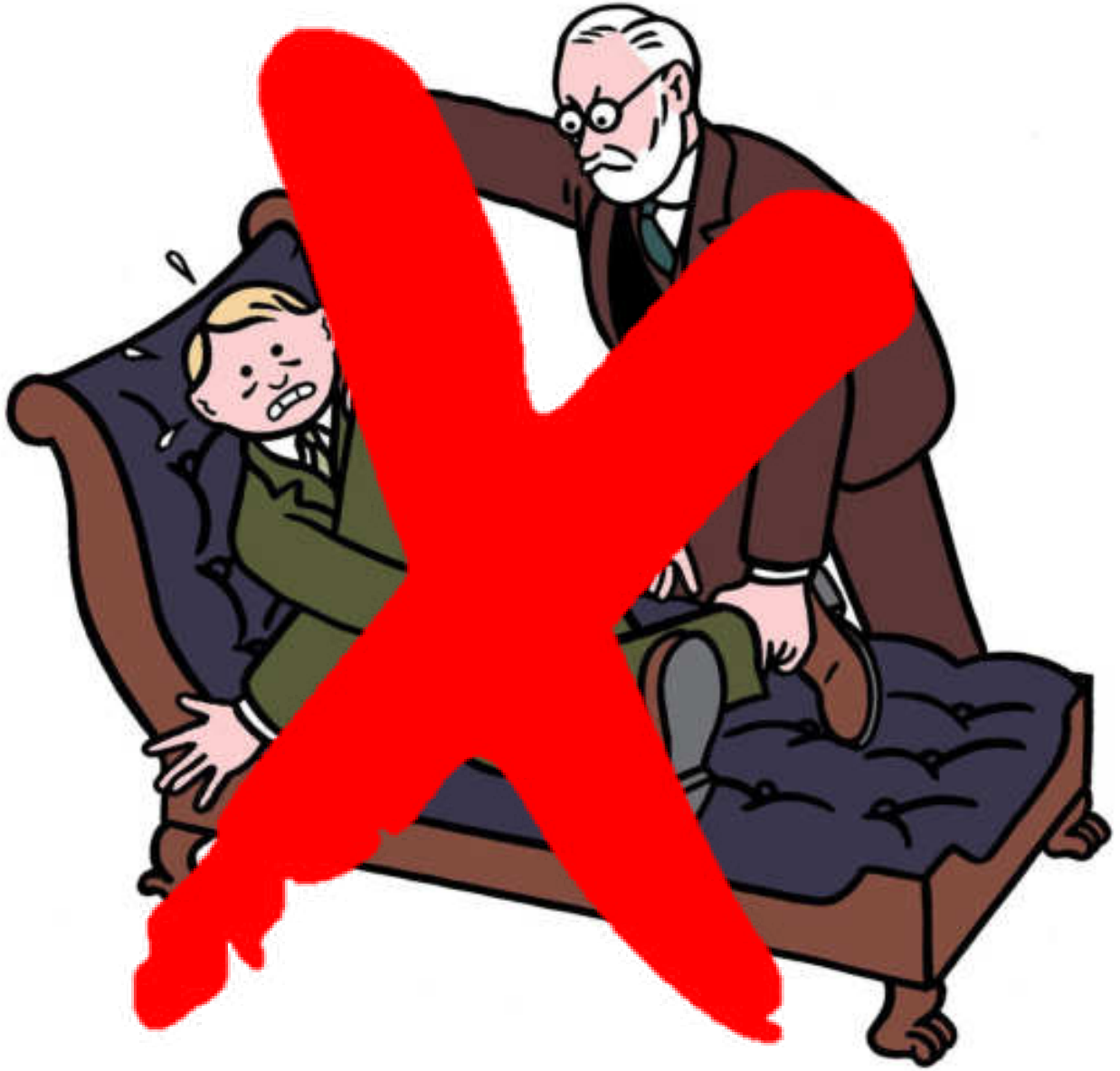
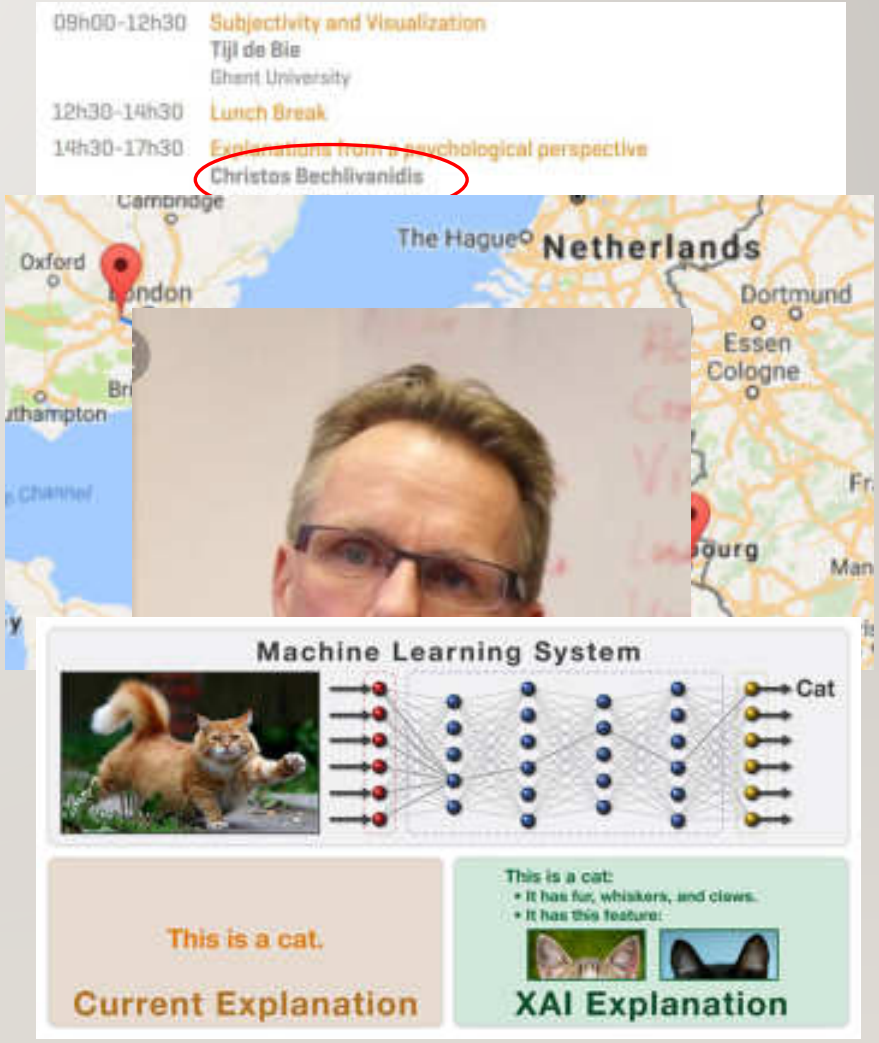


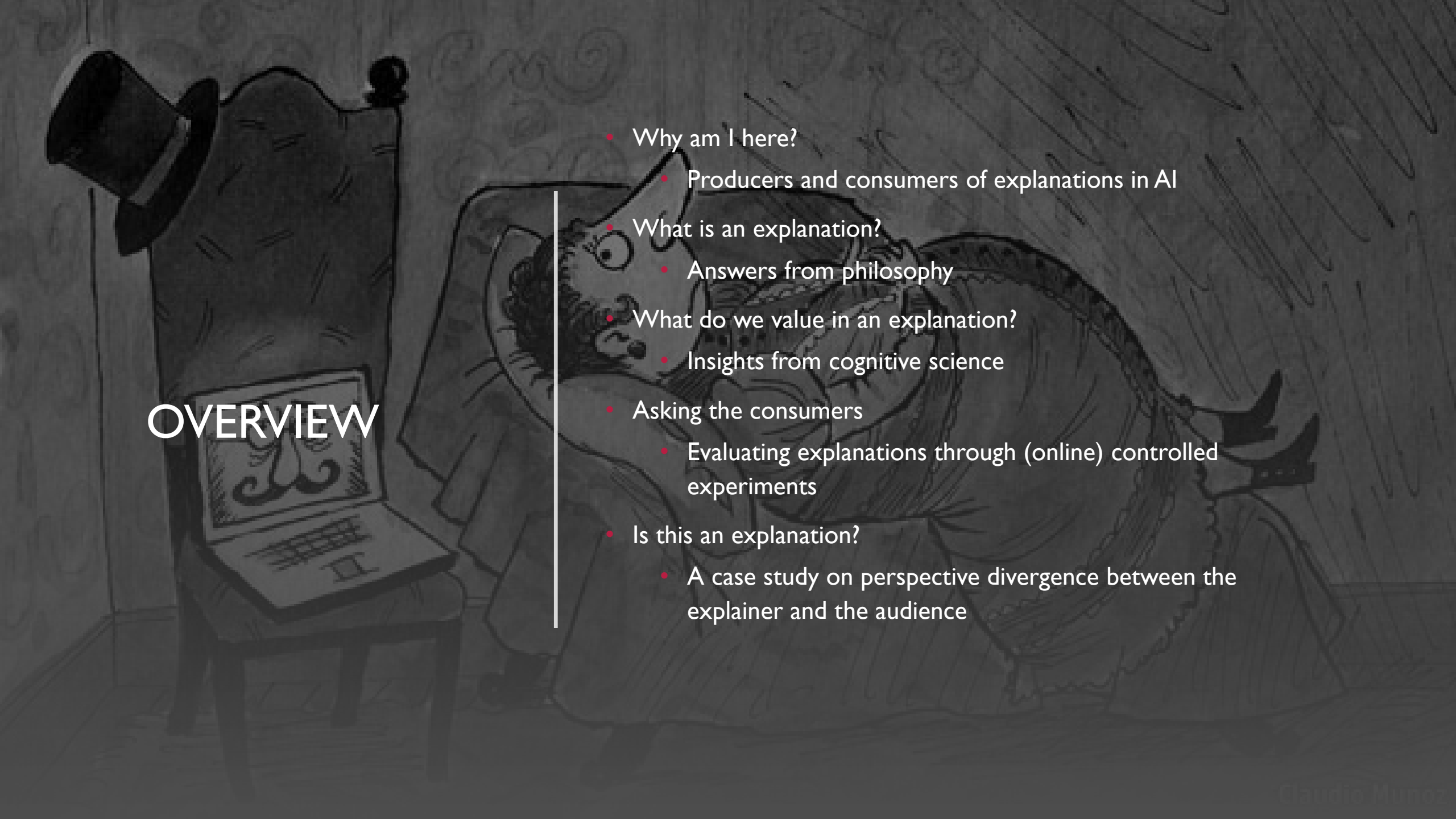
UCL

WHY AM I HERE?

- Because my name is next to 14:30 in the timetable
- Because I took a plane from London to Luxembourg and then Denise took me to my hotel and then I walked here
- Because Peter invited me
- Because it'd be interesting and useful to see how other disciplines think about the things that you're trying to generate.

- Explanations are always contrastive
- The contrast class is usually implied by the context





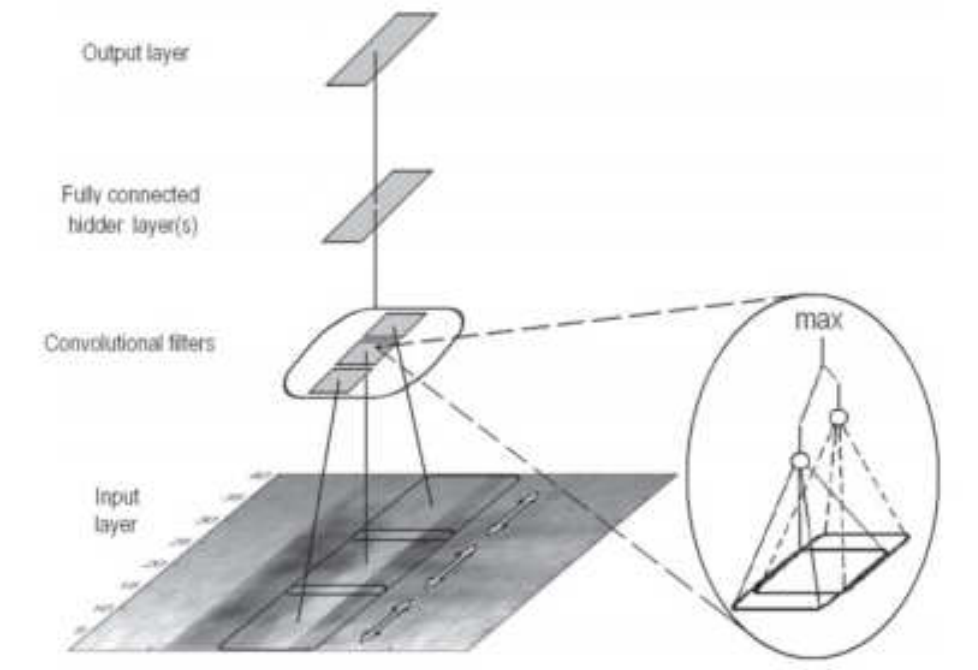
OVERVIEW

- Why am I here?
  - Producers and consumers of explanations in AI
- What is an explanation?
  - Answers from philosophy
- What do we value in an explanation?
  - Insights from cognitive science
- Asking the consumers
  - Evaluating explanations through (online) controlled experiments
- Is this an explanation?
  - A case study on perspective divergence between the explainer and the audience

WHAT IS THE X IN XAI?

- Who decides?
  - Miller et al (2017): The inmates are running the asylum
- The developer both produces AND evaluates the (form of) explanations
- e.g. "The recommender system produced the given output because the latter shared types and aspects with your stated preferences"
- The developer has deep knowledge about the system
  - is this a good thing?
  - i.e. different understanding, different perspective, different aims compared to the user

ON COMPLEXITY



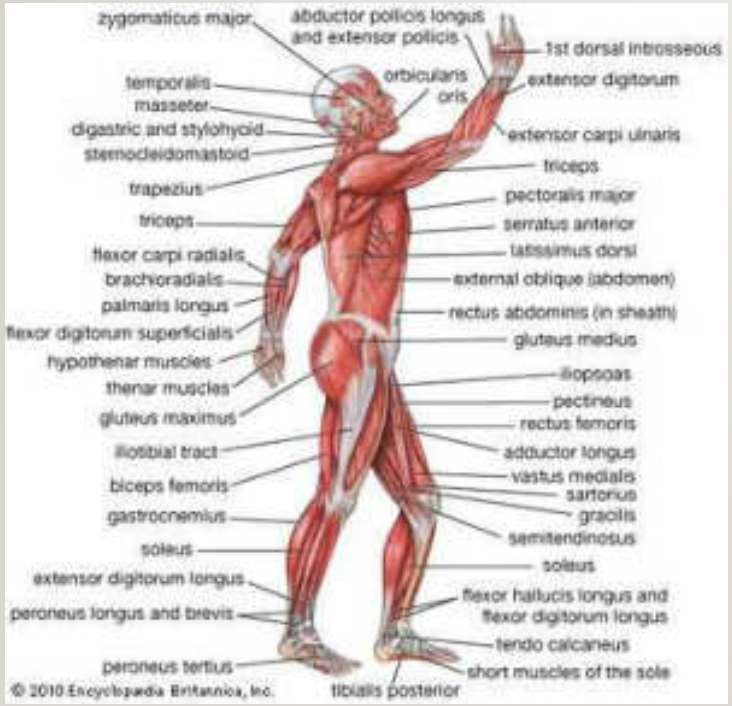
Hierarchical Convolutional Deep Maxout Network

VS

Fred

(not Fuzzy Reasoning Edge Detection, just Fred)

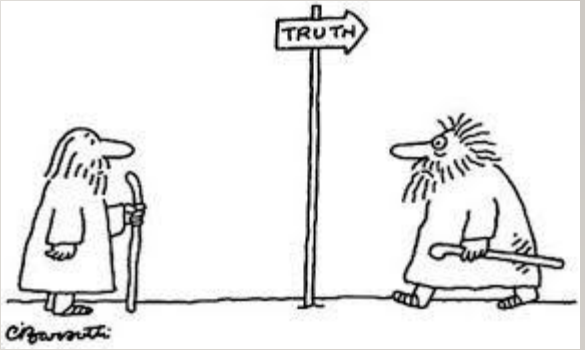
WHAT IS AN EXPLANATION?

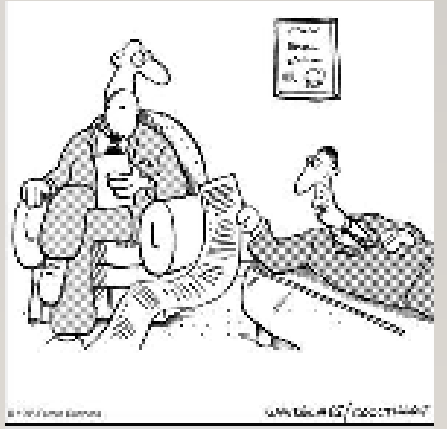


not a description

WHAT IS AN EXPLANATION?

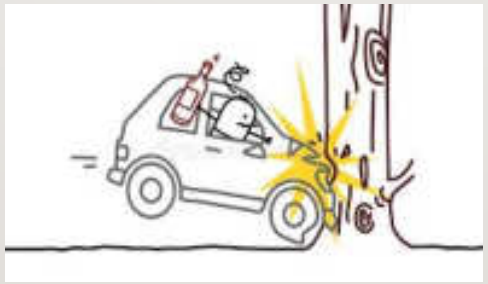
- Who should we ask?
- Philosophers are mainly interested in the **normative** side: what is a (good) explanation
  - Mainly discuss scientific explanations (but assume continuity)
  - Mainly set at the type level – seeking generalizations
  - A very long like of thinkers proposing solutions to this problem
- Psychologists care about the **descriptive** side: what do people see as an explanation & what properties do they value in an explanation
  - Mainly care about everyday explanations
  - Mainly set at the token level: why did this happen here and now
  - Became the focus of research much more recently – last 20-30 years

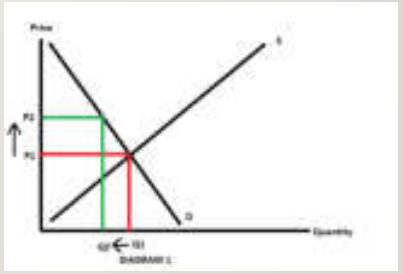




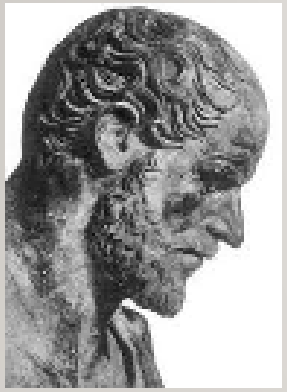
"My wife gave me this list of character flaws she wants you to fix."

WHAT IS THE TASK?





WHAT IS AN EXPLANATION FOR X?


Aristotle


Hempel


Salmon


Kitcher

- Citing the function, the material, the category or the (efficient) cause of X
- Producing an argument whose conclusion is X
- Stating everything that affects the probability of X
- Stating the causal history of X
- Showing how X fits a more general state of affairs

ARISTOTELEAN VIEW

Why is this round?

- **Efficient** explanation: cites the proximal mechanism of change - whatever brings it about
  - e.g. it was made by Pirelli
- **Final** (Teleological/Functional) explanation: cites functions and goals
  - e.g. it takes less energy to move it
- **Material** (Mechanistic) explanation: appeals to material or parts and processes
  - e.g. the molten metal is cast into a round mould
- **Formal** (Categorical) explanation: cites kind or category membership
  - it's a wheel

DEDUCTIVE-NOMOLOGICAL MODEL OF EXPLANATION

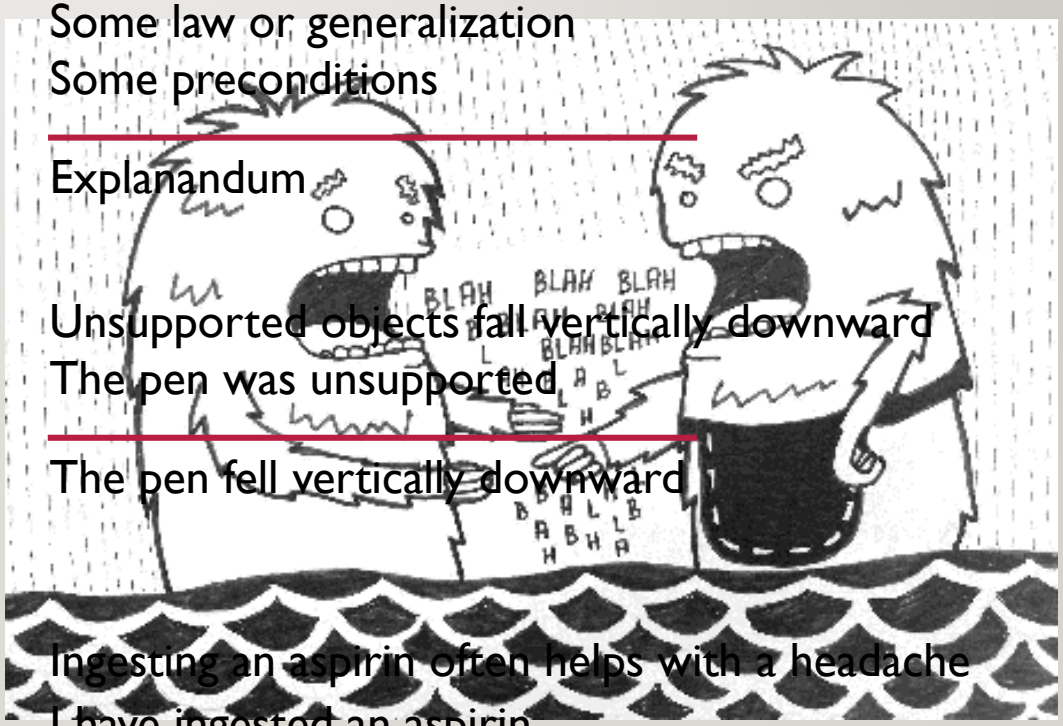
- An explanation is a logical (deductive) argument having a particular form:
  - The explanandum (the statement that we're explaining) is the conclusion
  - The explanans is a set of true statements that are the premises of the argument
  - The premises contain a law of nature
  - The explanandum must follow logically from the explanans
- A DN argument shows why the explanandum is to be expected. This expectation leads to understanding
- There is also a non-deterministic version known as the Inductive-Statistical (IS) model where the law states some statistical regularity.

Some law or generalization
Some preconditions

Explanandum

Unsupported objects fall vertically downward
The pen was unsupported

The pen fell vertically downward



Ingesting an aspirin often helps with a headache
I have ingested an aspirin

My headache has receded

WHAT IS A LAW?

Why do I wear trousers?

All men in this room wear trousers
I am a man
I'm in this room

I wear trousers

- Hempel warns against confusing laws with accidental generalizations
- But what exactly is a law is another point of contention
- What about exceptions?
  - All special sciences (biology, psychology, economics) make use of generalizations that allow for exceptions.
  - The IS model does not help
  - e.g. Why do pigeons fly?

PROBLEMS FOR DN MODEL

- "The glass tipped over because I hit the table"
- Is there a law?
- Hidden structure strategy: behind every explanation there is a DN argument in disguise – a sketch of a full argument
- Does the user of the explanation need to know the underlying details?
- If something is hidden, why should that be the DN model?

PROBLEMS FOR DN MODEL

Why am I not pregnant?

Men who take contraception pills daily don't stay pregnant

I am a man

I take contraception pills daily

I am not pregnant

Why did the salt dissolve in the water?

When salt is touched by the magic wand it dissolves in water

The salt was touched by the magic wand

The salt was emerged in water

The salt dissolved

Problem of explanatory relevance

PROBLEMS FOR DN MODEL

Why does the shadow has length s?

Why does the person has height h?

Linear propagation of light

Length of shadow is s

Angle of sun above the horizon is θ

Height of person is h

Problem of explanatory asymmetry

IS THE DN MODEL NECESSARY?

- The problems of relevance and asymmetry indicate that the DN model is not sufficient
- What is missing?

- Is the model necessary?

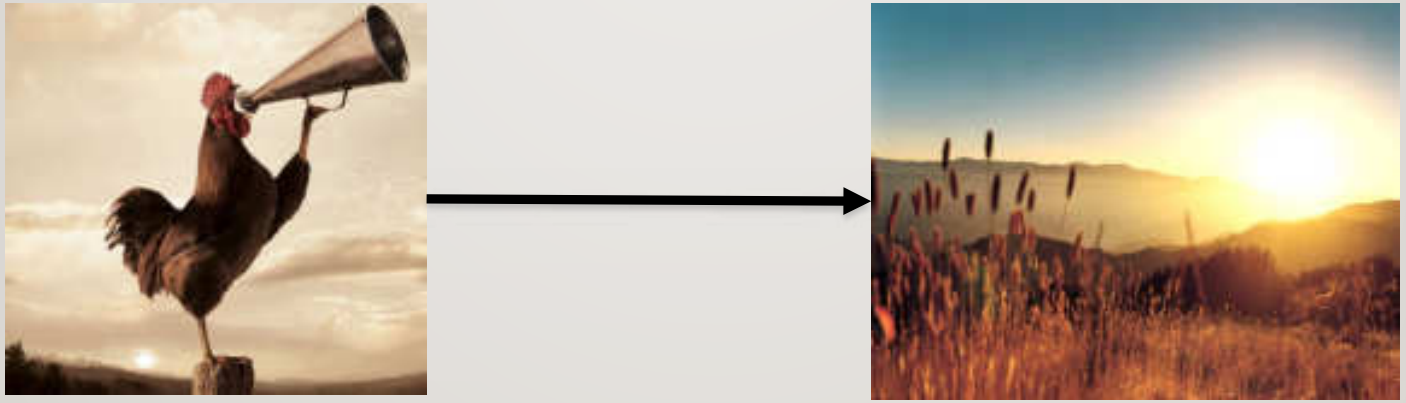
STATISTICAL RELEVANCE MODEL

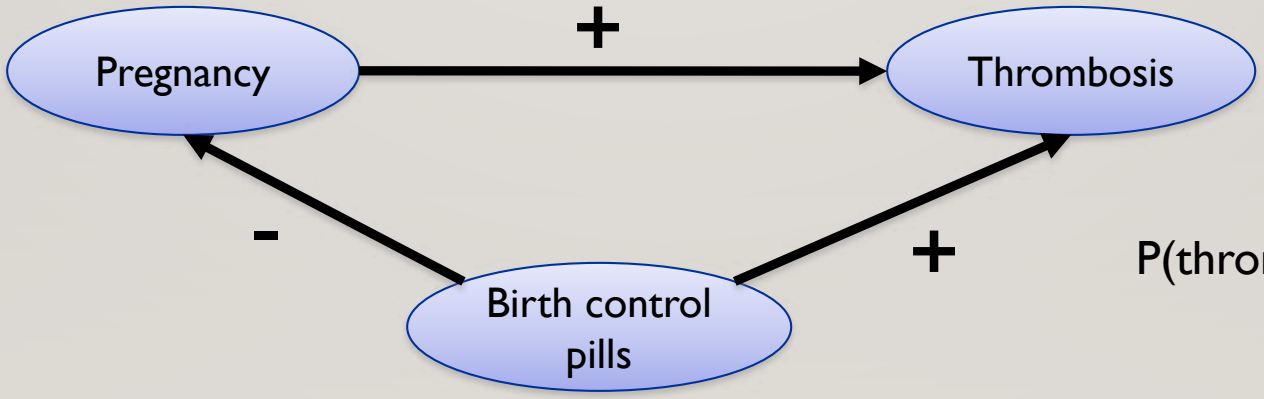
- Attempts to solve the problem of relevance through conditional dependence
- A is explanatory relevant to B if $P(B|A) \neq P(B)$
- Since P(pregnant | male & contraceptives) = P(pregnant | male), contraceptives are explanatory irrelevant to male pregnancies.
- Unlike the DN model an explanation is not an argument but the set of statistically relevant information: "irrelevancies are harmless in arguments but fatal in explanations"
- Unlike the IS model, it does not require for the explanandum to follow with high probability
  - if $P(B|A) - P(B) = 0.000001$, A is still explanatory

LOW PROBABILITY EVENTS

- Consider a very biased coin: P(heads | tossed)=0.9.
- Suppose that before tossed it is resting in either heads or tails, P(heads)=P(tails) = 0.5
- If it is tossed and comes up heads, the tossing explains the result since P(heads | tossed) ≠ P(heads)
- If it is tossed and comes up tails, the tossing explains the result since P(tails | tossed) ≠ P(tails)
- So, tossing explains both results!
  - Perhaps, we're missing statistically relevant variables, but what about macroscopic event (i.e. not quantum phenomena?)

CORRELATION AND CAUSATION





P(thrombosis | pills) = P(thrombosis)

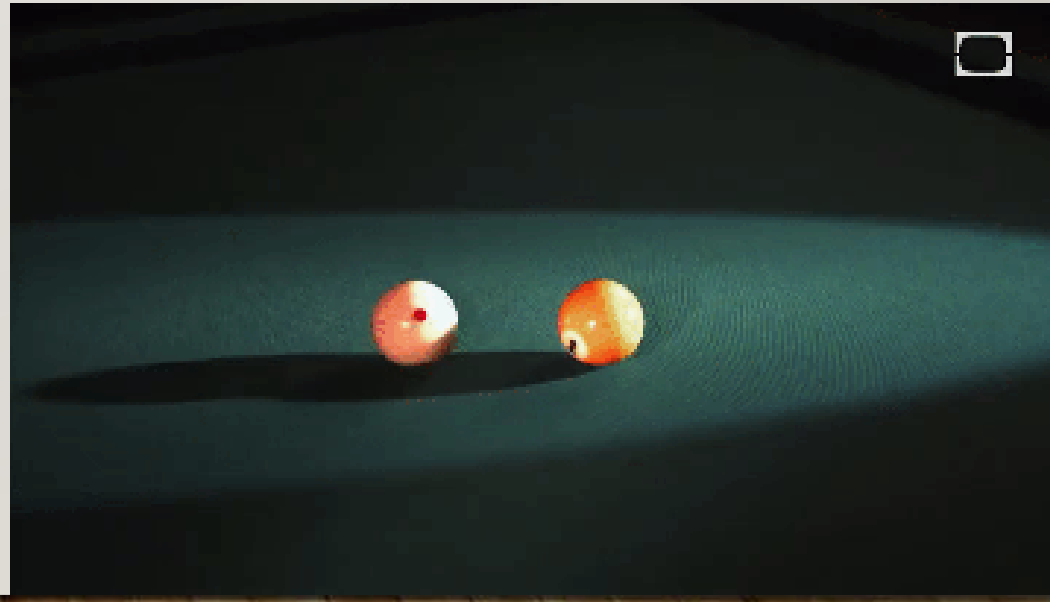
CAUSAL-MECHANICAL MODEL OF EXPLANATION

- Most current theories of explanation include some notion of causation.
- But what is a causal interaction?
- Causation is a notoriously difficult concept to define
- Most algorithms in AI (Feature importance, Partial Dependence, SHAP etc) use a counterfactual notion of causation
  - A→B, if in the absence of A, B would not occur
  - problems with pre-emption, overdetermination
- Wesley Salmon suggested a specific notion of causal explanation based on a process theory of causation (also Bertrand Russell, Philip Dowe)

CAUSAL-MECHANICAL MODEL OF EXPLANATION

- An explanation of E will trace the causal processes and interactions that lead to E, thus showing how E fits into a causal nexus
  - A causal process is a physical process that transmits a mark
  - A causal interaction is an interaction between two causal processes that leaves a mark (modifies the structure) in both

- A mark is any modification that persists without further interaction



PROBLEMS FOR CM MODEL

- But which property should we pick out?
- Why pick the momentum or direction and not the chalk trace?
- Talking contraceptive pills leads to a permanent biochemical change. Problem of explanatory relevance, again.
- In later versions, Salmon redefined the mark to refer to the transmission of some conserved quantity (e.g. angular/linear momentum, charge etc)
- But, in general, not all causal events in the past of E are explanatory relevant to E
  - (although the theory does a good job distinguishing causal processes and pseudo-processes)

PROBLEMS FOR CM MODEL

Like other process theories it has trouble with certain types of causation:

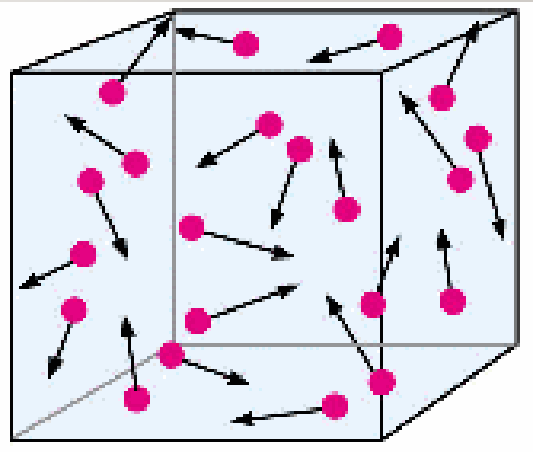
Causation by omission

common in legal, medical, scientific explanations

Double Prevention

PROBLEMS FOR CM MODEL

- Level of abstraction (complex system equilibria)
- Explain the temperature in the container
  - only molecules are processes
  - but the ideal gas law moves up a level of abstraction
  - + the final temperature would be the same irrespective of the initial positions/momenta (explanatory irrelevant?)
- Explain the price of orange
  - Depends on millions of interactions (with cash, card etc)
  - None of the physical interactions look explanatory relevant



UNIFICATIONIST ACCOUNT OF EXPLANATION

- An explanation unifies different phenomena
- "Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same pattern of derivation again and again, and in demonstrating this, it teaches us how to reduce the number of facts we have to accept as ultimate." (Kitcher)
- An explanation shows connections between previously unrelated things
  - Newton's theory of gravitation: terrestrial and celestial things attract each other in the same way
- The aim is to use the fewer sentences possible to derive the most phenomena possible
- Gives us a way of comparing competing explanations: Reject explanations for phenomena already explained unless the new explanatory pattern explains more phenomena

UNIFICATION AND CAUSATION

- Explanation is more fundamental (primitive) than causation.
- Kitcher argues that we call "causal" things that are explanatory
- The explanation that salt dissolves because it was hexed is worse than the explanation that it dissolved because of its molecular structure because the latter explains all cases of salt dissolution = more unifying.

- We should reject the explanation that the shadow length explains the person's height because not all things have shadows.

- Causal asymmetry is due to explanatory asymmetry: the maximally unifying explanations show us the direction of causation.
- Causal information is culturally transmitted and is the result of a long process of explanatory unification
  - People make arguments about different systemizations, and the most unifying ones win out. Causal beliefs are justified based on whether they fall out of the best systematization

SHADOWS AGAIN

- The theory by which causes are explained by their effects makes use of the same assumptions as the theory in which effects are explained by their causes
- Compare explaining the current position of the sun from its previous position, mass, speed etc with explaining its current position from its future position: The sun is here now because it will be there in the future!

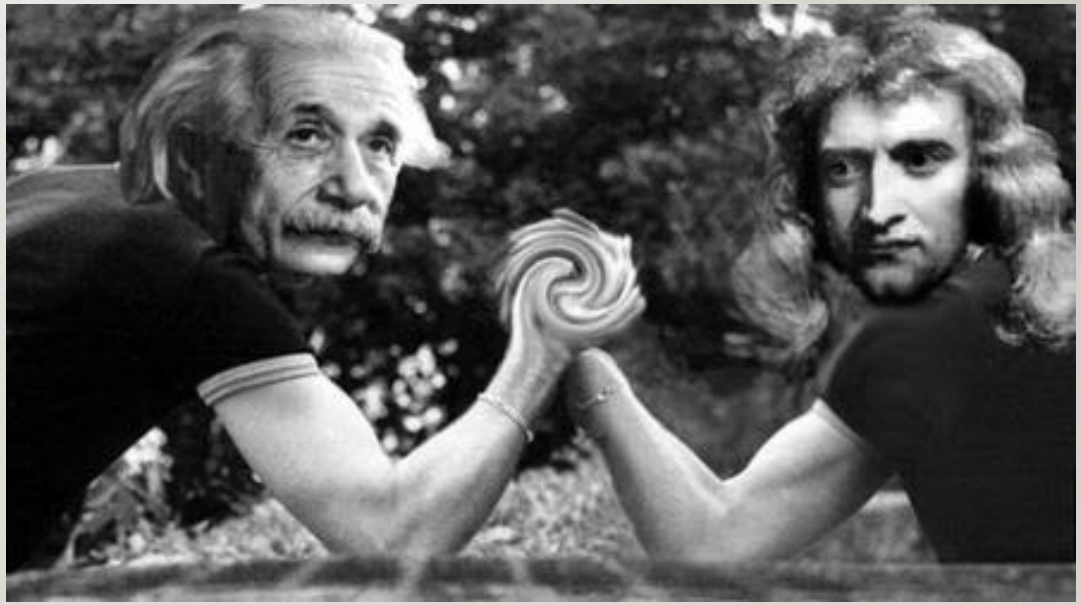
DO ALL UNIFICATIONS EXPLAIN?

- The same mathematical framework can be applied to different phenomena (unifying)
  - e.g. Lagrange's equations describe both gravitation and electrical circuits. Is there a common explanation?
  - e.g. a system of classification (e.g. x is a bird) allows prediction of properties (e.g. x has wings), but does the classification explain the properties?
  - e.g. factor analysis allows reduction of statistically relevant factors but correlation is not causation

Unification is about descriptive economy or information compression, not necessarily about explanation.

ARE NON-UNIFYING EXPLANATIONS NON-EXPLANATORY?

- Is Newton's theory of gravitational forces non-explanatory because general relativity is more unifying?
- If we say that perhaps Newton's theory is not as good but still explanatory we lose our criterion for picking explanations
  - perhaps contraception for males is a fine explanation but not as good



CAUSAL KNOWLEDGE FROM EXPLANATION?

- Is the source of all causal knowledge cultural transmission?
- What about causation from experience or even perception?
- Do we really compare different systematizations to find the most unifying one?

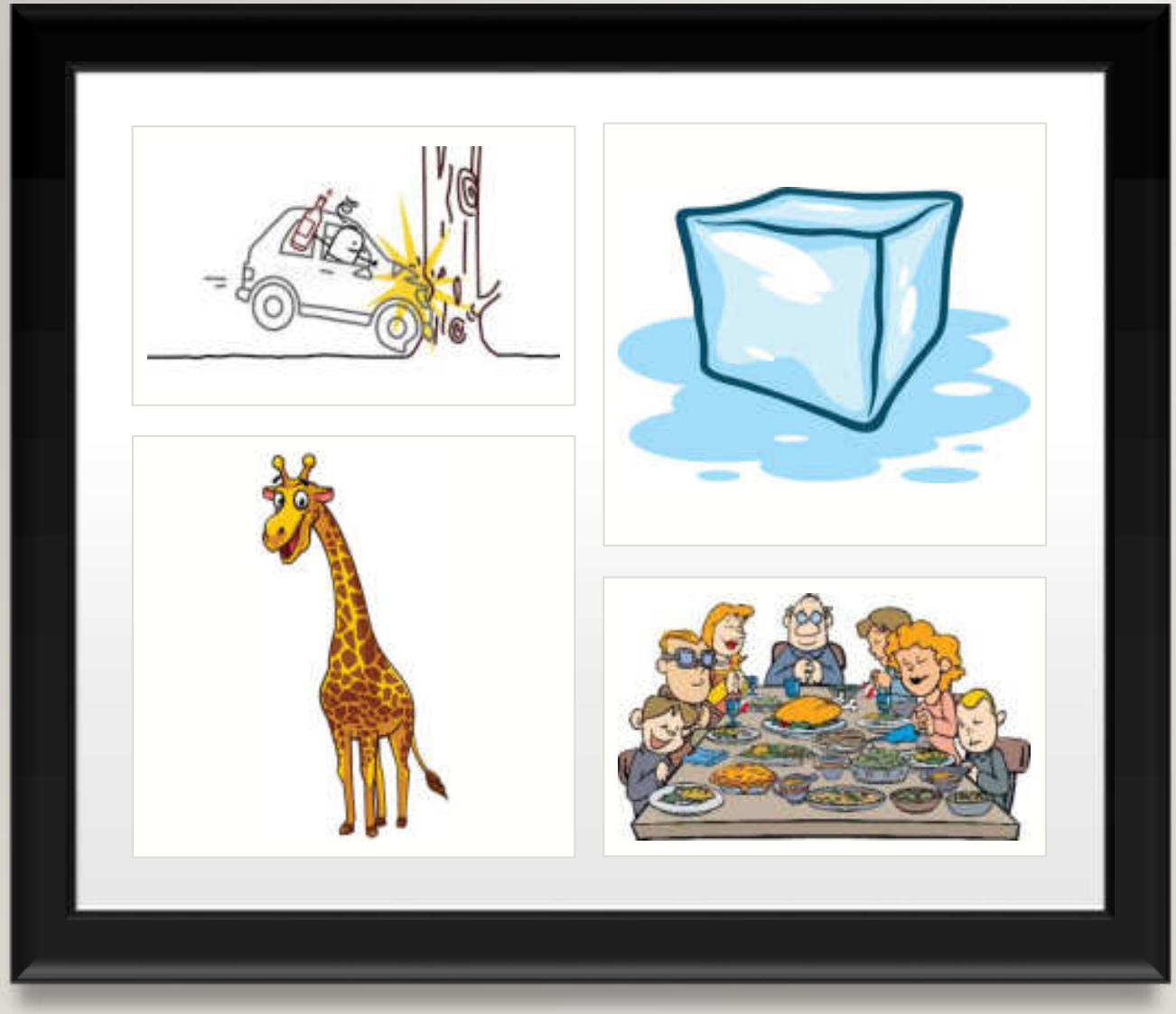
WHAT ABOUT PRAGMATIC CONSIDERATIONS?

- All the theories discussed assume that we can formalize explanations in a way that does not depend on the domain or the context or the psychology of the participants (interests, aims, background knowledge):
  - Who is giving and who is receiving the explanation?
    - Try explaining a neural network to a child, a psychologist or a developer.
    - Explaining an accident to a policy maker or to an insurer
- After all, we can imagine situations where the length of a shadow does explain the height of the object, if, for example, the audience wants to know how did the explainer reached the conclusion: "Why do you say that the height is X?"

EXPLANATIONS IN PSYCHOLOGY

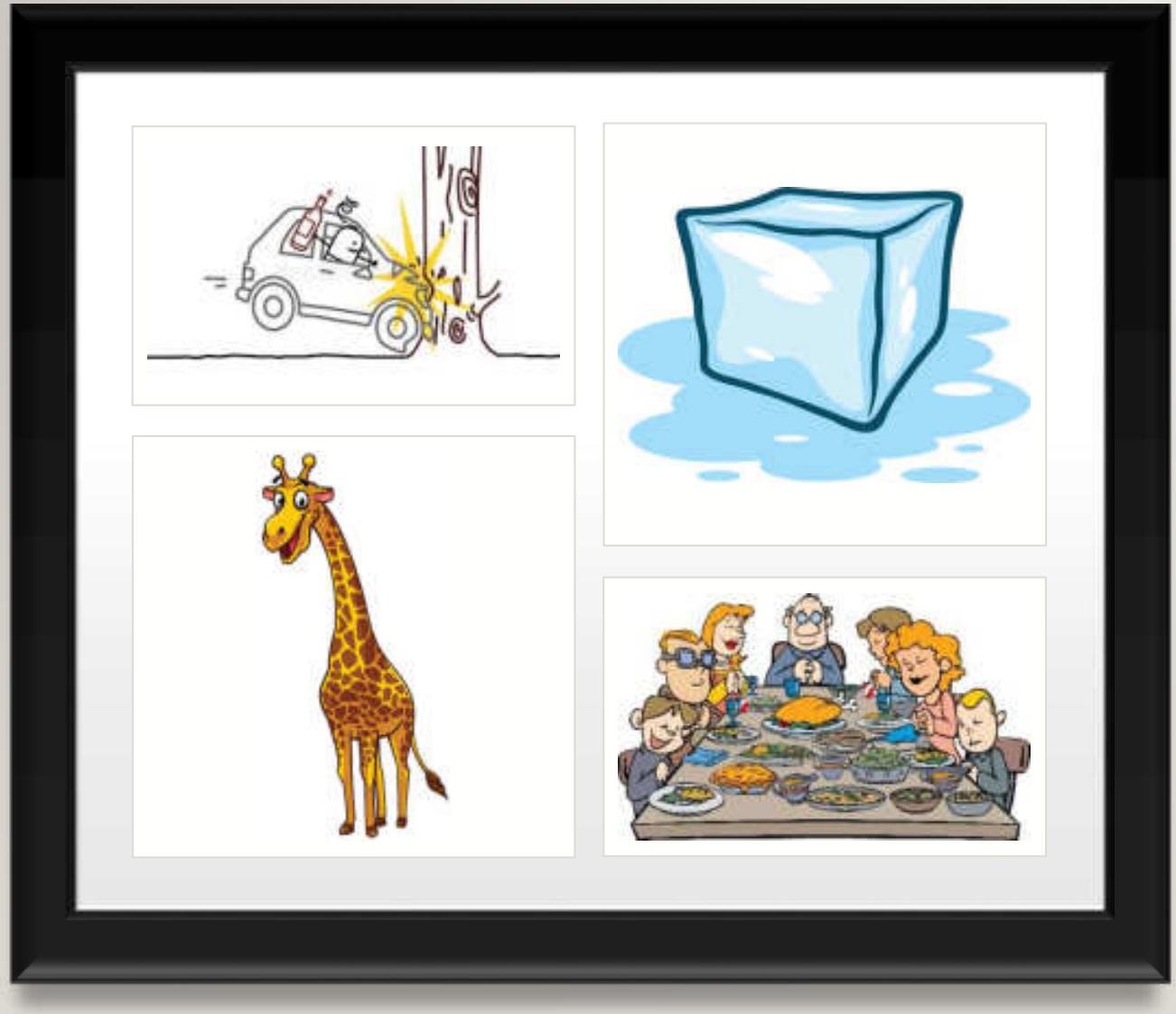
- What do psychologists think about explanations?
  - Answers to why or how questions (Wellman 2011)
  - Judgements about why an outcome has occurred (Krull & Anderson 2001)
  - Hypotheses about the causes of the explanandum
  - Explanations accommodate novel information in the context of prior beliefs, and do so in a way that fosters generalization (Lombrozo, 2006)
  - A family-resemblance term picking out a cluster of related phenomena (Lombrozo 2012)
- Explanation as a product or as a process?
  - product: proposition or judgement that addresses an explicit or implicit request for explanation (that's what philosophy cares about)
  - process: a cognitive activity that generates explanation products

- Complete or selected explanations?
  - Partial structure (cf hidden structure strategy) or big-bang included?
  - Do we know the complete explanation?
- Contrastive nature of explanations:
  - Explanations typically identify conditions that differentiate what is being explained (the explanans) from a counterfactual case in which the contrast holds.



TYPES OF EXPLANATIONS

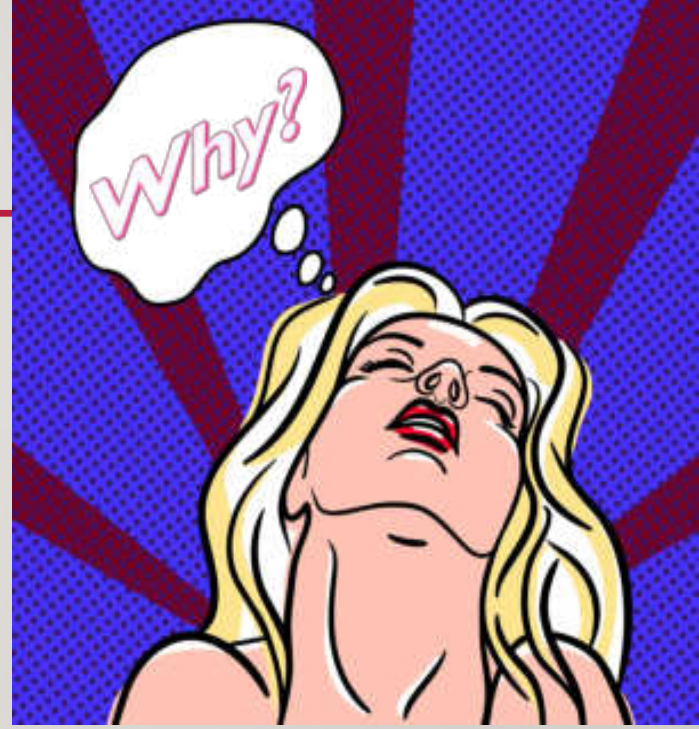
- Why seat belts prevent traffic fatalities?
- Why salt melts ice?
- Why giraffes have long necks?
- Why Aunt Edna insulted Uncle Billy at the family holiday dinner?



TYPES OF EXPLANATIONS

- Different domains
- Different causal patterns (common cause, chain etc)
- Different explanatory stances / modes of construal
- Emotion/value-laden: same properties but different threshold for acceptance

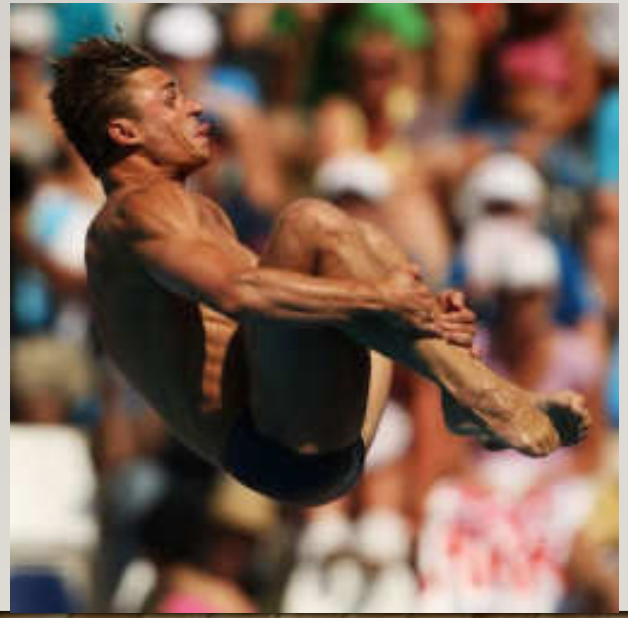
WHY DO WE SEEK EXPLANATIONS?



- Prepare for the future:
  - "a kind of distance-receptor in time, which enables organisms to adapt themselves to situations that are about to arise" - Craik (1943)
  - Why is the road slippery? Because of the cold wet weather.
  - Why does the remote fail? Because it has not batteries (diagnostic)
- But we seek explanations even for one-off events: e.g. why did the assassination of Duke Ferdinand led to WWI?
  - Understanding is not the same as predicting – common warning in the study of history
  - Constellation of unique and unusual factors do not afford predictions
  - Deciding responsibility – assigning blame
- Rationalizing actions: didn't vote because my vote wouldn't matter anyway, punished the child for its own good
- Aesthetic pleasure: explain a poem or a painting – a more polished lens through which to view the explanandum (Keil 2006)
- Male (2004): We seek explanations to find meaning (reconcile contradictions) and to share meaning (manage social interactions)
- Gopnik: explanations as orgasms. Evolution gives us satisfaction from explanations as an incentive to engage in theory formation
- Whatever the functions of explanations, explainers don't necessarily have the goal of fulfilling them (Lombrozo , 2012)

EXPLANATORY STANCE / MODE OF CONSTRUAL

- Dennett (1987) – ref Aristotelian causes
  - Mechanical stance: objects and interactions
  - Design stance: purposes and functions (functional/teleological)
  - Intentional stance: requires beliefs/desires/intention (BDI calculus)



- Mechanical stance: physics of rotating objects
- Design stance: purpose of pulling in the limbs close to body
- Intentional stance: he wants to achieve the best dive

A PREFERENCE FOR TELEOLOGY

- Why do trees shed their leaves?
  - Because the winter wind or snow will damage them.
- Teleological Bias: People and children often default to teleological explanations (Kelemen 1999; Kelemen & Rosset, 2009; Piaget 1929)
  - What are tigers made for?
    - For walking and being in a zoo (Heider & Simmel 1944)
  - Why are rocks pointy?
    - So that animals won't sit in it and smash it (Kelemen 1999)
  - Mountains are for climbing, clouds for producing rain (Kelemen, 1999)
  - Adults under time pressure: The sun makes light so that plants can photosynthesize (Kelemen & Rosset, 2009)
- Perhaps because we care predominantly about intentional agents
- Kelemen & Rosset argue that teleological explanations are an explanatory default that is suppressed but not replaced by scientific alternatives
- But must be consistent with causality: must play a causal role (Lombrozo & Carey 2006)
  - Why does the tiger have stripes? For camouflage!
  - Camouflage is teleological but also a shorthand for a causal process

EXPLANATORY VIRTUES: WHAT MAKES FOR A GOOD EXPLANATION?

- The following properties have been suggested as properties of good explanations:
  - No Circularity
  - Coherence
  - Relevance
  - Match the epistemic status of the audience
  - Simplicity
  - Generality
- Are people sensitive to these "virtues"?

CIRCULARITY

- "This diet pill works because it helps people lose weight,"
- Children from the age of 5-6 detect circularity – robustly from age 10 (Baum et al 2008)
- But not always easy to detect
  - Allen: The Evanston City Council should make it illegal to tear down the city's old warehouses.
  - Beth: What's the justification for preserving them?
  - Allen: The warehouses are valuable architecturally.
  - Beth: Why are they so valuable?
  - Allen: The older buildings lend the town its distinctive character.
  - Beth: But, what's the reason the warehouses give it character?
  - Allen: The warehouses are important architectural examples

Beth: But, anyway, why do you personally like these warehouses?

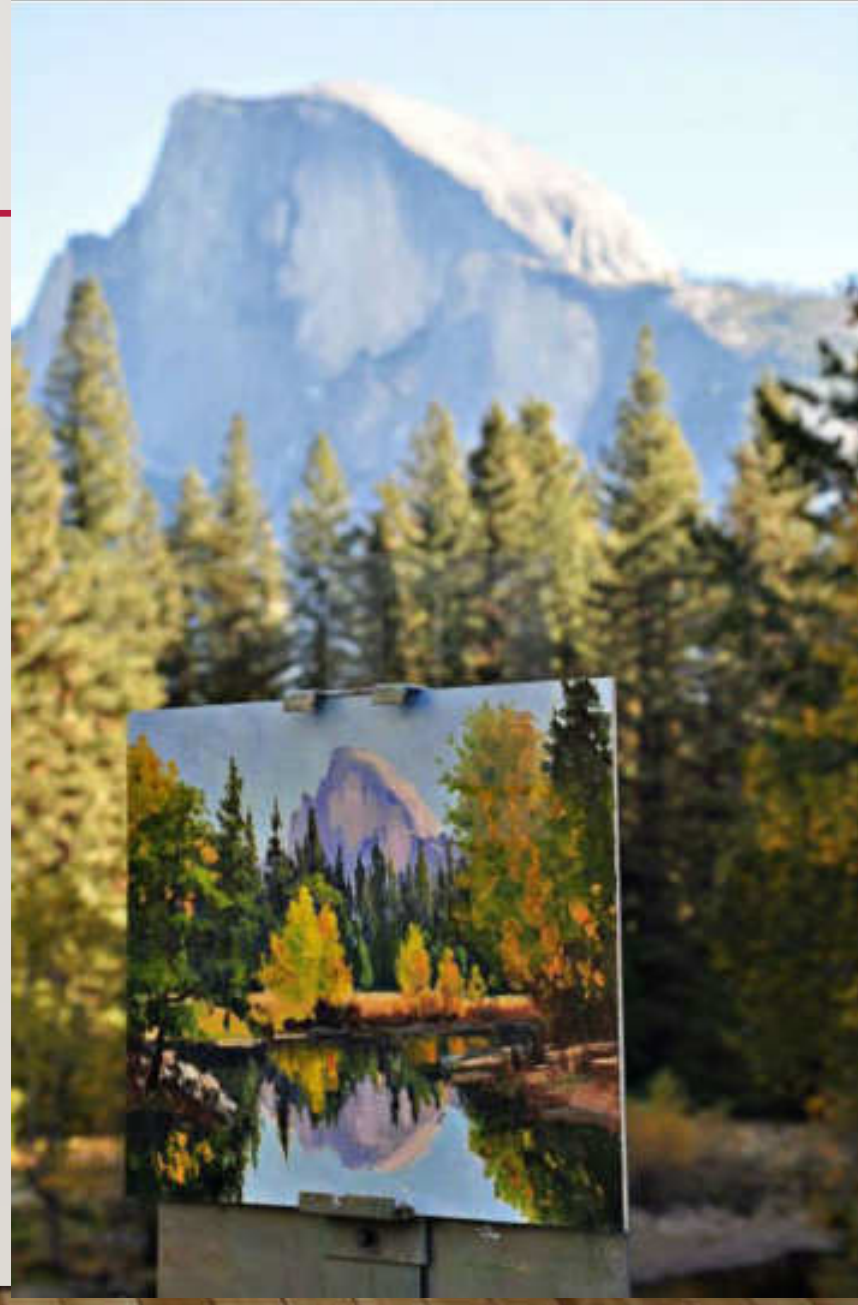
Rips 2002

COHERENCE

- Different elements of explanation must hang together, have internal consistency
- An explanation must be consistent with prior knowledge and current evidence
- A set of elements is coherent to the extent that each element in a set, positively constrains other ones, often causally (Thagard 2000)
- But are we sensitive to coherence?
  - Holism Problem (Fodor 1998): explaining how a bicycle works we need to say how its various mechanical elements interact and constrain each other. But these elements are also constrained by human anatomy, physiology and goals. Also rigidity of surfaces, economics of construction etc

PARTIAL EXPLANATIONS

- For most things natural or artificial, the full set of relations to be explained is huge, much beyond what an individual can grasp, represent or remember.
- Even in science, explanations are always incomplete, science is often driven by hunches and vague impressions
  - We often feel that a theory is crystal clear in our head until we start writing (Rozenblit & Keil 2002)
- Unlike scientists, everyday people rarely write down their understanding of a theory, rarely doubt their understanding.

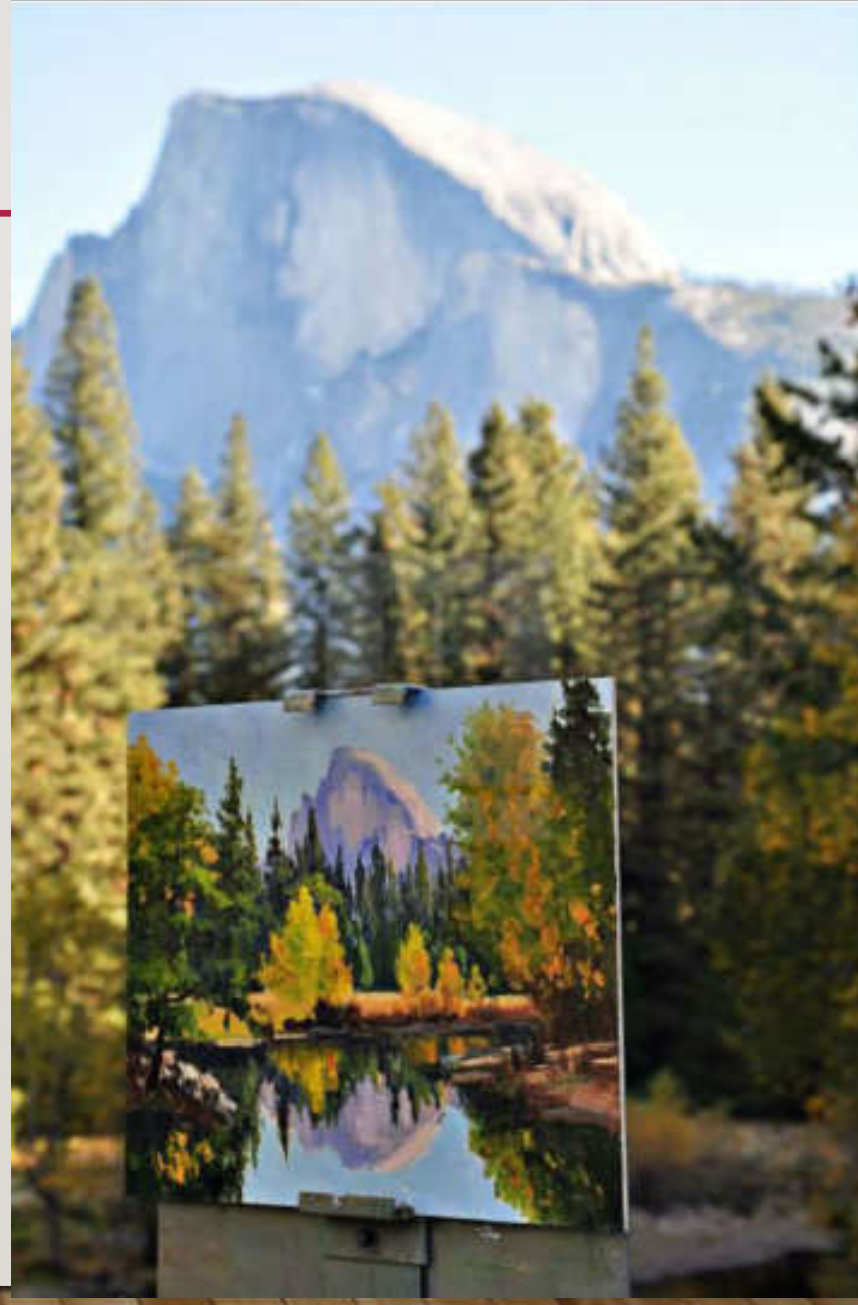


TIGER EXAMPLE

Our mental representations are skeletal and incomplete – but we don't have this impression

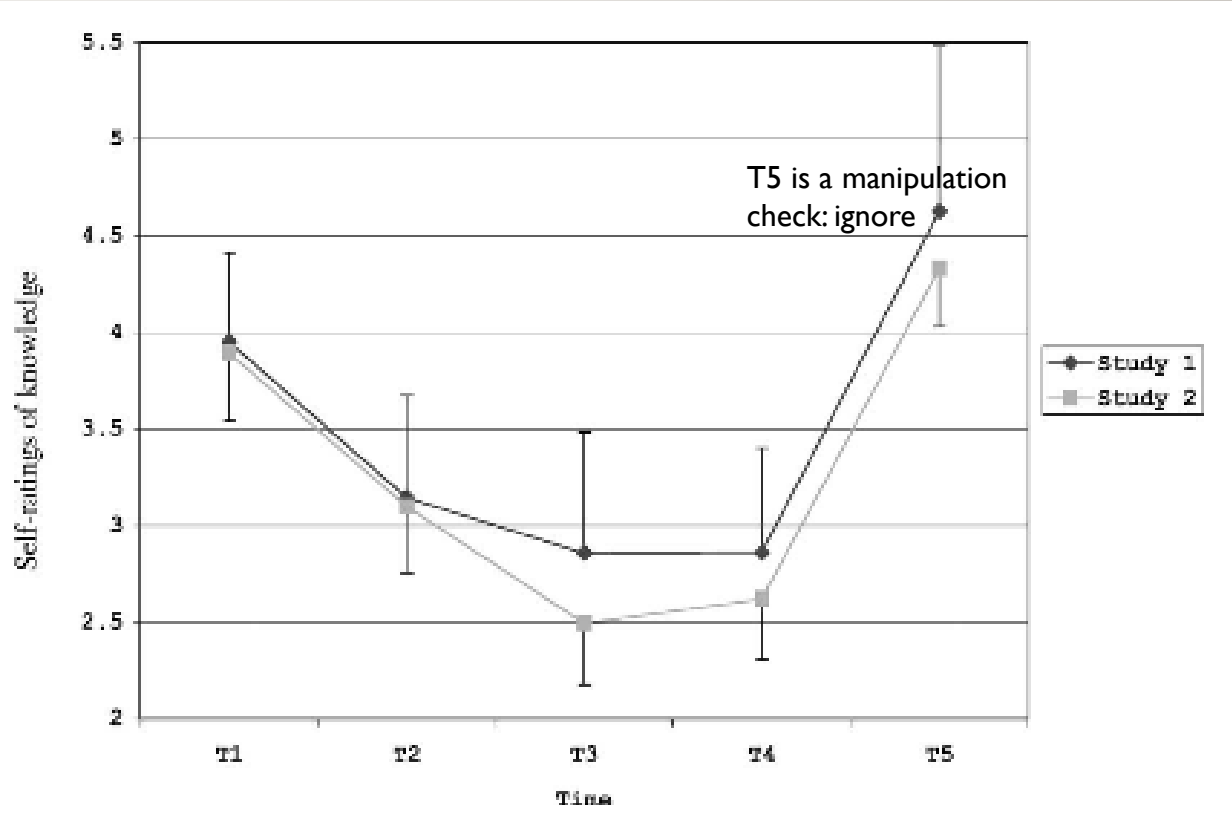
PARTIAL EXPLANATIONS

- But how is it that are we most often successful with our explanations?
- Some compression is needed
- We tend to overestimate the depth of our own understanding



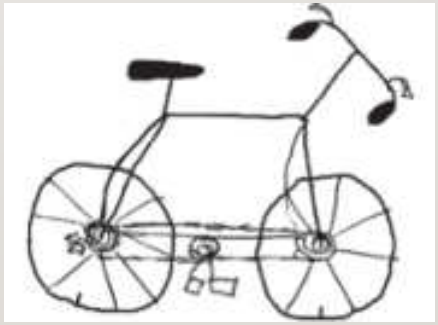
ILLUSION OF EXPLANATORY DEPTH (IOED)
(ROZENBLIT & KEIL 2002)

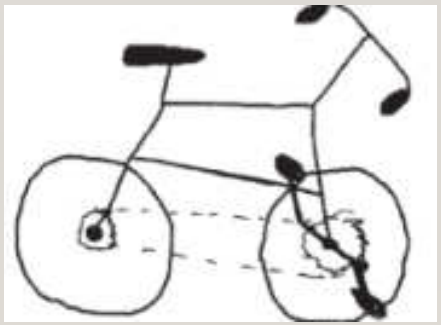
1. Rate how well you understand how a zipper, a flush toilet, a piano key or a speedometer works?
2. Give a detailed, step-by-step causal explanation of how it works. Rate again
3. Answer a diagnostic question that shows your knowledge. Rate again
4. Read expert explanation. Rate again



ILLUSION OF EXPLANATORY DEPTH

1. How much do you know about how bicycles work?
2. Draw pedals, chain and missing part of frame on a sketch

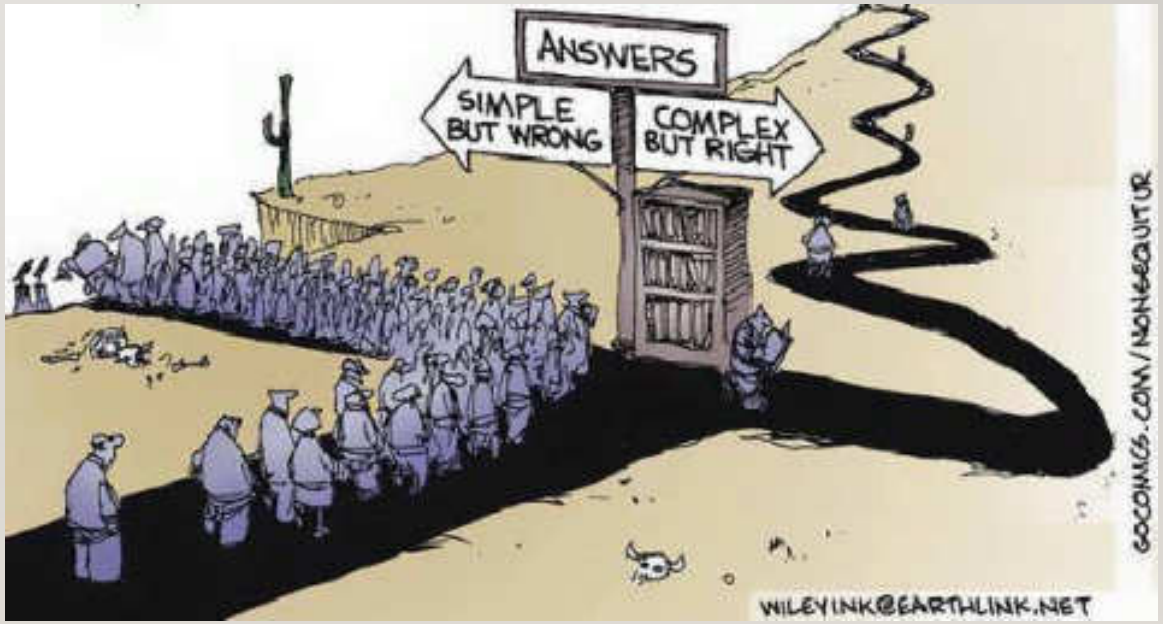




Rebecca Lawson (2006)

ILLUSION OF EXPLANATORY DEPTH

- Fernbach et al (2013) asked participants to (a) express their agreement (b) rate their understanding before/after they give an explanation about 7 policies such as
  - imposing unilateral sanctions on Iran for its nuclear program,
  - establishing a cap-and-trade system for carbon emissions,
  - instituting a national flat tax
- Not only ratings of understanding went down after the explanation (IOED) but position extremity also went down.



ILLUSION OF EXPLANATORY DEPTH

- Even if our functional understanding is fragmentary and shallow, it (mostly) works. How?
- Groupthinking: we rely heavily in expertise of other minds
- In Zemla et al (2017) we took explanations from Reddit "Explain like I'm 5" and asked participants to judge their overall quality and also rate them across a number of dimensions (e.g. complexity, articulation, coherence, generality, truth). One of the most important predictors of explanation quality was "perceived expertise", i.e. whether participants believed the explanation was written by an expert.
- Perhaps, when we believe the explainer is an expert and provided we trust them, our threshold for acceptable explanations is reduced.

RELEVANCE

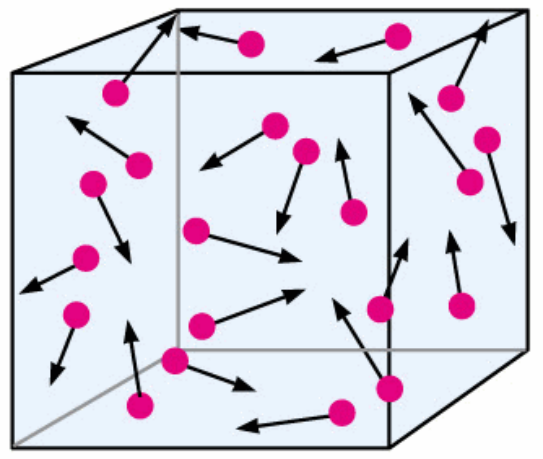
- Relevance: Central problem in the philosophy of science, and a hard question in most cases
- Causality is central, but again what is causally relevant:
  - Does the oxygen explain the fire?
  - Is the Bing Bang a cause for everything?

WHAT GOES INTO A GOOD EXPLANATION?

with Dave Lagnado (UCL), Steve Sloman and Jeff Zemla (Brown University)

LEVEL OF DETAIL / LEVEL OF ABSTRACTION

- For some philosophers including everything that has a causal role is ideal but unattainable.
  - Railton (1981): Abstraction is a compromise
    - Nowak (1992): Science works through concretization: We start from a vague description and keep adding information until we get "the true causal story"

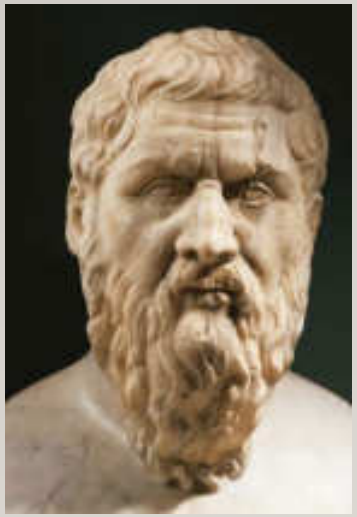


HYPERCONCRETE EXPLANATIONS

- Garfinkel (1981): Hyperconcrete explanations are not merely too good to be true (impractical) but are actually "too true to be good"
  - Think of maps
- Strevens (2007): Good explanations must lie
  - In his Kairetic account, Strevens describes that in order to generate the optimal explanation
    1. we include every imaginable event
    2. we remove and abstract everything that makes no difference to whether or not the explanandum occurred.
- Garfinkel, Strevens, Woodward, Hitchcock , Weslake: **What matters is counterfactual dependence not causal influence**

WHAT IS A GOOD EXPLANATION?

How much detail is needed?



Philosophical view
What should be included in a good explanation?

Only factors that made a difference to the explanandum

Everyday view
How people evaluate explanations depending on what is included in a good explanation?

THE SEDUCTIVE ALLURE OF NEUROSCIENCE

Weisberg et al (2008) asked people to rate the following two explanations about the curse of knowledge:

The researchers claim that this "curse" happens because subjects have trouble switching their point of view to consider what someone else might know, mistakenly projecting their own knowledge onto others.

vs.

Brain scans indicate that this "curse" happens because **of the frontal lobe brain circuitry known to be involved in self-knowledge**. Subjects have trouble switching their point of view to consider what someone else might know, mistakenly projecting their own knowledge onto others.

Adding irrelevant neuroscientific information increased the judged quality of explanations for both naïve adults and neuroscience students (not experts though).

CONCRETENESS AND ABSTRACTION

- Present the description of an event
- Ask participants to evaluate 3 explanations

Concrete	Abstract	Irrelevant
Only causally relevant events with high precision	Only counterfactually relevant events	Same as concrete but includes irrelevant information

- Ask participants to evaluate the causal relevance of each factor mentioned in the description and the explanations.

CONCRETENESS AND ABSTRACTION

Times & Citizen

WEDNESDAY, APRIL 15, 2015

LANDSLIDE IN ROCHESTER

In a village outside Rochester, New York, a landslide caused the destruction of one home and the evacuation of 35 others. The destroyed house was built in 2012.

Nobody has been injured. According to statement from the governor's office, state officials are working with local municipalities in the affected areas. County road crews have already begun the cleanup effort.

Here's what is currently known about the landslide:

The hill, situated 5 miles north of the premises of the annual Lilac festival, had a 37 degree slope.

The hill consisted mainly of light brown sandy particles with diameter 2/64 of an inch.

The vegetation was non-edible and covered 13% of the hill.

Times & Citizen

WEDNESDAY, MAY 15, 2015

POOR WEATHER AND BUGS AFFECT STRAWBERRY CROPS

The strawberry market is growing 10-15% a year, and fresh berries are now consumers' most popular fresh fruit. However, British strawberry growers in some parts of the country are having their worst season in years.

Strawberry grower Sandy Booth, from Hampshire's New Forest says his crop usually produces more than 2000 tonnes. But he says he's probably lost between 50-100 grams of berries per plant in recent months.

Here are some facts about this year's strawberries:

The mean temperature when the white strawberry flowers started to grow was 2 degrees Celsius.

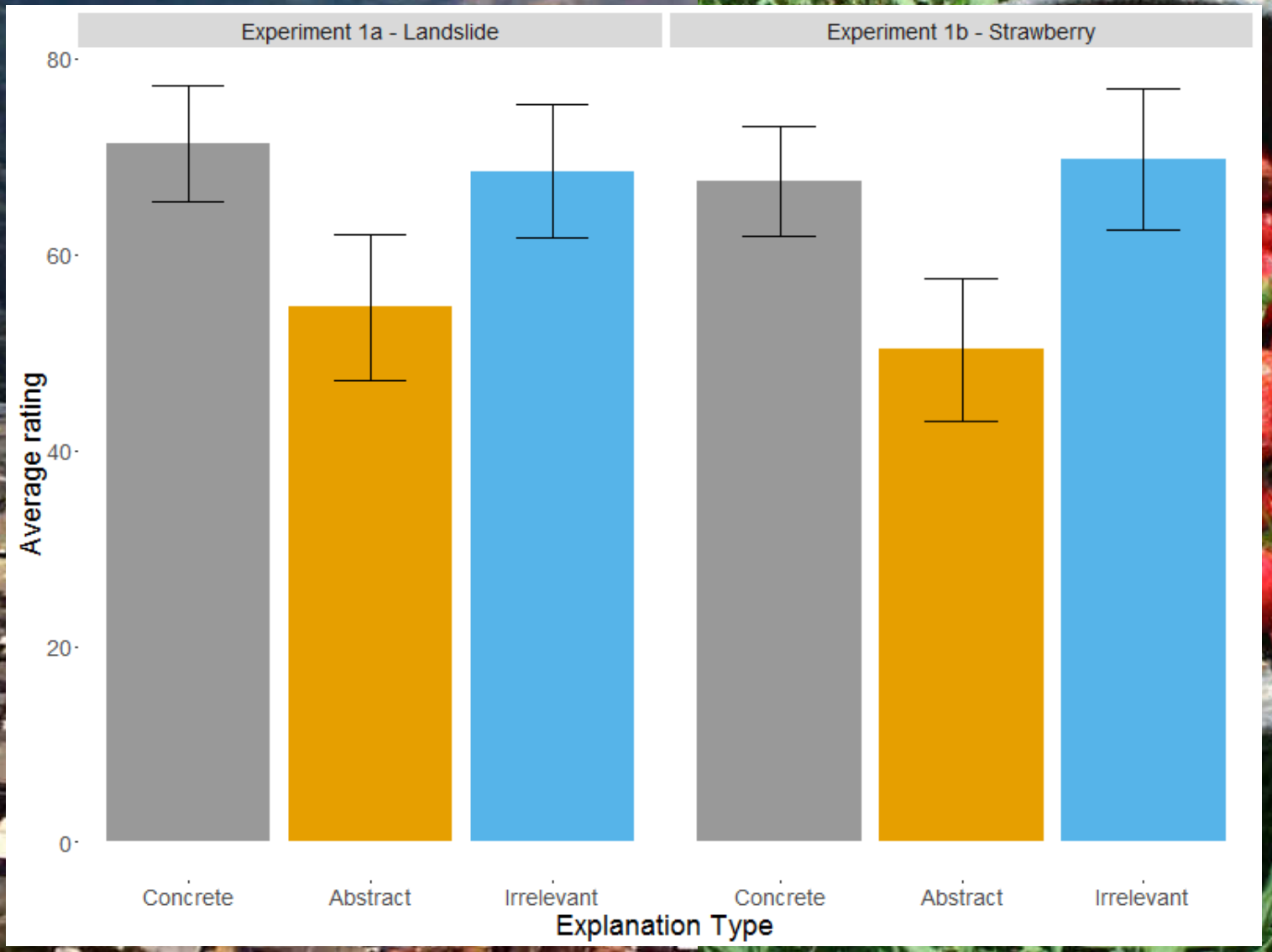
There has been a 27% increase in the frequency of attacks by the strawberry bug (phytonemus pallidus) which is only about 0.25mm is size.

77mph winds were blowing from the east just as the fruit was starting to ripen.

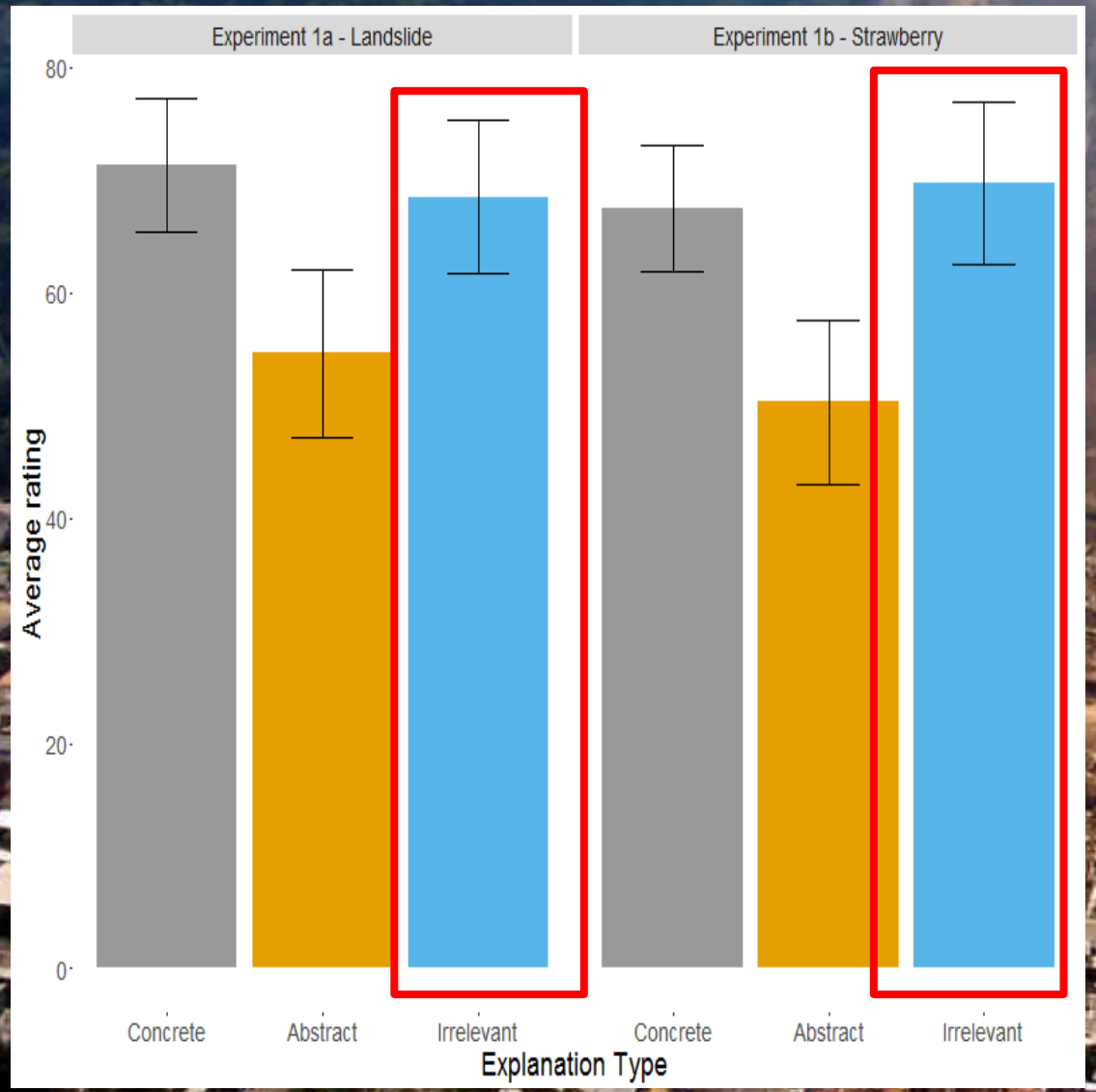
EXPLANATIONS

Concrete	Abstract	Irrelevant
The fact that the hill, consisted mainly of sandy particles with diameter 2/64 of an inch meant that the soil was unstable. The vegetation covering 13% of the hill did not withhold the rainwater causing soil erosion. Finally, the force of gravity acting down the 37 degree slope overcame the resistance of friction thus triggering the landslide.	The fact that the hill, consisted mainly of fine sandy particles meant that the soil was unstable. The sparse vegetation did not withhold the rainwater causing soil erosion. Finally, the force of gravity acting down the steep slope overcame the resistance of friction thus triggering the landslide.	The fact that the hill which was 5 miles north of the premises of the annual Lilac festival, consisted mainly of light brown sandy particles with diameter 2/64 of an inch meant that the soil was unstable. The non-edible vegetation covering 13% of the hill did not withhold the rainwater causing soil erosion. Finally, the force of gravity acting down the 37 degree slope overcame the resistance of friction thus triggering the landslide.

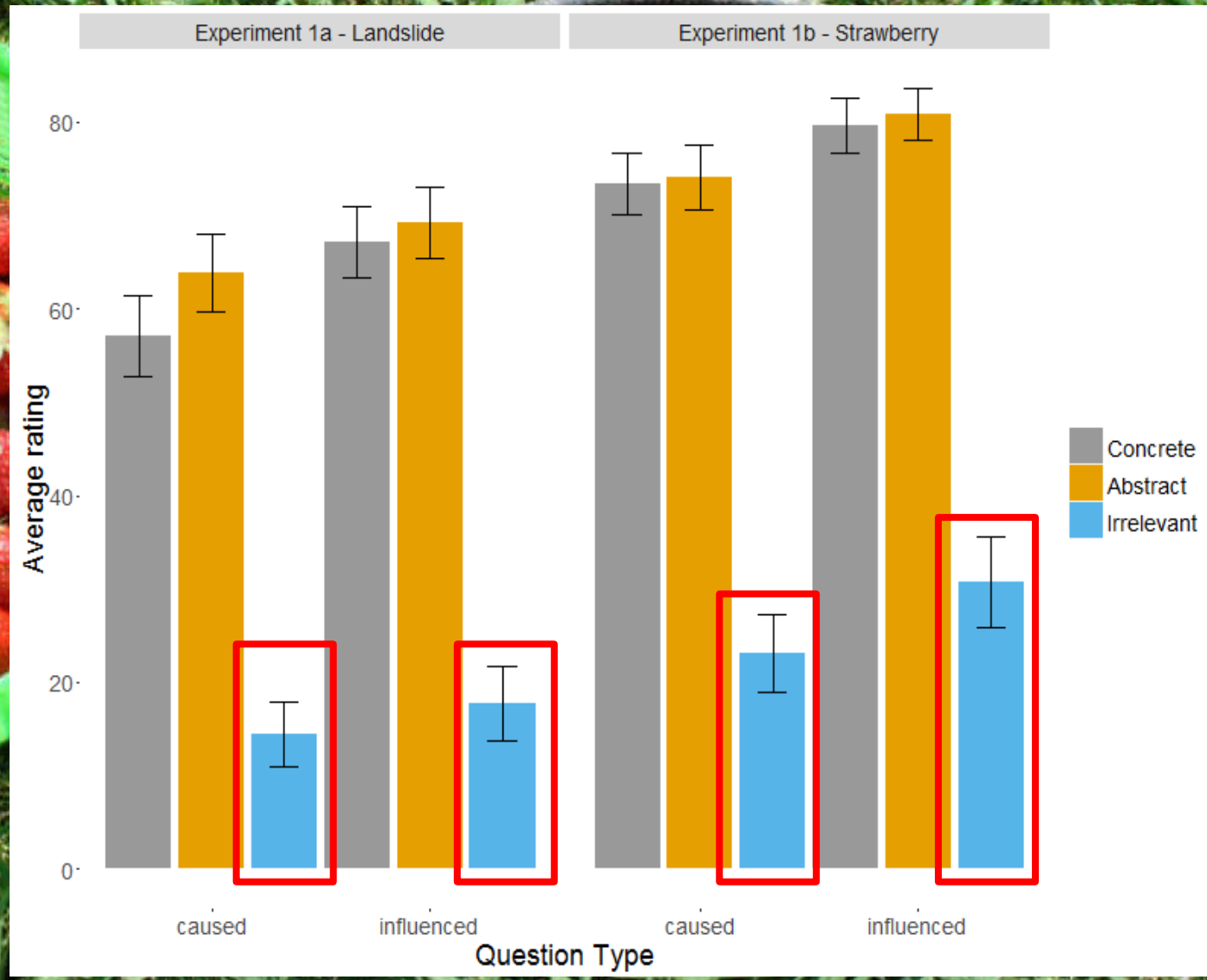
EXPLANATION RATINGS



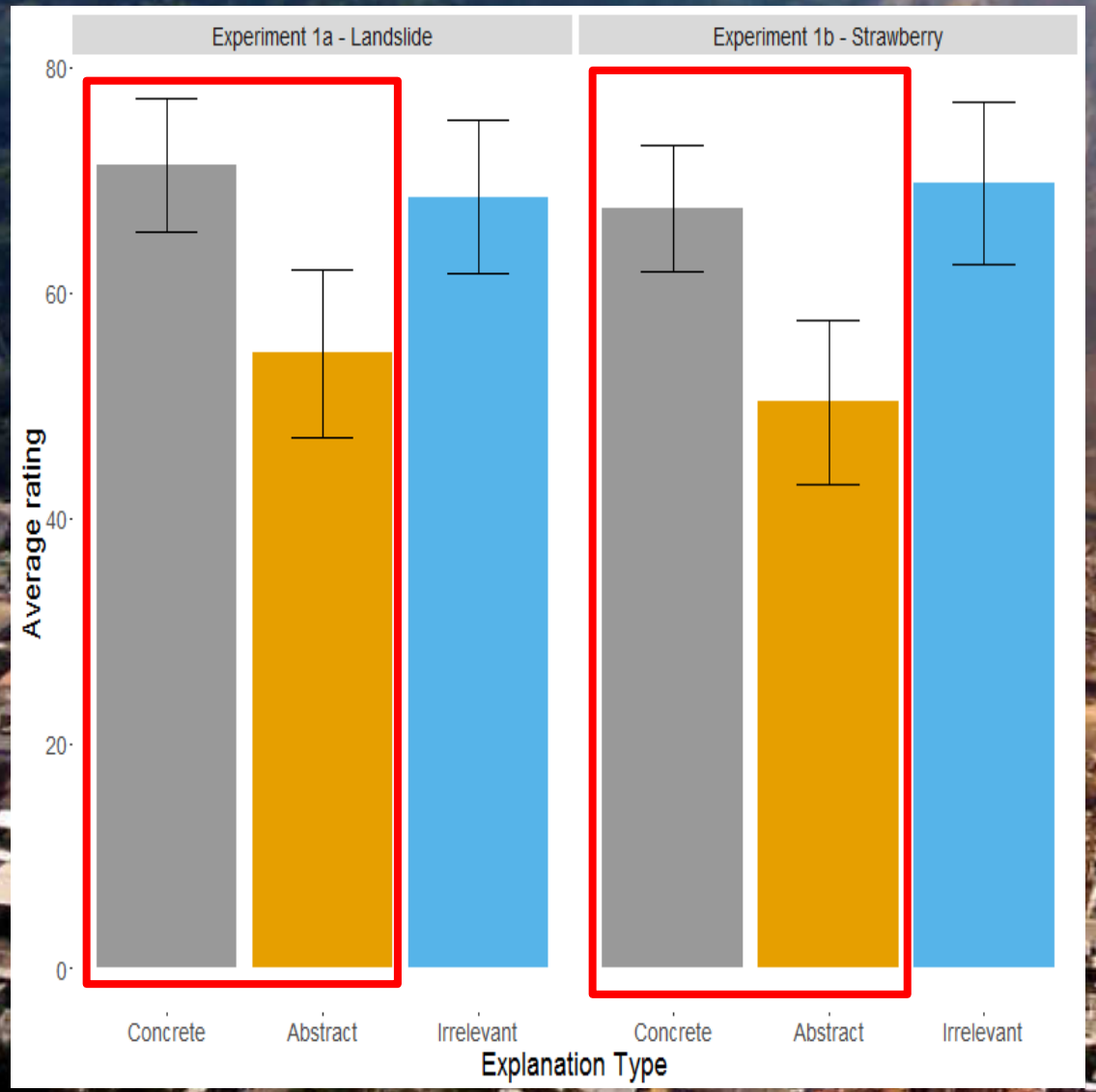
EXPLANATION RATINGS



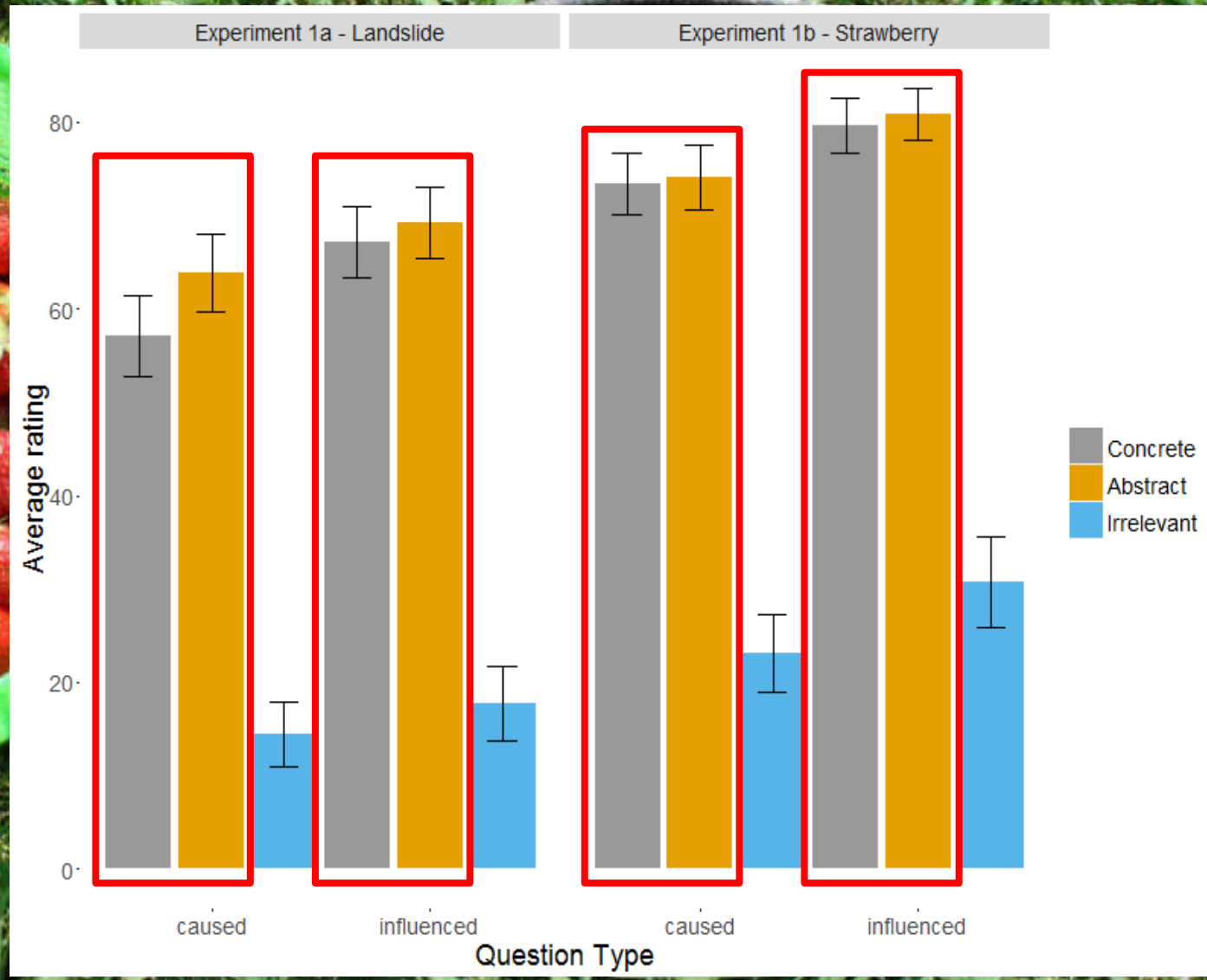
Causal Ratings



EXPLANATION RATINGS



Causal Ratings



ON THE PREFERENCE FOR CONCRETENESS

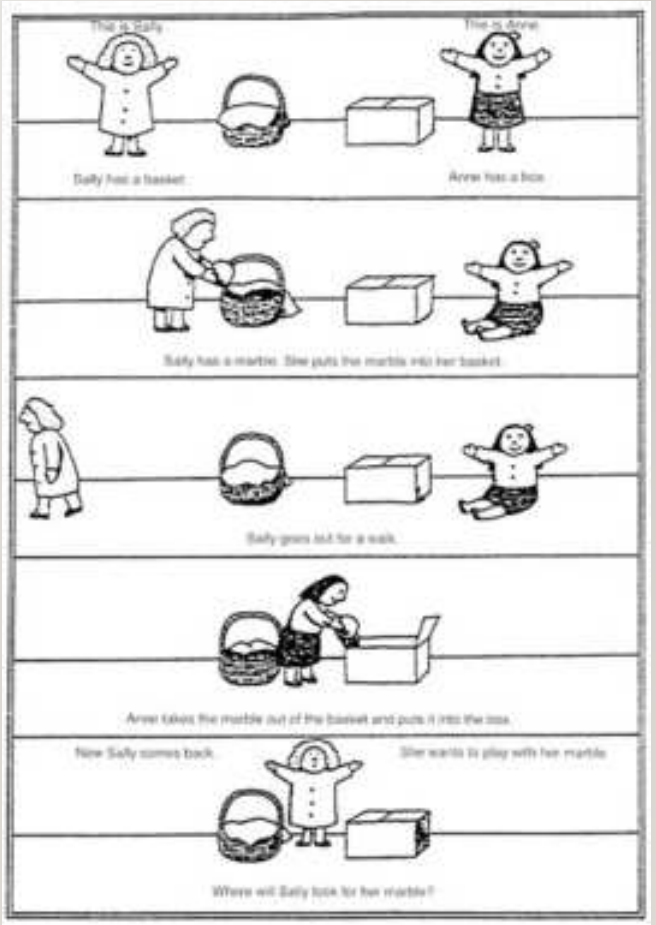
- Explanations (especially of token events) might differ depending on the aim:
  - We may try to understand why this particular event happened (e.g. to attribute responsibility) – backwards-looking
  - We may try to avoid or generate future instances – forwards-looking

VS.

  - People might default to backwards-looking explanation, in which concreteness matters
  - Vasilyeva et al (2017) show that rated explanation quality depends on the evaluator's current task

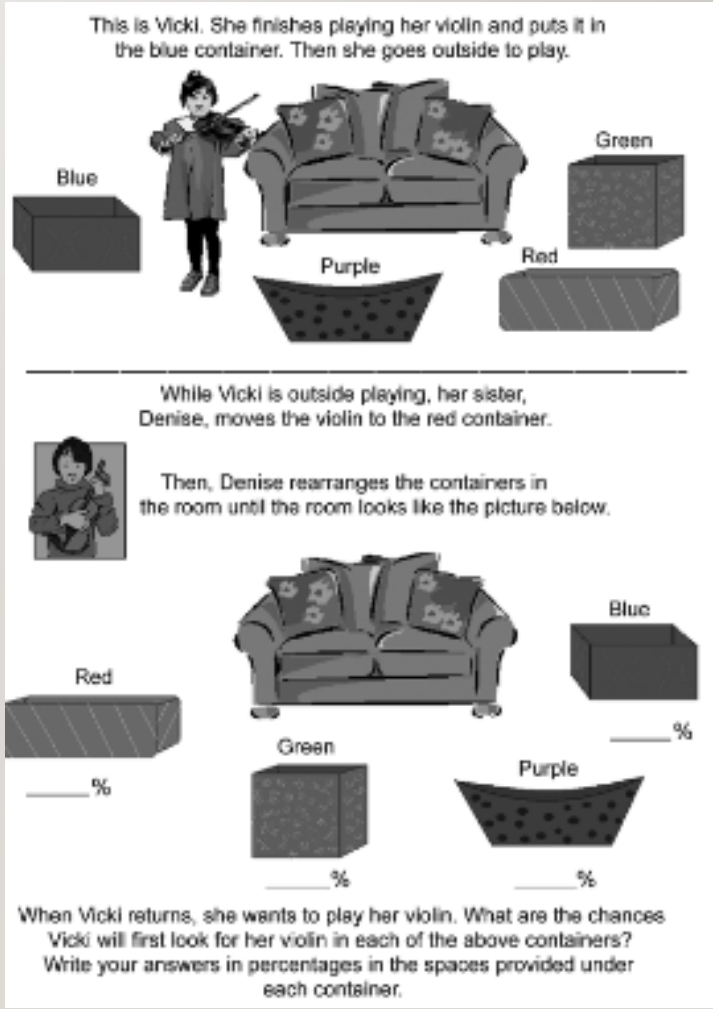
EPISTEMIC STATUS

- "The recommender system produced the given output because the latter shared types and aspects with your stated preferences"
- When producing explanations we need to carefully assess the knowledge and beliefs of the audience
- Children, up to the age of 4-5 have trouble taking into account other people's knowledge (Baron-Cohen et al 1985)
  - Theory of mind?
  - Memory?



CURSE OF KNOWLEDGE (Birch & Bloom 2003)

- Vicki puts the violin in the blue container
- Denise rearranges boxes and moves the violin to:
  - another container: original(B) = 71%, location(R)= 23%
  - the red container: original(B) = 59%, location(R)= 34%
- We try to take into account what the other person knows but by using our own knowledge as a point of departure, we may egocentrically distort it and miscalculate the informational common ground



SIMPLICITY

- Unification accounts of explanation are centred around the notion of simplicity
- But what is simplicity?
  - number of entities?
  - number of entity types?
  - shortest description?
  - most inflexible (least degrees of freedom)?

Aristotle: "We may assume the superiority *ceteris paribus* of the demonstration which derives from fewer postulates or hypotheses" (Aristotle - Posterior Analytics)

Aquinas: "If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments where one suffices"

Kant: "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances."

Newton: "Nature is pleased with simplicity, and affects not the pomp of superfluous causes"

Galileo: "Nature does not multiply things unnecessarily; that she makes use of the easiest and simplest means for producing her effects; that she does nothing in vain, and the like"

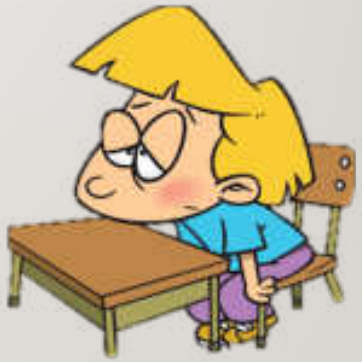
Lavoisier: "If all of chemistry can be explained in a satisfactory manner without the help of phlogiston, that is enough to render it infinitely likely that the principle does not exist, that it is a hypothetical substance, a gratuitous supposition. It is, after all, a principle of logic not to multiply entities unnecessarily"

Einstein: "The supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience" (Einstein, 1934)

Baker, Alan, "Simplicity", *The Stanford Encyclopedia of Philosophy* (Winter 2016

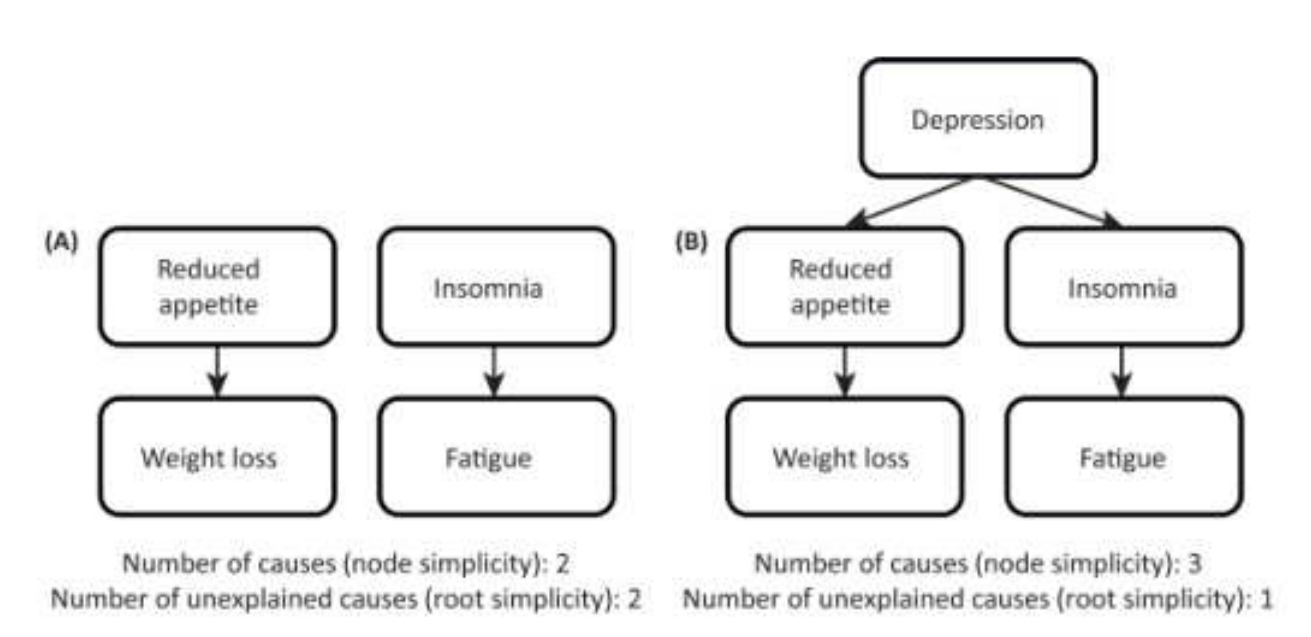
SIMPLICITY

- Paul Thagard (1989) defined simplicity as the number of special assumptions required
- He asked participants to select between explanations that accounted for observations equally well but differed in the number of assumptions
- Cheryl suffers from nausea, weight gain, and fatigue. Is it because:
  - she is pregnant, or
  - she has a stomach virus, has stopped exercising, and has mononucleosis"



- People preferred the 'simpler' explanation
  - but it could be on probabilistic grounds: the conjunction of 3 causes is less probable than a single cause
- When mentioning the probabilities (Lombrozo 2007), people favour the most likely explanation even if it's not the simplest.
- When candidate causes were equally likely, most went for the simpler explanation.
  - The complex one is preferred by the majority only when it is 10 times more likely than the simple alternative.
- Same results with 5 year old children! (Bonawitz & Lombrozo 2012)

NODE SIMPLICITY OR ROOT SIMPLICITY?



Even when explanation (B) was less probable, people preferred it

Fewer assumptions = fewer unexplained causes

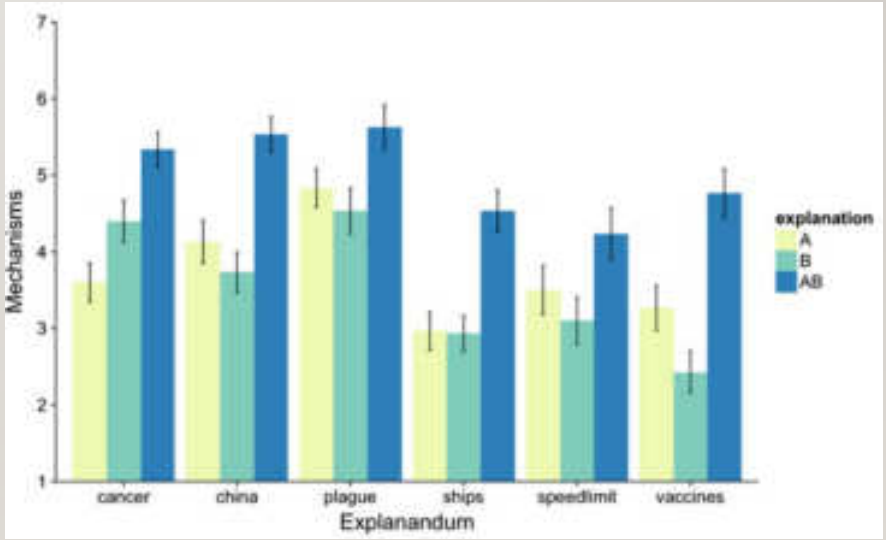
Lombrozo & Vasilyeva (2017)

SIMPLICITY?

- In the Zemla et al (2017) study I discussed earlier (Reddit) we found that judged complexity (rather than simplicity) was positively correlated with explanation quality.
- We found that explanation quality was positively correlated both with the number of root/unexplained causes and its length (level of detail)

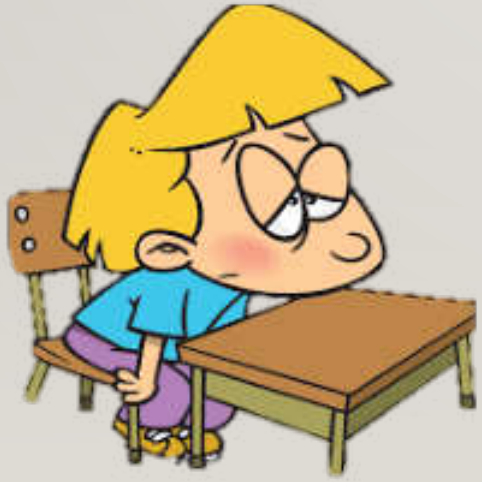
- Why isn't China's population declining if they have had a one-child policy for 35 years?
  - ethnic minorities and rural populations are exempt from the rule
  - Chinese are living longer on average and wealthy couples can afford to pay fines associated with violating the policy
  - both the above

People preferred the 3rd option in most cases.



GENERALITY – BREADTH – SCOPE - COVERAGE

- "The hallmark of a deep explanation is that it answers more than you ask" (Max Tegmark)
- This goes against the preference for hyperconcrete explanations
- Again hard to pin down:
  - Is it better to explain more things but less precisely?
  - Or fewer things but more diverse?
- "If hypothesis H1 explains two pieces of evidence and H2 explains only one, then H1 should be preferred to H2" (Thagard, 1992)



- Cheryl's suffers from nausea, weight gain, and fatigue. Is it because:
  - she has a stomach virus (explains only nausea)
  - she is pregnant (which explains all three symptoms).
- Participants favoured narrow explanations when presented with a single fact but the broad explanations when multiple facts were presented. (Read & Marcus-Newhall, 1993)
- Again this can be probability driven: 3 symptoms are stronger evidence for the broad than the narrow explanation.

LATENT SCOPE

- Why did Lois paint her nails in the shower?
  - She is obsessive-compulsive
  - She was afraid of spilling nail polish on the new bathroom rug
- The former is much broader but also less probable
- Latent scope: The number of effects that an explanation could account for but which have not been observed - Khemlani et al (2011)
  - Delimenta causes lumps and spots 72%
  - Homorula causes lumps, spots, and bumps 7%
  - Nothing else is known to cause lumps, spots, or bumps
  - Daryl has lumps; we don't know whether or not Daryl also has spots or bumps

YOU VS. YOUR USER

- What makes a good explanation for you is not necessarily what makes a good explanation for the (majority of) your users
- You may have a preference for a particular:
  - form of explanation (e.g. logic-based, DN)
  - explanatory stance (mechanical vs intentional)
  - level of abstraction (detailed over abstract)
- You may be cursed by your knowledge
- You may have different aims

EVALUATING EXPLANATIONS

- You need to ask the public to evaluate your explanations
- How?
  - Ask friends/family/colleagues who are naïve to the system and the purposes of your study
    - May still be biased
    - Responses will be noisy (individual differences) so a large sample is needed
  - Work with a HCI expert / exp. psychologist / behavioural scientist
  - Run an (online) study through:
    - Amazon Mechanical Turk – mturk.com
    - Prolific Academic - prolific.co
    - Gorilla - gorilla.sc
    - Testable - testable.org

WHAT TO ASK?

- Is this a good explanation?
- How well do you understand this explanation?
  - but beware of IOED
- Ask users to explain the system
  - but open-ended answer can be hard to analyze
- Ideally you want a behavioural measure:
  - what is the user expected to know after the explanation?
  - measure performance in a task before and after the explanation

EVALUATING EXPLANATIONS

The system made this recommendation because blah blah blah, yada yada yada

You ask 200 people: Is this a good explanation?

Average rating: 90.7%

Did you prove it is a good explanation?

COMPARE!!!

- Getting ratings for a task the users haven't seen before is not informative
- Participants may:
  - be happy/unhappy with their reward
  - think that positive ratings will ensure their payment
  - …….
- Users don't know how to evaluate a novel task
- Is this a good apple? vs Is this a good starfruit?

COMPARE!!!

- When evaluating explanations you must:
  - compare different versions of your explanations (keeping everything else the same)
    - different groups of participants see different explanations (between-groups)
    - same group of participants sees different explanations (within-group)
      - must randomize the order in which explanations are presented
  - compare the performance in a task (e.g. understanding) before and after presenting the explanation

EXPERIMENT EXAMPLE

- Get ratings for 70 movies
- Using a Quantitative Bipolar Argumentation Framework (QBAF) we generate recommendations based on inferred preferences and similar users
- Then we explain our recommendations:

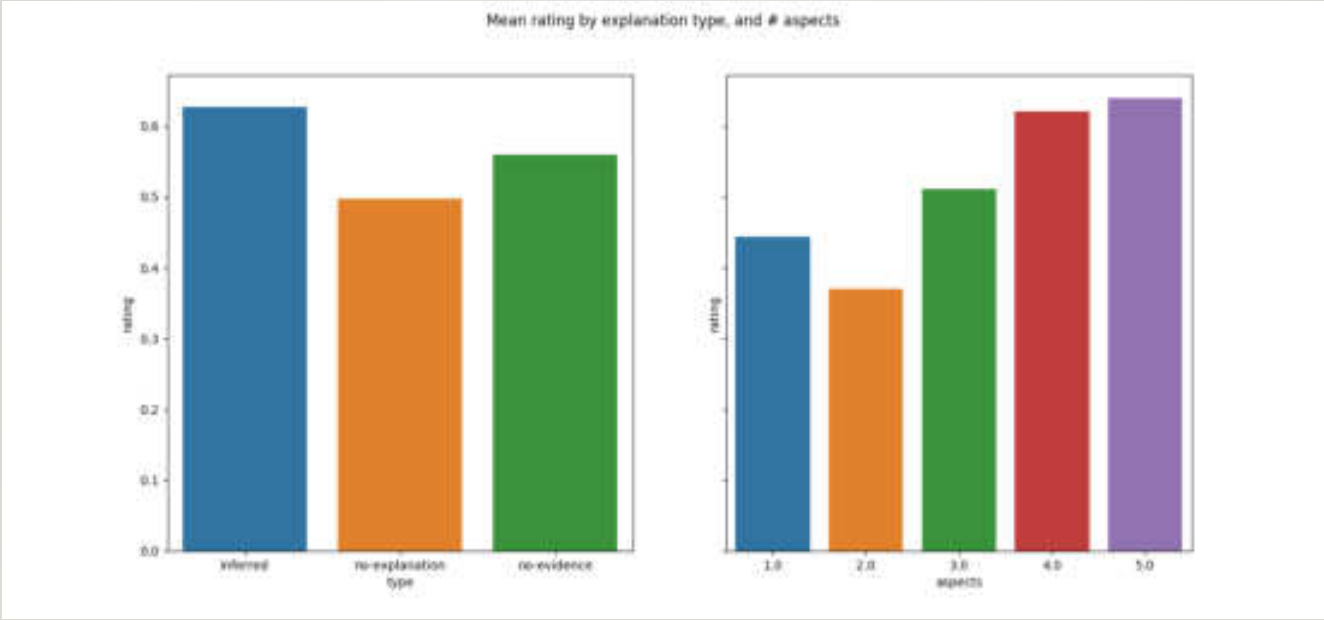
We think you should watch "The Tailor of Panama" because you liked the film "Excalibur" which is also directed by John Burman and you liked films with Pierce Brosnan such as "Mama Mia" and "Golden Eye" and also you like spy films such as "Tinker, Tailor, Soldier, Spy" and "Mission Impossible"

- Please rate this explanation

EXPERIMENT EXAMPLE

We think you should watch "The Tailor of Panama" because you liked film "Excalibur" which is also directed by John Burman and you liked films with Pierce Brosnan such as "Mama Mia" and "Golden Eye" and also you like spy films such as "Tinker, Tailor, Soldier, Spy" and "Mission Impossible"

- We were comparing (1) different phrasings (2) providing evidence or not and (3) number of type/aspect combinations



EXPERIMENT EXAMPLE

We think you should watch "The Tailor of Panama" because you liked film "Excalibur" which is also directed by John Burman and you liked films with Pierce Brosnan such as "Mama Mia" and "Golden Eye" and also you like spy films such as "Tinker, Tailor, Soldier, Spy" and "Mission Impossible"

But is this an explanation?

- Can treat it as an Inductive Statistical explanation through "Hidden Structure":

  If one likes a movie of genre G they might also like a movie of the same genre

  You like movie X of genre G

  Movie Y is of genre G

  ……..

  You might like movie Y

EXPERIMENT EXAMPLE

We think you should watch "The Tailor of Panama" because you liked film "Excalibur" which is also directed by John Burman and you liked films with Pierce Brosnan such as "Mama Mia" and "Golden Eye" and also you like spy films such as "Tinker, Tailor, Soldier, Spy" and "Mission Impossible"

But is this an explanation?

But do participants treat it as an explanation?

- Why the recommender chose "The Tailor of Paname" (which is a fact)
  - the statements are a way of clarifying how the recommender reached that conclusion, how it works.
- Why you should like "The Tailor of Panama" (which is not a fact)
  - the statements become premises in support of that, an attempt to persuade users (to agree/buy etc).
  - In this case, participants would probably not judge the quality of the explanation in terms, for example, of its simplicity, coherence or coverage but rather report the degree to which they are convinced about the suitability of the recommendation.

EXPERIMENT EXAMPLE

We think you should watch "The Tailor of Panama" because you liked film "Excalibur" which is also directed by John Burman and you liked films with Pierce Brosnan such as "Mama Mia" and "Golden Eye" and also you like spy films such as "Tinker, Tailor, Soldier, Spy" and "Mission Impossible"

But is this an explanation?

But do participants treat it as an explanation?

- Repeated the experiment but one group was first told:

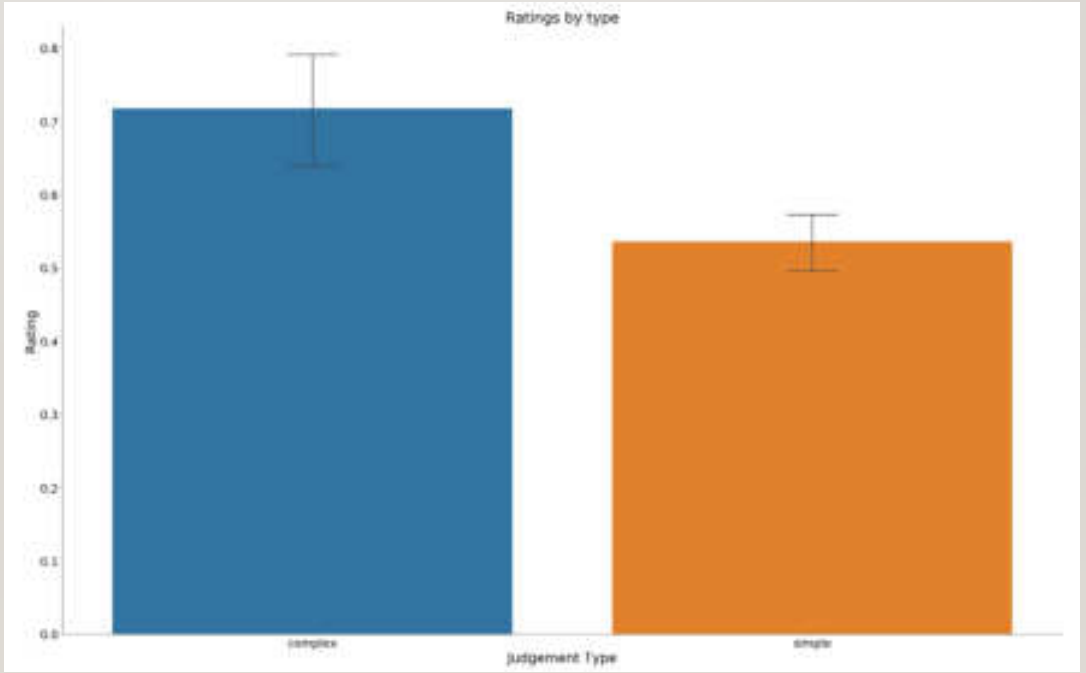
  In the past, there has been the suspicion that our recommendations are not genuinely based on users' preferences but result from promotion agreements between the recommendation system and production companies.
  Some users believe, in other words, that the films recommended to them are just commercials dressed up as personalized suggestions

EXPERIMENT EXAMPLE

We think you should watch "The Tailor of Panama" because you liked film "Excalibur" which is also directed by John Burman and you liked films with Pierce Brosnan such as "Mama Mia" and "Golden Eye" and also you like spy films such as "Tinker, Tailor, Soldier, Spy" and "Mission Impossible"

But do participants treat it as an explanation?

In the past, there has been the suspicion that our recommendations are not genuinely based on users' preferences but result from promotion agreements between the recommendation system and production companies.
Some users believe, in other words, that the films recommended to them are just commercials dressed up as personalized suggestions



SUMMARY

- Normatively, explanations are deductive or inductive arguments that unify different aspects of knowledge but almost always point to a causal connection
- People prefer explanations that contain no cycles, are simple (but not always), coherent (but that's difficult), perhaps providing additional background information
- People also might be impressed by length, jargon and teleological explanations
- But their judgements will depends on the explanatory stance and their aims – perhaps they don't even treat your explanation as such!
- When you generate explanations, it is a good idea to have explanatory virtues and vices in mind.
  - Even if you don't include all virtues and don't exclude all vices, you should at least keep an eye for them
- In experiments you should aim for behavioural measures as close as possible to the real task and you should ALWAYS use comparative measures

Thanks!