).

Personally relevant example

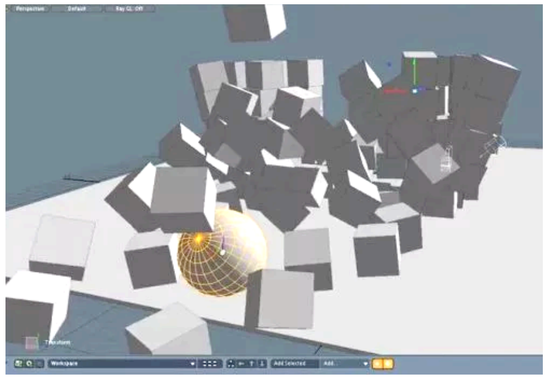


Despite recent progress in artificial intelligence…

…still lacks human level **competency** or **flexibility** in basic interactions with physical world



Theory: Humans learn internal “intuitive physics” engine + use this to reason robustly through mental simulation (Battaglia et al, 2013; Smith & Vul, 2013; Tenenbaum et al, 2011; Unman et al, 2017)



I.e. rather like a game engine, embodying Newtonian physical laws…

$$F = G\frac{m_1 m_2}{d^2} \qquad v_2' = \frac{2m_1}{m_1 + m2}v_1 \quad \text{etc.}$$

…and latent parameters (masses, forces, friction, elasticity)?

Personally relevant example

PNAS

Simulation as an engine of physical scene understanding

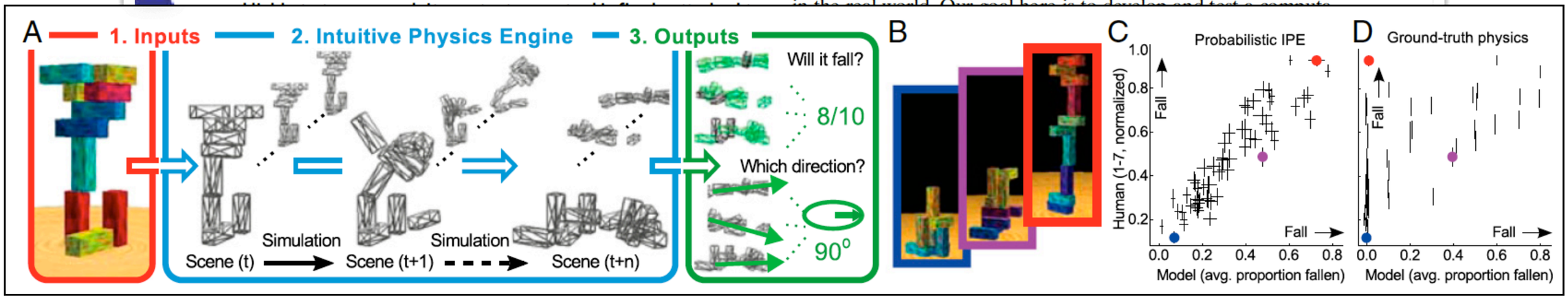
Peter W. Battaglia[1], Jessica B. Hamrick, and Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved September 20, 2013 (received for review April 8, 2013)

In a glance, we can perceive whether a stack of dishes will topple, a branch will support a child's weight, a grocery bag is poorly packed

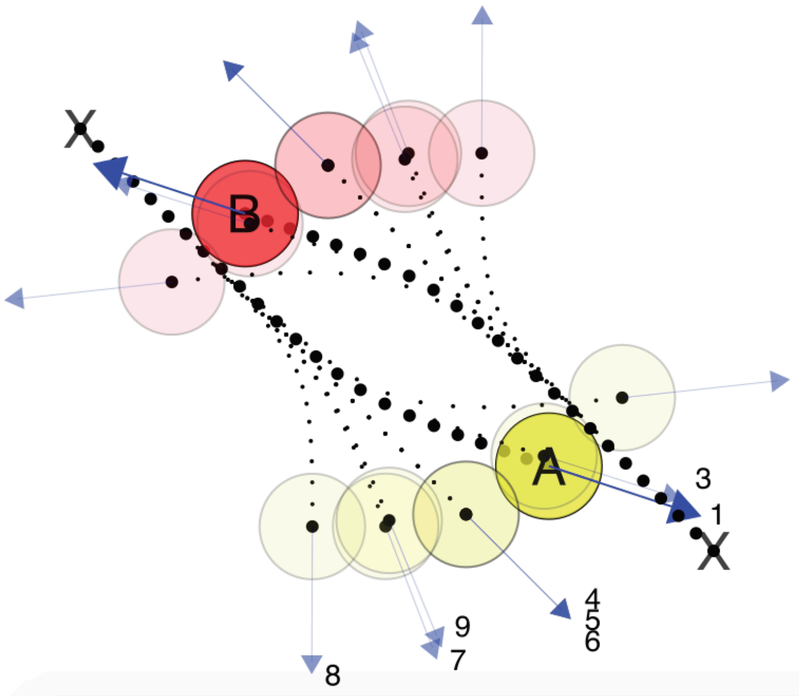
very simple, idealized cases, much closer to the examples of introductory physics classes than to the physical contexts people face



Personally relevant example

- To implement this theory, Battaglia et al assume learners integrate over uncertainty (e.g. perceptual / latent properties) by running many slightly different simulations in parallel (i.e. Monte Carlo integration)

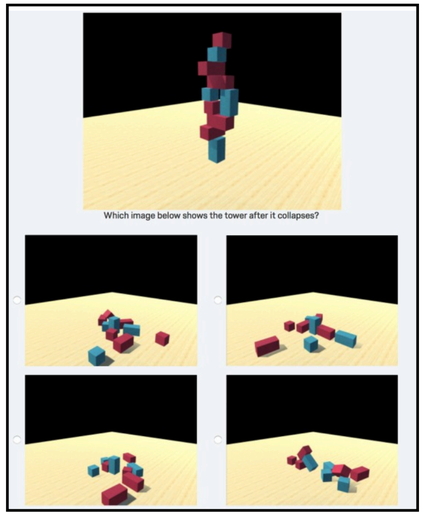
- Necessary to derive their experimental predictions but seemingly incompatible with other behavioural phenomena… (Ludwin-Peery, Bramley, Davis & Gureckis, 2020/2021)



Personally relevant example

- All objects must be consistently represented, simulated in synchrony and outcomes aggregated over independent simulations

- But Ludwin-Peery et al showed people systematically violate all three model features



People systematically select fallen block towers that do not contain same objects

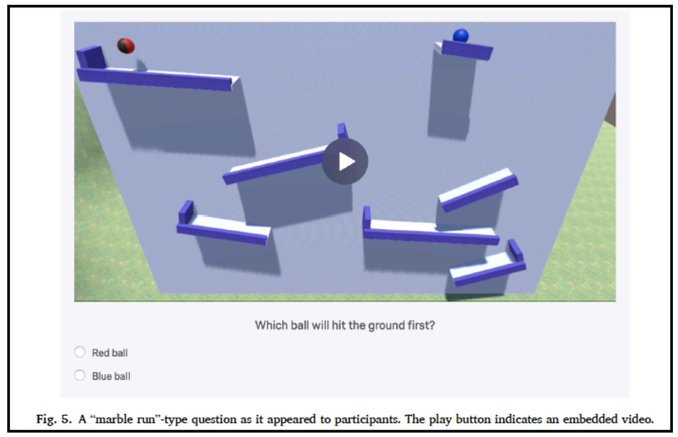
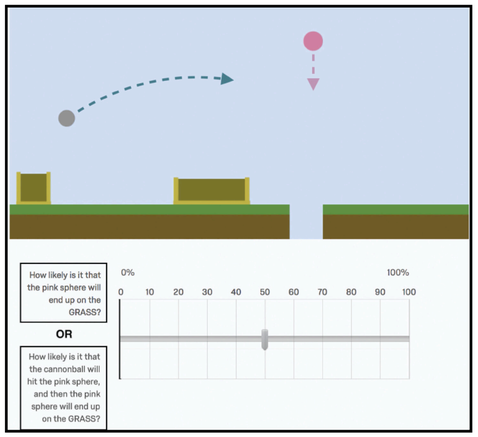


Fig. 5. A "marble run"-type question as it appeared to participants. The play button indicates an embedded video.

Predicting marble run outcomes that are chronologically impossible



Judging conductive outcomes as more likely than their constituent elements

Personally relevant example

- Original theorizers rise to our challenge and adjust and refine their theory of mental simulation

COGNITIVE NEUROPSYCHOLOGY

Routledge
Taylor & Francis Group

Cognitive Neuropsychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pcgn20

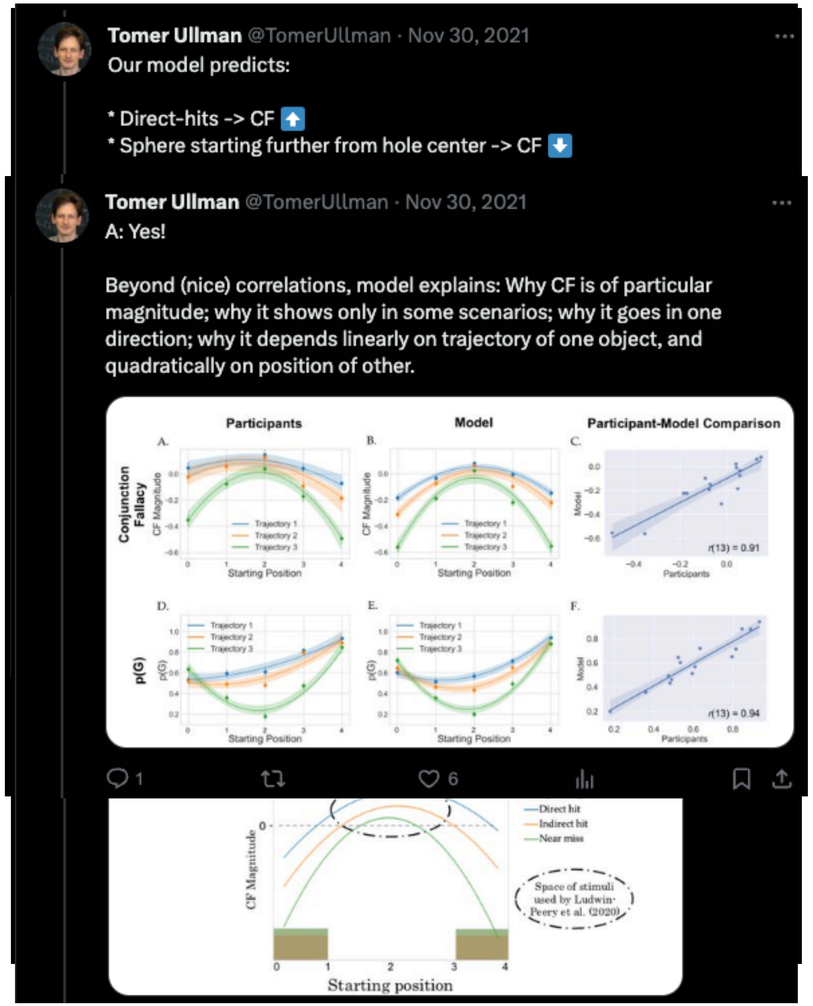
Partial mental simulation explains fallacies in physical reasoning

Ilona Bass, Kevin A. Smith, Elizabeth Bonawitz & Tomer D. Ullman

To cite this article: Ilona Bass, Kevin A. Smith, Elizabeth Bonawitz & Tomer D. Ullman (2022): Partial mental simulation explains fallacies in physical reasoning, Cognitive Neuropsychology, DOI: 10.1080/02643294.2022.2083950

To link to this article: https://doi.org/10.1080/02643294.2022.2083950

- **Progress achieved!**



Tomer Ullman @TomerUllman · Nov 30, 2021

Our model predicts:

* Direct-hits -> CF ⬆️
* Sphere starting further from hole center -> CF ⬇️

Tomer Ullman @TomerUllman · Nov 30, 2021

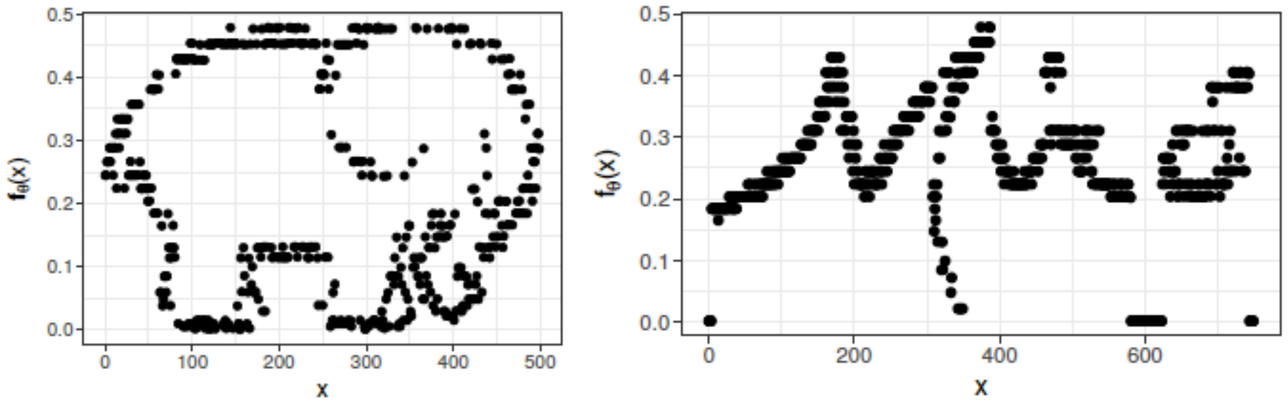
A: Yes!

Beyond (nice) correlations, model explains: Why CF is of particular magnitude; why it shows only in some scenarios; why it goes in one direction; why it depends linearly on trajectory of one object, and quadratically on position of other.

1 6

Explanation vs Fit

- A tight fit between data and a model not always what we care about
- Can be due to the flexibility of a model (Myung)
- Heuristics for penalising model fit (i.e. parameter counting in BIC/AIC) can be pushed to failure…



From Piantadosi, (2018) -"One parameter is always enough"

FIG. 1: A scatter plot of f_θ for $\theta = 0.24468472667347454582275406 56\cdots$ plotted at integer x values, showing that a single parameter can fit an elephant (left). The same model run with parameter $\theta = 0.00242654180550004019353876 20\cdots$ showing a fit of a scatter plot to Joan Miró's signature (right). Both use $r = 8$ and require hundreds to thousands of digits of precision in θ.

Explanation vs Fit

- "Between the devil and the deep blue sea" (Navarro, 2018): Beyond question of over- or under-fitting, there is the question of what we care about, different objective functions reflect different values:
  - Maximum Likelihood Estimation $P(\mathbf{X} \mid M, \hat{\theta})$ where $\hat{\theta} = \text{argmax}_\theta \prod_{x \in \mathbf{x}} \mathscr{L}(x \mid M, \theta)$ — How well can the model do when given its best shot
  - Bayes factors driven by prior predictive distribution $P(\mathbf{x} \mid M)$ — Does model capture the phenomenon averaging over prior on possible values of params
  - Cross validation closer to posterior predictive $P(\mathbf{x}' \mid \mathbf{x}, M)$ — Does the model, once fit to some observations, capture other/future observations?
  - Correlations more about matching qualitative patterns
- Navarro gives example of experiments on "sensitivity to sampling" where she agonises over the most appropriate scientific objective…

Explanation vs Fit

Narrower generalisation of novel properties from learning samples selected because they had that property than because they belonged to same category (category sampling).”

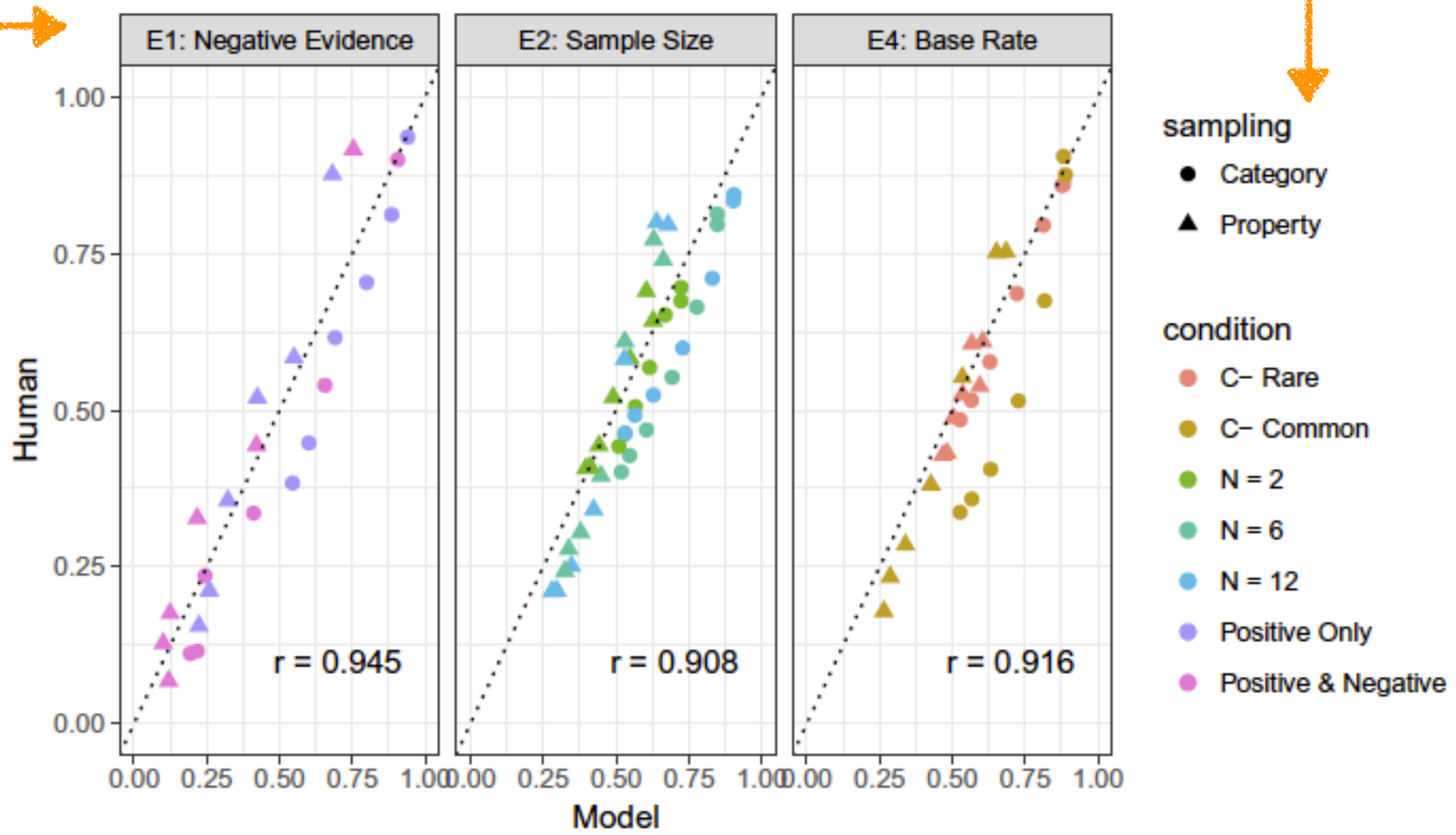
…and various conditions

Across several experiments

…their proposed & fitted model has a tight relationship with behavioural results:

A consistently high correlation between likelihood human assigns to new case and model-predicted probability of case having property

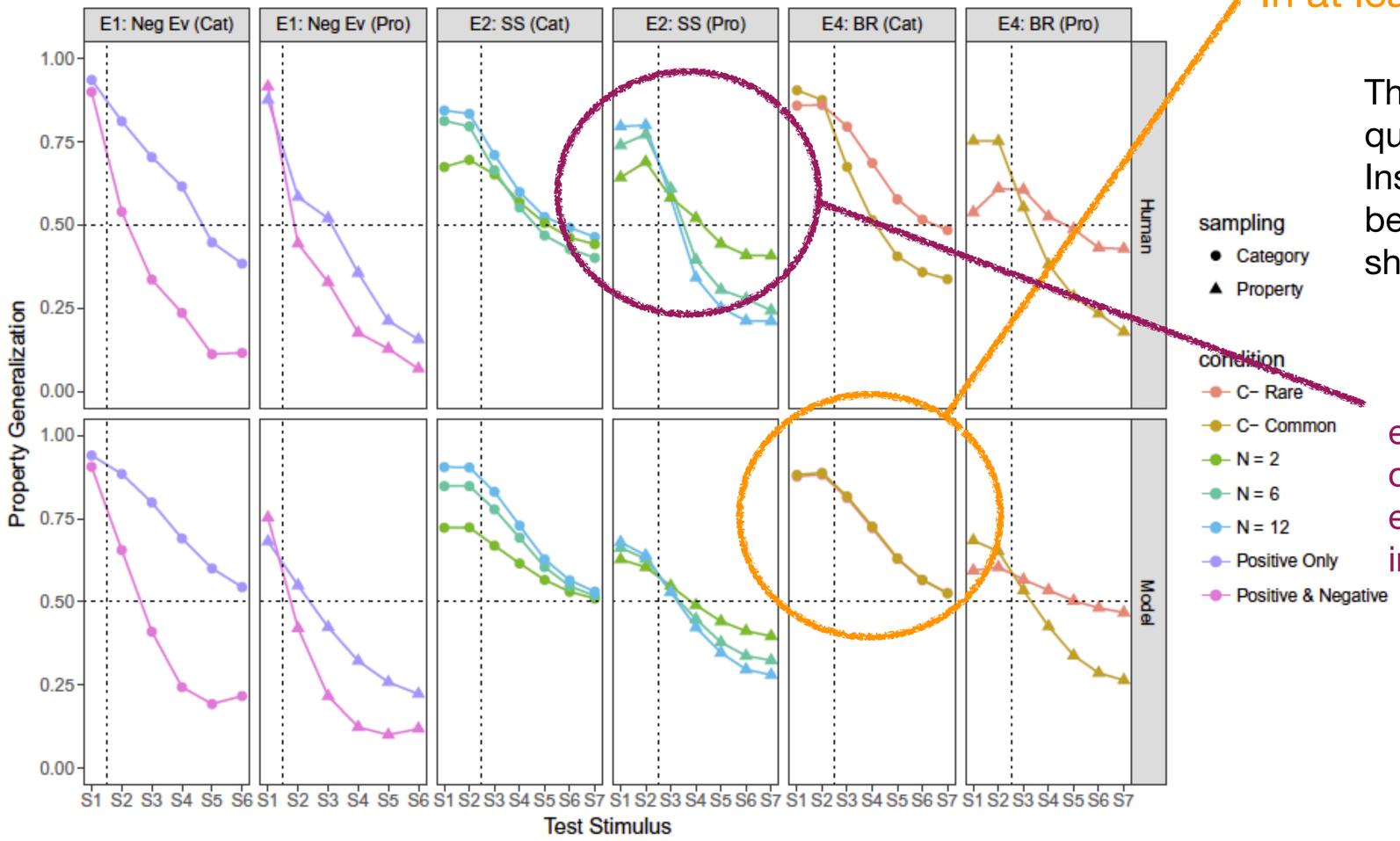
Almost suspiciously high..



Explanation vs Fit

"[…] Some choices (e.g. how smooth is an unknown generalisation function?) can be instantiated as model parameters, but others (e.g. what class of functions is admissible to describe human generalisation?) not so simple.…

Does model explain the qualitative empirical phenomena they set out to explain?



In at least one case it really does not!

Therefore, they elect to put little weight on quantitative measure of fit
Instead derive ordinal constraints from behavioural data a satisfactory model should mimic

e.g. "Increasing number of observations must cause crossover effect under property sampling irrespective of model parameters" etc

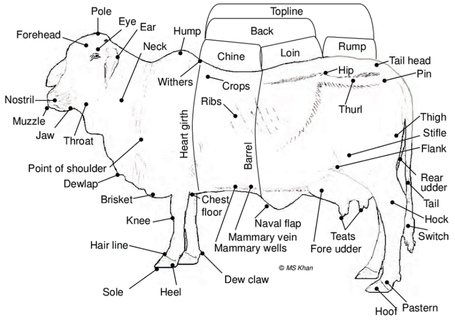
Scope

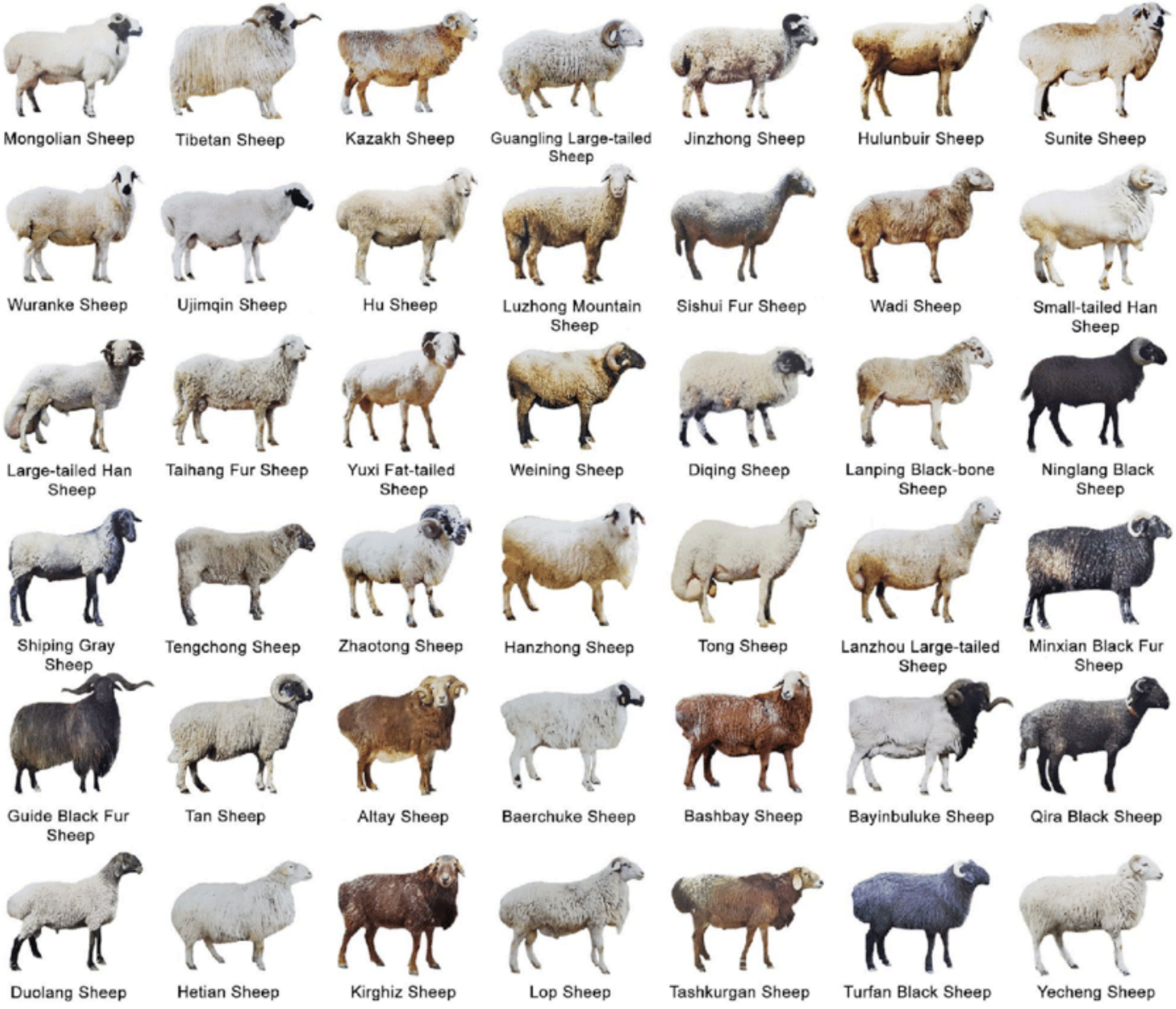
Weak generalisation

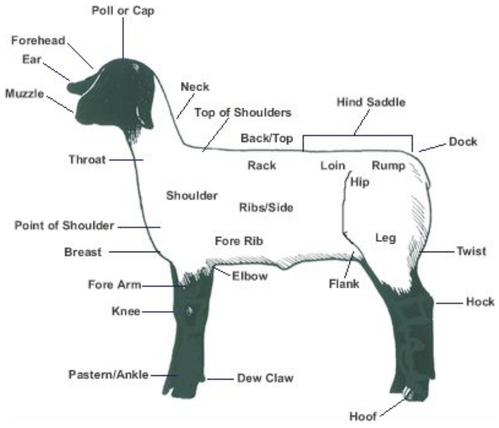
- The training set and evaluation set are drawn from the same generative model for the same task

Strong generalisation

- The evaluation set falls outside the training distribution
- The model is evaluated on a different task
- Ultimately depends on your theory aligning with reality within the scope it is applied


Pole
Eye
Ear
Forehead
Hump
Neck
Topline
Back
Chine
Loin
Rump
Withers
Crops
Hip
Tail head
Pin
Ribs
Thurl
Nostril
Muzzle
Jaw
Throat
Heart girth
Barrel
Thigh
Stifle
Flank
Point of shoulder
Dewlap
Rear udder
Brisket
Chest floor
Tail
Knee
Naval flap
Hock
Mammary vein
Mammary wells
Teats
Fore udder
Switch
Hair line
© MS Khan
Dew claw
Sole
Heel
Hoof
Pastern


Mongolian Sheep
Tibetan Sheep
Kazakh Sheep
Guangling Large-tailed Sheep
Jinzhong Sheep
Hulunbuir Sheep
Sunite Sheep
Wuranke Sheep
Ujimqin Sheep
Hu Sheep
Luzhong Mountain Sheep
Sishui Fur Sheep
Wadi Sheep
Small-tailed Han Sheep
Large-tailed Han Sheep
Taihang Fur Sheep
Yuxi Fat-tailed Sheep
Weining Sheep
Diqing Sheep
Lanping Black-bone Sheep
Ninglang Black Sheep
Shiping Gray Sheep
Tengchong Sheep
Zhaotong Sheep
Hanzhong Sheep
Tong Sheep
Lanzhou Large-tailed Sheep
Minxian Black Fur Sheep
Guide Black Fur Sheep
Tan Sheep
Altay Sheep
Baerchuke Sheep
Bashbay Sheep
Bayinbuluke Sheep
Qira Black Sheep
Duolang Sheep
Hetian Sheep
Kirghiz Sheep
Lop Sheep
Tashkurgan Sheep
Turfan Black Sheep
Yecheng Sheep


Poll or Cap
Forehead
Ear
Muzzle
Neck
Top of Shoulders
Hind Saddle
Back/Top
Dock
Throat
Rack
Loin
Rump
Hip
Shoulder
Ribs/Side
Point of Shoulder
Fore Rib
Leg
Breast
Twist
Elbow
Fore Arm
Flank
Hock
Knee
Pastern/Ankle
Dew Claw
Hoof

Take homes

- Science is a garden of forking paths
- Computational modelling is a way of keeping track of the forks
- When we eschew this, we make consequential choices implicitly, or blindly
- Model evaluation can occur at all points on the path, not only at the level of data
- Model evaluation against data is often not what we care most about —yet sociologically it is what we do

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327-18332.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413-424.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407.

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science*, *21*(10), 1363-1368.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, 16(4), 789-802.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., ... & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, 5(8), 180448.

Hardwicke, T. E., & Wagenmakers, E. J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, 7(1), 15-26.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review,* 121(1), 1.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. Trends in cognitive sciences, 14(7), 293-300.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, 31(12), 1602-1611.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, 127, 101396.

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, 2(1), 28-34.

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories: A tutorial by dialogue. *Social Psychology, 51*(5), 285–298.

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, 32(6), 939-984.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Peirce, C. S. (1898/1992). *Reasoning and the logic of things: The Cambridge conferences lectures of 1898.* Harvard University Press.

Piantadosi, S. T. (2018). One parameter is always enough. *AIP Advances*, 8(9).

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10), 421-425.

Settles, B. (2012). *Active learning*. Springer.

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, 115(11), 2632-2639

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in cognitive sciences*, 24(2), 94-95.

Szollosi, A., Grigoras, V., Quillien, T., Lucas, C., & Bramley, N. R. (2023). How do instructions, examples, and testing shape task representations? In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45, e1.