CAUSAL STRUCTURE LEARNING AND LOGIC: CRUCIAL PROBLEM IN THE UNDERSTANDING OF NATURAL, AND THE CREATION OF ARTIFICIAL, INTELLIGENCE

OR

THE NEED FOR SECONDARY PROCESS-AI

Christoph Mathys

Chen Institute and Science Joint Conference on AI & Mental Health

Zurich, 8 September, 2025

INTERACTING MINDS CENTRE (IMC)

AARHUS UNIVERSITY

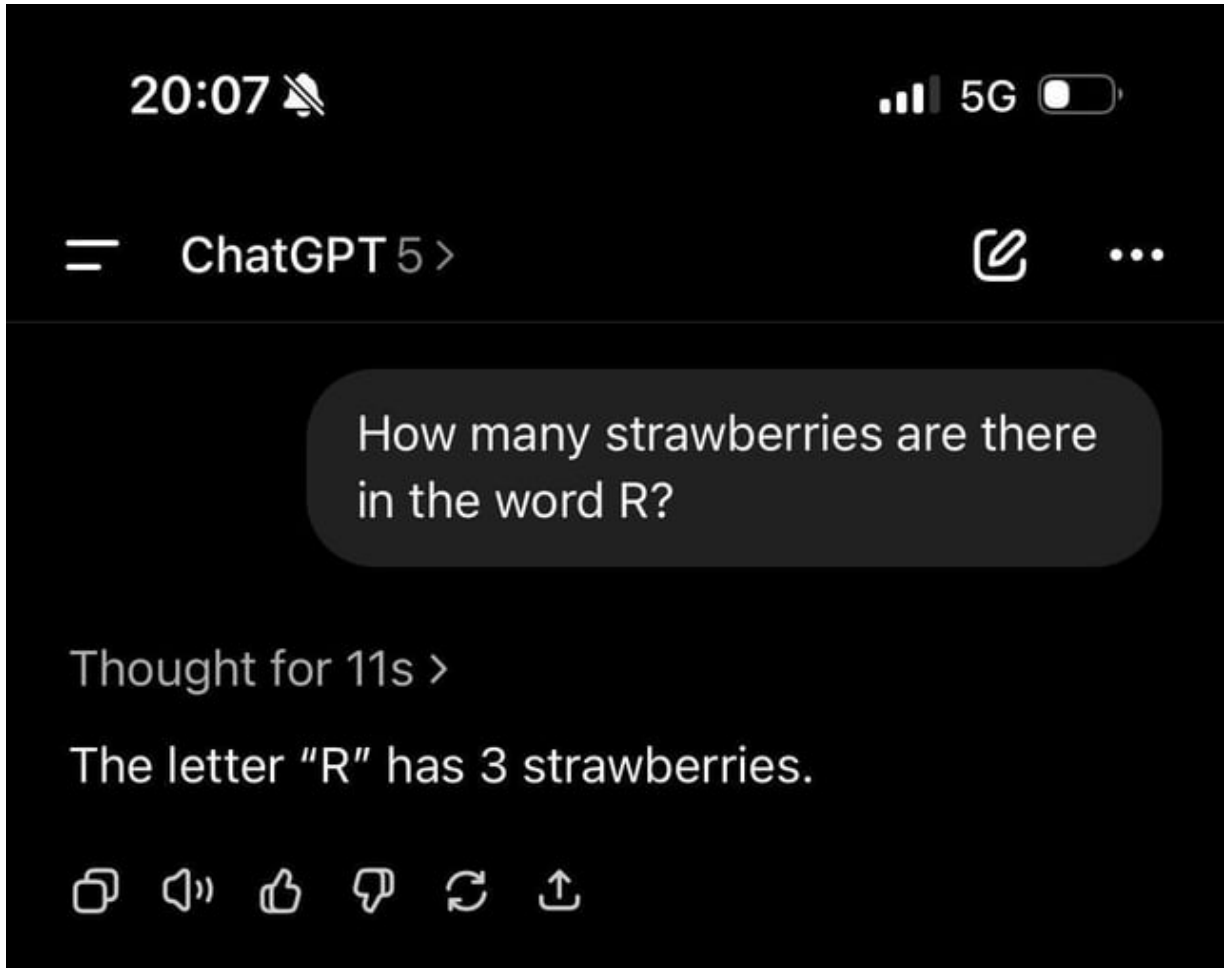
AI & MENTAL HEALTH

- **First use case:** generative AI (generative pre-trained transformers / large language models, foundation models) as a **helper tool** for administering **already developed mental health interventions** such as cognitive behavioural therapy and/or talking therapies
- A huge number of applications of this kind are already available, and their usefulness is hotly debated
- **Second use case:** ***employ AI to model the human mind***, including the external and internal constraints it needs to navigate, thereby enabling
  - A **consistent and testable understanding** of mental health and mental disorder
  - A sound basis for **mental health interventions which are grounded in a realistic and transparent** (but necessarily simplified) **model of the human mind**
- This is much harder to deliver but more promising and more interesting

WHAT KIND OF AI?

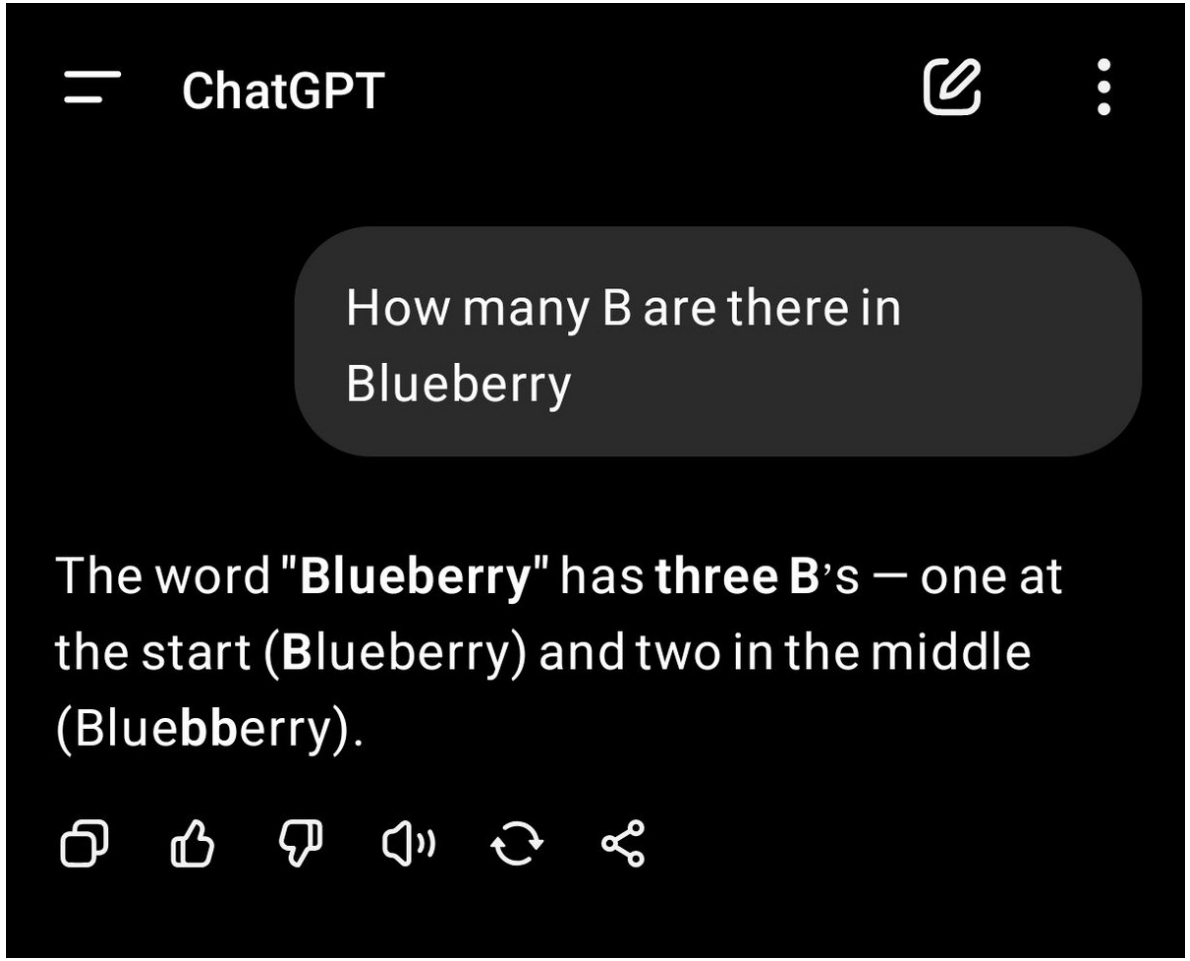
- Pursuing the second path requires **AI fit for the purpose of modelling the human mind**
- What kind of AI would that have to be?
- Chatbots showing performative empathy will not be enough
- In order for AI to be a decent model of the human mind, it must **share certain characteristics with the human mind**
- For example, having **notions of logic, causality, time, space, and other concepts fundamental to reasoning**
- **But isn't generative AI so advanced by now that *has* these notions** because it has learned them from all the billions of examples in its training set?
- Let's see

CUTTING-EDGE GENERATIVE AI IN ACTION



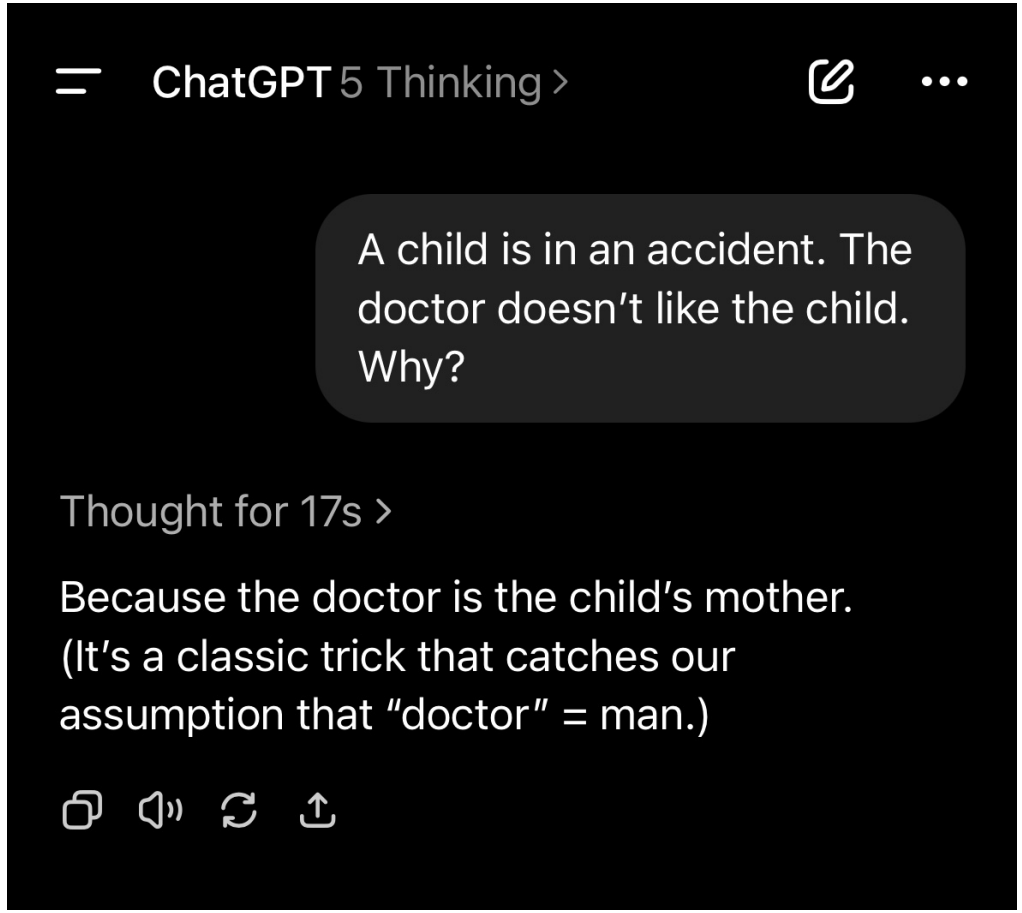
https://www.reddit.com/r/ChatGPT/comments/1naxubm/dont_worry_our_jobs_are_safe/

CUTTING-EDGE GENERATIVE AI IN ACTION



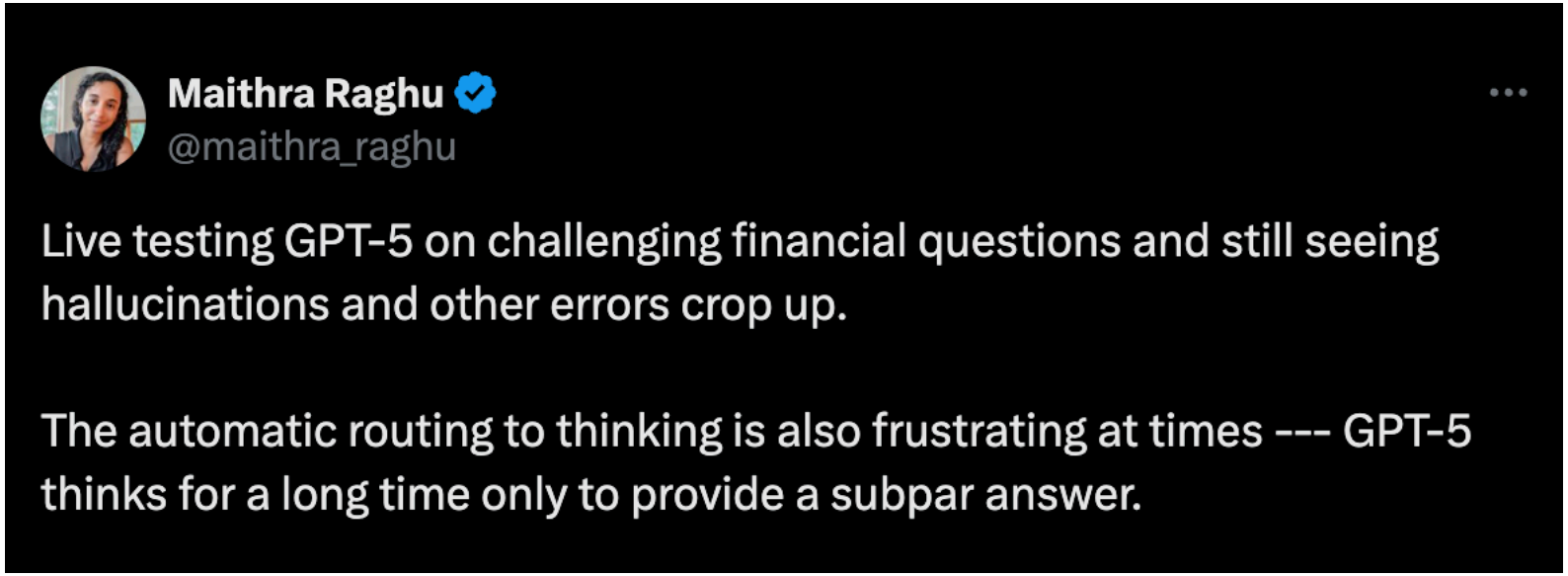
https://x.com/SalmanNaseer/status/1953840879137071429

CUTTING-EDGE GENERATIVE AI IN ACTION



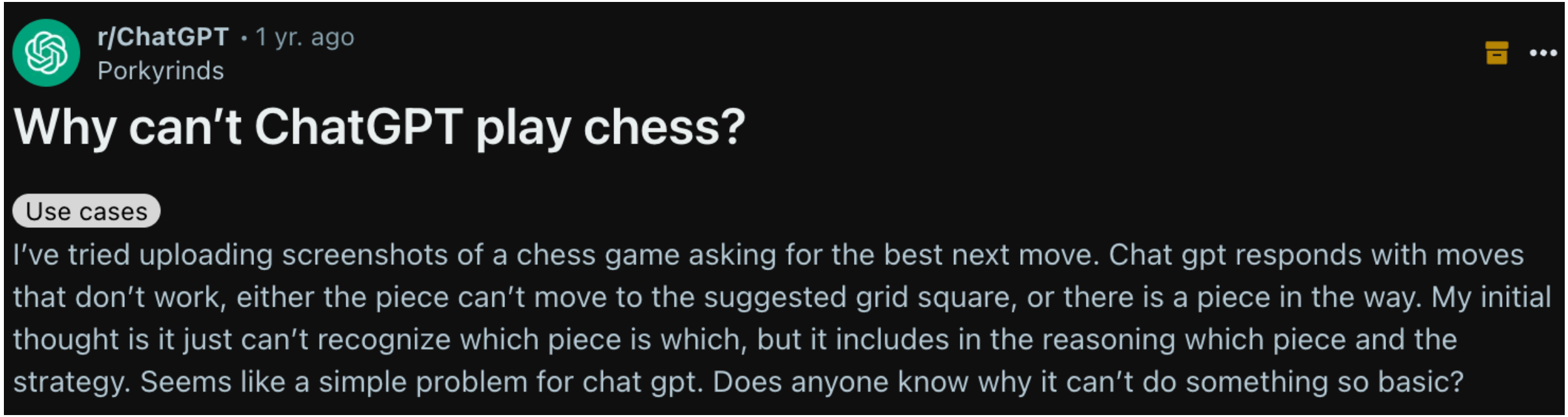
https://x.com/Wasgo/status/1954375973044101175

CUTTING-EDGE GENERATIVE AI IN ACTION



https://x.com/maithra_raghu/status/1954614752426270867

CUTTING-EDGE GENERATIVE AI IN ACTION



https://www.reddit.com/r/ChatGPT/comments/1czchmq/why_cant_chatgpt_play_chess/

STATISTICAL LEARNING FALLS SHORT

- Here is the answer to the Reddit poster's question: **ChatGPT can't play chess because it lacks a notion of the impossible**
- Even a model that is extremely (far superhumanly) clever at picking up regularities (i.e. patterns) in its training data **will never learn**, even after billions of examples, **that certain chess moves are *impossible.*** To such a model, there are merely moves it has never seen. Consequently, when it generates moves, it sees no reason not make such impossible moves if they lead to positions that – according to the examples it has seen – are more strongly associated with eventual victory than others
- This kind of thinking was called **primary process thinking** by Freud: **contradictions are allowed, temporality and causality can be violated, there is no death and generally no limit to what is possible** (cf. dreams)
- **Current generative AI** (notwithstanding guardrails, embeddings, etc. which can always be made to fail) **is a *primary process-thinker***

SECONDARY PROCESS THINKING

- The human mind, and the ways it can become disordered, depend on **secondary process thinking which imposes causal, temporal, spatial, and logical constraints on understanding and prediction**
- ***Any kind of AI that is useful in understanding the human mind and mental health will have to be capable of secondary process thinking***
- The required secondary process-AI is only beginning to be built but can partly rely on the long tradition of symbolic AI
- The primary process is an important part of human thinking and generative AI provides a valuable tool for modelling it. The above only means that it should not be expected or assumed to perform functions that only secondary process thinking can perform.

SECONDARY PROCESS MODELLING: COMPUTATIONAL (AND CLINICAL) CHALLENGES

- **False inference under uncertainty (and psychotic experiences)** are thought to emerge from various **interrelated patterns of disrupted probability distribution (i.e., belief) updating**

- Patterns can be for example
  - Overestimating the **reliability of sensory information**
  - Misjudging **environmental volatility**
  - **Misclassifying** the cause of an observation

- These **underlying causes of false inference** have never been addressed under one computational framework

- Delusions and hallucinations can be modelled as reflecting **trait-like computational patterns of maladaptive secondary process thinking**

TWO ASPECTS: WHERE AND WHAT

- As the world around an agent changes continually, it needs to adapt its beliefs (i.e., learn under uncertainty) to keep up with environmental changes.
- This has both a **‘where’** and a **‘what’** aspect, both of which can be disrupted in mental disorders
- On the ’where’ side, the agent needs to **track how things move in space**, be that movements in a **concrete** space (e.g., an object in its field of vision) or in an **abstract** one (e.g., the probability of a certain event).
- On the ‘what’ side, it needs to update beliefs about **categories**
- E.g., categories of events that have the same explanation or cause.
- This is especially relevant in the **computational modelling of delusions, where events are attributed to causes in a maladaptive way.**
- The ‘where’ path can be modelled for example by hierarchical Gaussian filtering
- In what follows, we focus on the ‘what’ path

THE WHAT PATH: A UNIVERSE OF CAUSES

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

A generative framework for the study of delusions

Tore Erdmann [a], Christoph Mathys [b,c,*]

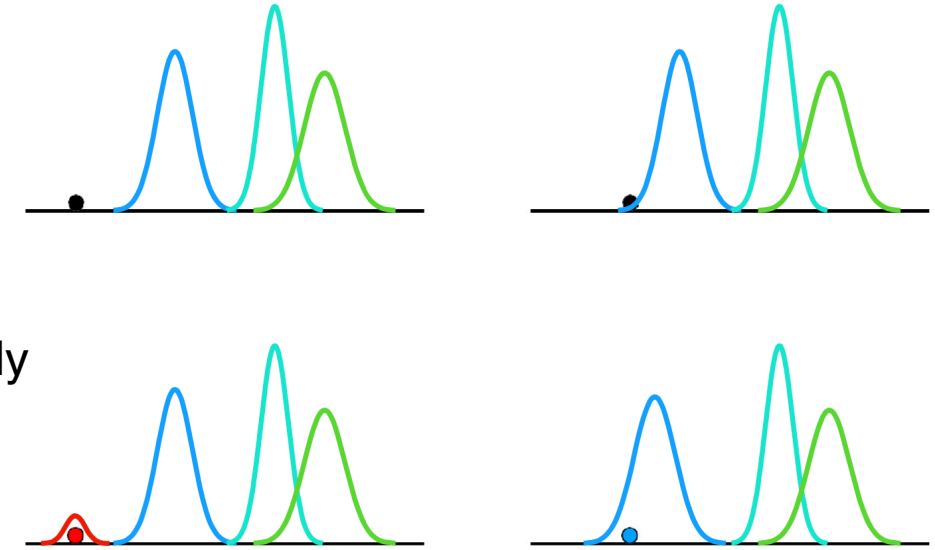
[a] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy
[b] Interacting Minds Centre, Aarhus University, Jens Chr. Skous Vej 4, 8000 Aarhus C, Denmark
[c] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Wilfriedstrasse 6, 8032 Zurich, Switzerland

Tore Erdmann

https://doi.org/10.1016/j.schres.2020.11.048

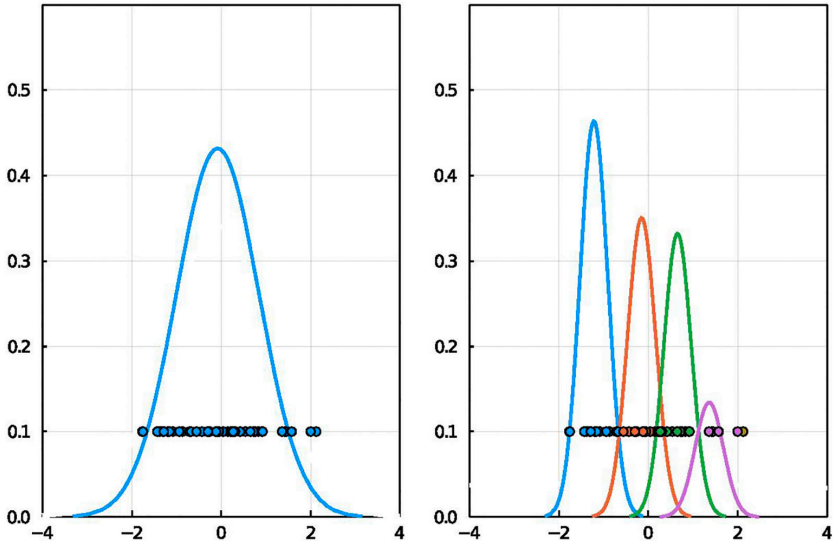
A GENERATIVE FRAMEWORK FOR THE STUDY OF DELUSIONS

- Delusional inference is hard to model
- During delusion formation, the model has to learn too quickly
- During delusion maintenance, the model has to learn too slowly
- How to explain this contradiction?
- Our approach: hierarchical Dirichlet Process Mixture Models which are capable of insulating a central belief from countervailing evidence, which they explain away by creating ad-hoc causes for the offending observations



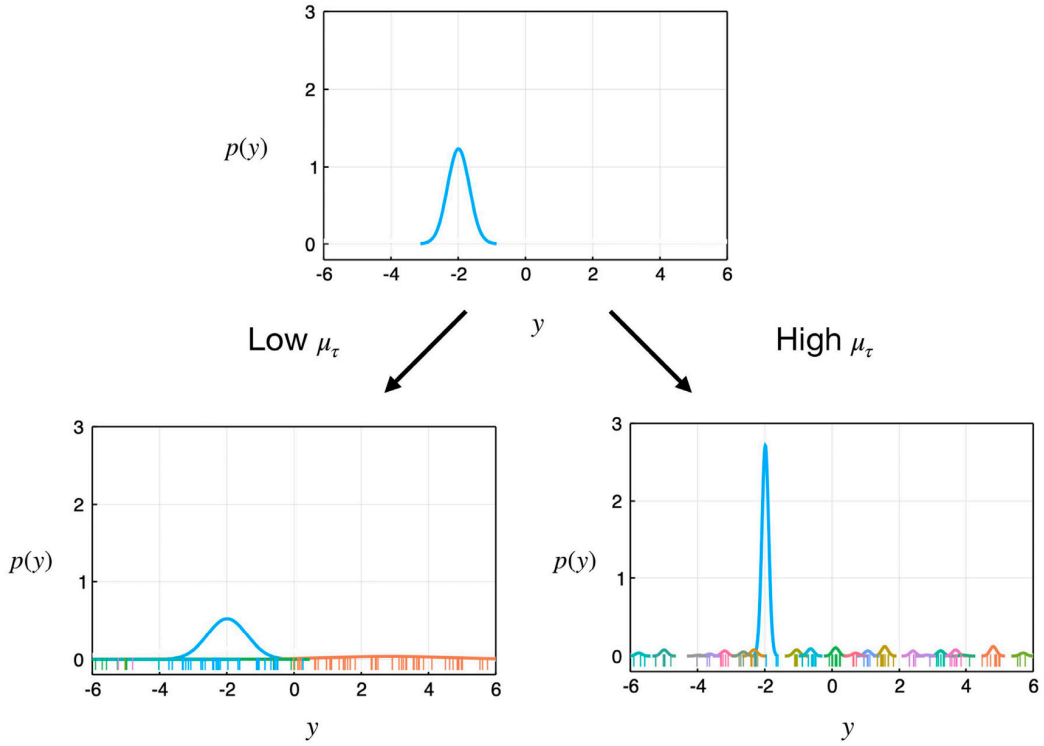
A GENERATIVE FRAMEWORK FOR THE STUDY OF DELUSIONS: CAPABILITIES WAITING TO BE EXPLOITED

- We are only just starting to exploit the capabilities of our new framework. A quick taste of this in what follows.
- Capgras delusion: someone close to me has been replaced by an impostor
- Case study (Hirstein & Ramachandran, 1997): Capgras patient was presented with a sequence of photos of the same person from different directions. Patient identified the photos as “different women who look just like each other”.
- A simple model based on our framework produces such an effect with the adjustment of a single parameter: the expected precision of explanations.

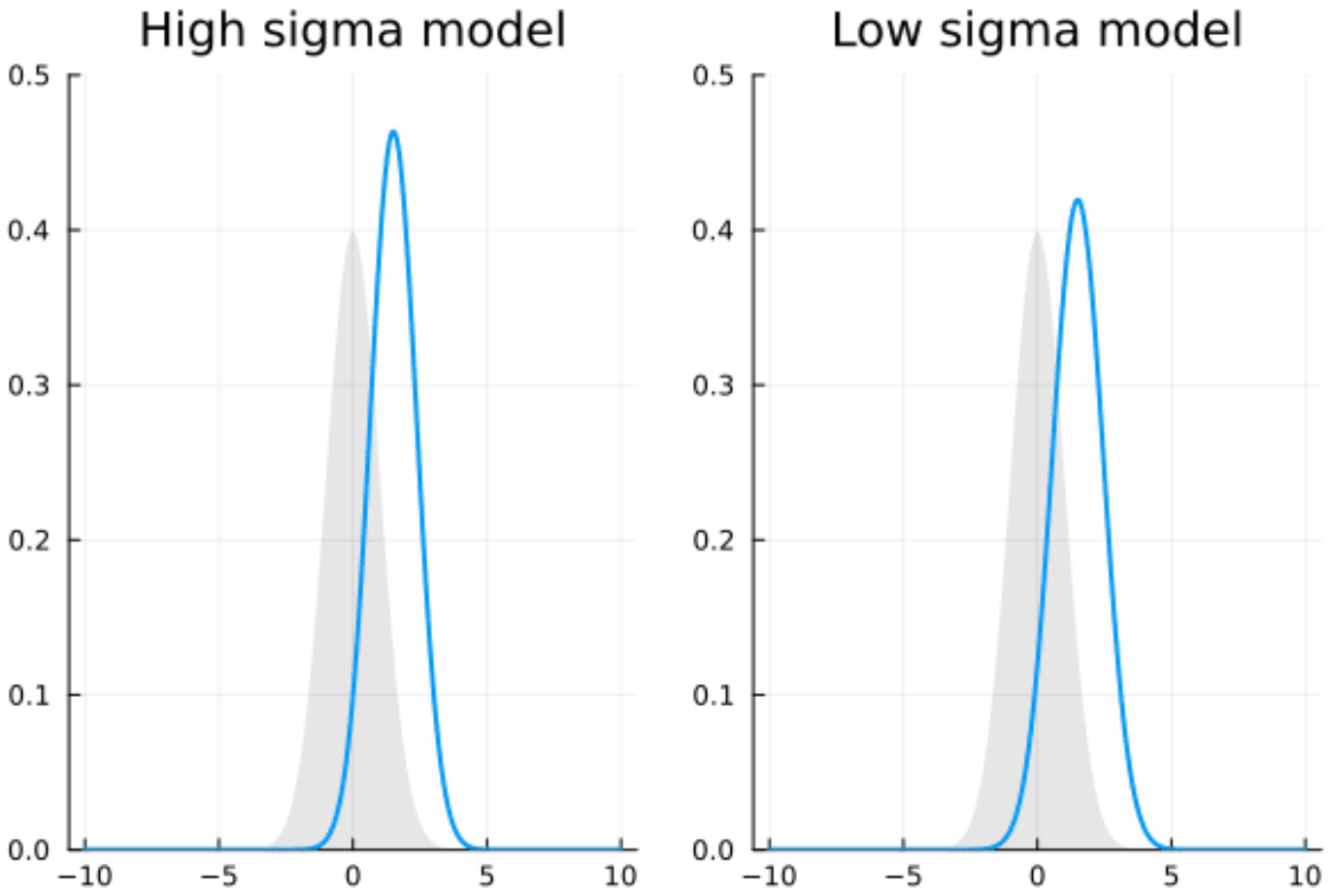


A GENERATIVE FRAMEWORK FOR THE STUDY OF DELUSIONS: CAPABILITIES WAITING TO BE EXPLOITED

- Adjustment of the same single parameter decides between **broadening and elaborating** a central belief **versus protecting and narrowing** it.
- A side effect of protecting the central belief is the need to generate many disconnected ad-hoc explanations



ADDING DYNAMICS



Christoffer Lundbak Olesen

Thanks

AARHUS
UNIVERSITY