Lecture 5: Case study

[Disclaimer: These informal lecture notes are not intended to be comprehensive - there are some additional ideas in the lectures and lecture slides, textbook, tutorial materials etc. As always, the lectures themselves are the best guide for what is and is not examinable content. However, I hope they are useful in picking out the core content in each lecture.]

Introduction

The lecture series up to this point has followed a pretty typical structure. After the introductory lecture, each of the three main "topic" lectures picked out a single aspect of cognition and discussed some of the key ideas relating to each of them:

- Attention
- Similarity
- Reasoning

Throughout, I've tended to avoid going into detail about the specific methodology used in each experiment, preferring to focus on core ideas and theories.

Obviously, these are only a few topics in cognitive psychology. Every textbook will cover many more topics, and later on in this class other lecturers will extend the topic list beyond these.

In any given one of these lectures, I've tried to highlight how it relates to many different tasks. For example, in the reasoning lecture I talked about the Wason selection task, property induction (inductive arguments), etc.

In today's lecture, I want to "reverse" this structure, picking a single research paper and talking about how it ties together different thoughts about attention, similarity and reasoning. In the process, we'll bring in some ideas about social cognition, learning and so on. In a somewhat self-serving move, I've selected one of my own papers that came out last year:

- Ransom, K., Perfors, A. and Navarro, D.J. (2016). Leaping to conclusions: Why premise relevance affects argument strength *Cognitive Science, 40*, 1775-1796. paper

You aren't expected to read the original paper for this class, but if you're interested, I've included the link.

Background

In our discussion of inductive reasoning we discussed the idea of **premise monotonicity**, which roughly corresponds to the very intuitive idea that "adding more facts should strengthen an argument"! As an example, this argument

> Dolphin cells contain TH4 hormone
> Therefore cow cells contain TH4 hormone

feels much weaker than this one:

> Dolphin cells contain TH4 hormone
> Mouse cells contain TH4 hormone
> Bat cells contain TH4 hormone
> Therefore cow cells contain TH4 hormone?

And as usual, we should note that this phenomenon isn't restricted to "arguments that are explicitly written as syllogisms", you can see something pretty similar in everyday arguments and conversations too. As a general rule, more evidence is better.

Except... not always. Consider this pair of arguments. As before, our "baseline" argument is this one:

> Dolphin cells contain TH4 hormone
> Therefore cow cells contain TH4 hormone

But this time around, the two "new" facts that we will introduce happen to pertain to other marine mammals:

> Dolphin cells contain TH4 hormone
> Whale cells contain TH4 hormone
> Seal cells contain TH4 hormone
> Therefore cow cells contain TH4 hormone?

Intuitively, this argument seems to feel much *weaker* than the original one. This phenomenon is referred to as **premise non-monotonicity** (adding more examples weakens the strength of the argument), and it tends to appear in situations where all of the premise items are all similar to another in a very particular way.

A theoretical perspective

Premise non-monotonicity is a rather curious thing. It appears to be a phenomenon in which all three of our lectures seem relevant: when a particular pattern of **similarity** appears (*dolphin*, *whale*, *seal*), it seems to call our **attention** to a particular category (*marine mammals*), and this in turn drives the way we **reason** about the situation. Here's what I mean. From the lecture slides, we have the following observations:

(1) Similarity shapes reasoning:

- Dolphins and cows are dissimilar. So it feels unreasonable to generalise from one to the other. We'll denote this argument as `dolphins -> cows`

- Bats and mice are dissimilar to cows and to dolphins. So when we add those to the argument we obtain `(dolphins + bats + mice) -> cows` and it feels like a much stronger argument. The fact that the property (TH4 hormone) is common to such a diverse set of animals makes it seem like it must be pretty widespread across mammals, so the generalisation to cows is sensible

- Seals and whales are *dissimilar* to cows, but they are *similar* to dolphins. So if we added those items to the argument we obtain `(dolphins + seals + whales) -> cows` and that seems to suggest that TH4 is restricted to marine mammals, thereby undermining the argument that the property should be generalised to cows.

- On the other hand, if we switched to conclusion item from cows to a marine mammal like dugong, we get a very strong argument indeed `(dolphins + seals + whales -> dugongs`. In this case, the very tight similarity among all four items seems to yield the strongest arguent of the lot.

(2) Similarity directs attention:

- In order to understand what's going on consider the features that *dolphins* have. A dolphin is "a marine mammal", an "intelligent animal", a "mammal", a "cute entity", etc. Each of these features implies a *category* that I might want to generalise TH4 to. Perhaps TH4 is a hormone associated with temperature regulation, which suggests mammals; or maybe it generates big googly-eyed baby faces that humans find "cute" (or whatever). The scientific details of what TH4 might correspond to don't really matter so much as what kind of category I can *imagine* connecting the TH4 hormone to.

- Adding extra premises to the argument helps direct my attention (internally!) to one category or another. Bats are not cute, mice are not smart, and neither one lives in the ocean, so the most plausible category for me to imagine when given those three examples is *mammal*. In contrast, when I select many marine mammals, my attention is driven to the narrowest everyday category that seems to cover all of them: *marine mammals*. The strength of the inductive argument seems to be tied to *which* category our attention is drawn to. If we are called to think about a category that includes the conclusion item (e.g., *dugongs* are marine mammals), it feels like a strong argument; if the attended category does not include the conclusion item (e.g., *cows* are not marine mammals) the argument feels weak.

- Notice that seals and whales are also intelligent cute mammals. These possibilities aren’t ruled out by the additon of extra evidence, it just encourages us to ignore or disregard those possibilities.

Which leads to the question...

Why does this happen?

There are several possibilities for why this happens, but the one that we focused on in the Ransom et al (2016) paper is the idea that in everyday life *arguments are made by other people*, and other people have a vested interest in communicating their ideas to you. Unlike nature (which is dumb and unhelpful), humans are intelligent agents with complex goals and a rich language. We “transmit” information to each other via a complicated mechanism... we aim to *persuade* each other of the truth or falsity of different ideas. This has an effect on the evidentiary value of a "human uttered statement", when compared to the evidentiary value of a "fact sampled randomly from the world.

Here's how it works...

1. Suppose I want to persuade you to believe that dugongs produce TH4 hormone.
2. I might think well, if I choose these similar animals (dolphins & whales), I can assume that you are not stupid, and you will notice this similarity.
3. I can also assume that you will notice that the common feature between them is "marine mammals".
4. On your end of the "communication channel", you *know* that I know this. You can assume that I *wanted* you to think of marine mammals, so you can assume that this similarity is *relevant*, and so...
5. You can take a hint.

... all of which highlights the importance of **theory of mind**. We We have intuitive theories about the workings of each other’s minds, so we can select relevant information that drives attention to the right answer.

It relies on an assumption that we are being *helpful* to one another. That is, it is perfectly reasonable human behaviour to tell as story like this....

> “I’ve studied TH4 hormone for many years… and I have discovered it in the cells of whales, seals and dolphins. I want you to believe that dugongs will produce TH4 hormone"

It is also perfectly reasonable human behaviour to tell a story like this...

> "I’ve studied TH4 hormone for many years… and I have discovered it in the cells of mice, bats and dolphins. I want you to believe that cows will produce TH4 hormone"

However, it violates the assumption of helpfulness if I were to try and tell a story about TH4 hormone that went like this...

> "I’ve studied TH4 hormone for many years… and I have discovered it in the cells of whales, seals and dolphins. I want you to believe that kittens will produce TH4 hormone"

If I were to do this in real life and expected you to take me seriously, you would (quite rightly) conclude that I am a jerk. This idea has a lot of analogs. Here are a few examples:

- Chekhov's gun: when telling a story, every element must be necessary. *"One must never place a loaded rifle on the stage if it isn't going to go off. It's wrong to make promises you don't mean to keep."*
- When giving sworn testimony we require people to *tell the truth, the whole truth, and nothing but the truth*
- Grice's maxims tell us that good communication is informative, truthful, pertinent and unambigious.
- Etc

Fundamentally though, they all boil down to a version of Wheaton's law, which (politely put) advocates that humans should try to be nice to each other.

So ultimately, the claim is that a lot of what's going on when we "let" the salient similarities that appear in someone's argument drive our attention to a particular category, and then go on to endorse that category is... we assume someone is trying to help us learn about the world.

An experimental test

Reseach goals

To generate an experimental test, we considered two different ways of constructing an argument, and we wanted to see what impact these might have on the premise monotonicity/non-monotonicity effect. Our two methods were:

- **Random sampling**, in which the world generates true facts facts largely at random (e.g., dophins are cute, unicorns are awesome) and with no particular goal in mind
- **Relevant sampling**, in which another (helpful) human selects true facts purposefully, with a deliberate intent to persuade you of the truth of some proposition

Key procedural detail, and the dependent variable

We did this within the context of a reasoning task that used the following procedure:

- First, rate the strength of an inductive argument (e.g., `dolphin->cow`) that has a single premise, using a slider bar to indicate how likely it is that the conclusion item possesses the property (e.g., TH4 hormone).
- Next, tell people about a second item to produce a new inductive argument (e.g., `(dolphins + seals) -> cow`) and use the slider bar to indicate how strong this argument is.
- The *dependent variable* is the *difference* between these two responses: a positive difference means the second argument is stronger (i.e., premise monotonicity), and a negative difference means that the second argument is weaker (i.e., premise non-monotonicity)

Our hypotheses

- When a helpful human makes an argument, the similarity between premise items will be deemed relevant, and the premise non-monotonicity effect will appear

- When an indifferent world generates random data, the similarity between premise items will be deemed irrelevant, and people will revert to premise monotonicity

More procedural information, and independent variables

One potential issue that we have in designing this study is the fact that people know it is a psychological experiment, and people know that psychological experiments are designed by psychologists... who aren't always trustworthy people! So it's not going to be easy to convince people that they're seeing "truly random" information.

To get around that, we designed a pretty "over the top" experiment to really drive home the idea. We did this by manipulating two *independent variables*.

1. The **cover story manipulation**. When introducing the reasoning problem, we tried the obvious trick of just *telling* people about how the argument was (supposedly) generated. In the **relevant cover story** condition we told people that the facts were selected by a previous participant who wanted to help them make good guesses; in the **neutral cover story** we didn't say anything in particular about the origins of the argument;and in the **random cover story** condition we told people that the premise items were chosen at random from a data base of facts.

2. The **experience** manipulation. To help ensure that people really did believe that the items were generated in a helpful way (relevant) or a random way, we also included a series of "filler" trials that looked exactly like the ones we actually cared about, but were designed to seem useful or useless. In the **relevant filler** condition, the arguments used in the filler items tended to be very convincing and consisted of highly relevant evidence (e.g., `eagles + hawks -> doves` seems to include useful evidence); whereas in the **random filler** condition the filler trials included some very unhelpful evidence (e.g., `eagles + NOT tortoises -> doves`)

The order of arguments was designed so that people tended to see the filler items before seeing the target items (see slide 61-62)

An incomplete 2x3 design

Taken together, these two manipulations yield four possible conditions:

- Relevant story, Relevant fillers (1)
- Neutral story, Relevant fillers (2)
- Neutral story, Random fillers (3)
- Random story, Random fillers (4)

Notice that there are two "logically possible" conditions that we did not include:

- Relevant story, random fillers
- Random story, relevant fillers

In most situations, it's considered good scientific practice to include all possible combinations of your independent variables, yet for this study we didn't. Why?

Hopefully the answer is obvious: those two conditions introduce a very nasty confound... namely that the cover story and the filler items contradict each other. We would end up lying to participants if we included these conditions, and participants might rightly conclude that we were being jerks.

We used two different "target" arguments that would be expected to produce premise non-monotonicity under normal circumstances...

Target argument #1

```
Grizzly bears produce TH-L2
--
Therefore, lions produce TH-L2
```

versus

```
Grizzly bears produce TH-L2
Black bears produce TH-L2

--
Therefore, lions produce TH-L2
```

- If this similarity is deemed relevant… it strongly suggests TH-L2 is a property of bears so the extra evidence weakens the conclusion... non monotonicity

- If this similarity is deemed irrelevant because the data are random … the extra "evidence" is almost useless. So we expect no difference or a weak monotonicity effect because at least there's some extra evidence that TH-L2 is not rare

Target argument #2

The second target argument had roughly the same structure and the same logic to it. The single-premise version of the argument asked people to generalise from `tigers -> ferrets` and the two-premise version asked people to generalise from `(tigers + lions) -> ferrets`

Control argument

Finally, in order to check that the effect doesn't happen all the time, we included a control condition where it really shouldn't make much of a difference, because the second premise should always strengthen the argument. We did this with an argument where people first rated `orangutan -> gorilla` and then rated `(orangutan + chimpanzee -> gorilla`

Predictions?

- So we have four conditions, and our hypotheses lead to a clear prediction about how the dependent variable should change across the conditions: for the target arguments, the difference score should be largest in condition 1, followed by conditions 2 and 3, and then by condition 4. For the control argument, there should be very little effect.

- Better yet, the alternative theory (i.e., that the method of argument construction doesn't matter and that similarity always has the same effect no matter what) makes a clear prediction too. The DV should be the same for all conditions, but it should be negative for the two target arguments (non-monotonicity), and positive for the control argument (monotonicity)
- Even better, the other possibility (that we screwed up our design and people responded randomly) also yields a different prediction, namely that the data should be random and there will be no systematic or meaningful changes across conditions

Slides 70-78 provide a graphical illustration.

A digression

What do you think happened? How confident are you that I'm about to tell you that our predictions came true? You're probably very confident, right? Because... if our predictions didn't come true, I wouldn't be writing these lecture notes about a cool study... and you wouldn't have read all these preliminaries. The fact that I have set up the story like this - assuming that I am a helpful human trying to persuade you of something - *strongly* hints at what the answer is going to be, doesn't it? This whole set up is very clearly a Chekhov's gun and it's about to go off in the results section that will inevitable reveal that I was right in my predictions.

In contrast, how confident do you think I was when I ran the study originally? Nature doesn't supply us with Chekhov's guns, and my predictions don't always come true (unfortunately). So when we ran the study, I had no such guarantee that the results would tell a nice story... and I was not at all confident that it would work.

Oh, hey... that's actually our hypothesis right? Human reasoning when a knowledgeable and helpful informant (teacher!) presents you with an argument is qualitatively different from when an uncaring and indifferent universe generates observations (science!).

... the inevitable pattern of results

As shown on slides 80-83, we got *exactly* the pattern of results we predicted. Obviously we did, otherwise I wouldn't have written this lecture!

Sadly, not all my experiments work out that nicely.

... another digression ...

- One thing that the paper talks about (but I've hidden from this lecture) is that we went a fair bit further than just making "qualitative" predictions about what would happen in the experiment. We also built a **computational model** that tries to use the same set of facts to reason in a human like way.

Specifically:

- In some conditions the model is equipped with a simple "theory of mind" (i.e., it assumes it is being given helpful evidence) and in other conditions it assumes it is being fed random evidence... and we show that it produces human-like reasoning
- The model is based on (Bayesian reasoning)[https://en.wikipedia.org/wiki/Bayesian_inference], a tool that is used quite widely in statistics, artificial intelligence and machine learning. However, it's beyond the scope of this class to talk about Bayesian reasoning - I just wanted to mention it for those students who are also taking AI and computer science classes, because there are some really neat connections between psychology and computer science in this particular domain.

Limitations?

- Sample size too small?
  - Probably not a problem. We collected data from 538 participants

- Sample not representative?
  - Maybe? Our participants were recruited online: diverse in age and gender, but narrow in nationality (USA) and probably above average in education
  - Important question: is there a plausible reason to think this might matter?

- Stimulus order not randomised?
  - Probably not. The non random ordering (i.e., fillers mostly first) was intentional, and was central to the experimental manipulation

- Factors not fully crossed?
  - Absolutely not. It would have been absurd to include a “relevant story + random experience” condition… this would introduce a confound colloquially referred to as “lying to participants”

- Limited range of arguments?
  - Definitely a limitation. We used a fixed set of six arguments, all of which were about animals.
  - Important question: why might this matter? (hint… people have different knowledge)

- Limited range of phenomena?
  - Definitely a limitation. We only looked at the premise (non) monotonicity effect.
  - There are good reasons to think the same manipulations should influence other inductive phenomena (and in more recent work we've found some evidence for that!) but this study did not consider them