PHIL246: Foundations of Cognitive Science

Instructors Toby Mintz & Alexis Wellwood

TAs Megha Devraj, Rachel Keith & Nadja Winning

5 The conscious, social, moral, essentialist

1 / 200

Hi! How going! Exam; reminder that we'll curve.

Now: Module 5 is A "sampler", touching on issues of what we (think we) know about ourselves, and others

The big question I want to open with is:

What is it to be a human?

So far in this course, we've spent the bulk of our time trying to explicate and understand aspects of the computational theory of mind, and applying it to specific cases of description and explanation in perception, conception, language, and so on.

In this module we begin talking about some of the phenomena of human experience that are more readily accessible to us, with the goal of examining those areas with the scientific viewpoint we've stressed throughout the semester.

So where we have focused on specific ASPECTS of our minds, we now broaden our scope a bit, to thinking about what it means to be human-

Now, we won't ask a question this big, of course; instead, we'll replace it with—

2 / 200

What is it to be a conscious, social, moral, essentialist?

... each of these nouns indicates important aspects of human experience that cognitive scientists have begun to consider in detail, such as:

- We know we are conscious; can we *explain* conciousness?
- How and to what extent do we "see ourselves" in others?
- How do we know what to do and what not to do?
- And, how do our experiences with other people affect our expectations about them in the future, and how does that differ depending on what KIND of person we take them to be?

I want to start by giving you guys a quick whirlwind introduction to each of these 4 topics, which we'll cover for the next 4 days.

3 / 200

conscious
social
moral
essentialist

The problem of consciousness is, as we'll conceive of it, the problem of saying what it is like to be the kinds of creatures we are.

The importance of this problem for cognitive science is aptly put by Thomas Nagel, a philosopher emeritus at NYU:

Without consciousness the mind-body problem would be much less interesting. With consciousness it seems hopeless. ... If we acknowledge that a physical theory of mind must account for the subjective character of experience, we must admit that no presently available conception gives us a clue how this could be done.*

... explaining our "experience" of ourselves, the "what it's like" to be us, in this sense, is widely considered the hardest problem in cognitive science - and in fact, many people think that it may very well be beyond the reach of scientific explanation.

* Thomas Nagel (1974), What is it like to be a bat? *The Philosophical Review*.

conscious
social
moral
essentialist

Now, part of our experience in life tells us that we have a mind, of course; we do have that knowledge of what it's like to be us. Of course, as some philosopher famously put it, "I think, therefore I am"

What's fascinating is that we FURTHERMORE suppose that others are in the same predicament; that they have their own experiences and motives.

6 / 200

As Susan Johnson, psychologist, highlights the issue:

One of the most interesting senses of the concept 'agent'... is the one entailing the possession of mental states. Mental states, and the minds that possess them, are necessarily unobservable constructs.... [Yet] The ability to construe ourselves and others as agents with minds... is critical. With this ability we can communicate referentially, predict and explain others' mental states for the purposes of complex problem-solving and learning, not to mention deception.*

... As we'll see, typically-developing children behave, pretty early on, as if they expect others to have minds and experiences of their own, just as they themselves do.

Signs point to this knowledge being common across the species, despite, as usual, wide variations in experience; how did that come to be so?

* Susan Johnson (2000), "The recognition of mentalistic agents in infancy", *Trends in Cognitive Science, 4(1)*, p22

And further: finding ourselves as one among many agents in society, the question of how we ought to treat one another becomes critical.

Interestingly, how we think we should treat others and how we expect to be treated shows common threads also across cultures that otherwise might diverge wildly...

As Kiley Hamlin, developmental psychologist, puts the question,

although there are considerable cross-cultural differences regarding which actions are considered obligatory, permissible, or forbidden, and a wide variety of prescribed responses to violations across societies, all normally developing adults in all cultures have the basic notion that some things are right and others are wrong .... Where does ~~this moral sense c~~ome from?

...given the commonalities both within and across cultures, it looks an awful lot like "nature" plays a significant shaping role in determining the organization of our "moral faculty".

* Kiley Hamlin (2013), "Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core". *Current Directions in Psychological Science*.

conscious
social
moral
essentialist

Yet of course, despite evident similarities in what we'll grant to other people - that they have experiences, expectations, and goals just like ours - we are at the same time predisposed to treat different groups of people in ways that we would never want to be treated, nor in ways that we would treat members of our own group.

Sarah Jane Leslie, reflecting on such facts, says:

The perception that members of a given group share highly predictive natures supports a willingness to generalize across the group, and the perception that these natures are distinctive produces reluctance to generalize to more inclusive categories —to the category human being, for example. [...] If ... an individual [who commits an appalling act] belongs to a highly essentialized group ... we may view the inclination to appalling action as part of the essence of that group.

... Her ocmments here reflect findings that we humans have a general tendency to *essentialize* things in our environment, including people, and believing that one group possesses a different *essence* from your own can lead you to draw, in particular, strongly negative inferences about a group on the basis of very little evidence that the group has the relevant properties.

Sarah Jane Leslie (2017), "The original sin of cognition: Fear, prejudice, and generalization". *The Journal of Philosophy*.

That's the whirlwind overview, here's what's upcoming:

Upcoming

Module 5 required readings

1. **Nagel (1974), "What is it like to be a bat?"** - Intro to the problems of explaining or reducing away conscious experience [today]
2. **Johnson (2000), "The recognition of mentalistic agents in infancy"** - The rudiments of representing and reasoning about other minds are present early in human ontogeny [Tues Nov 1]
3. **Hamlin (2013), "Moral judgment and action in preverbal infants and toddlers"** - criteria for a "moral sense" and evidence that prelinguistic infants meet those criteria [Thurs Nov 3]
4. **Leslie (2017), "The original sin of cognition"** - A way of understanding cognitive biases towards "essentialism" in reasoning and generalization about social kinds [Tues Nov 8]

Assignments

1. **SA2** late submissions available through Nov 1
2. **Materials** for lab group project due Nov 4
3. **SA4** will be posted Friday, due Nov 11 (will make 50 points of extra credit available)

In this module we will engage directly with some texts, these are listed here as on the Blackboard site.

Today we're discussing Nagel (1974), ...

... Questions?

Agenda: Module 5

1. Consciousness
   - Return of the mind-body problem
   - The hard problem
   - The "easy" problems
2. Sociality
3. Morality
4. Essentialism

Today We'll look at ideas from the Nagel paper, as well as some discussion from David Chalmers, a philosopher of mind and cognitive science at NYU, widely considered one of the foremost contemporary thinkers on consciousness.

First, we'll bring the mind-body problem back to the fore, and then consider how to underestand "the hard problem" of consciousness, and then some putatively related problems that are not so hard.

The "hard problem" of consciousness is explaining the "what it's like" to be an organism of a particular sort. This is DIFFERENT from various problems that we loosely associate with consciousness, all of which can or should find natural explanation in terms of the implementation of FUNCTIONS as we've talked about so far in this class.

Nagel's essay, "What is it like to be a bat?" stresses the difficulty of even **formulating** the hard problem, and the challenges that an inability to formulate it properly pose for cognitive science.

Let's have a warm-up reminder of the mind-body problem, from Jonathan Westphal, a philosopher of mind, metaphysics, and science:

Consider the human body, with everything in it, including internal and external organs and parts — the stomach, nerves and brain, arms, legs, eyes, and all the rest. Even with all this equipment, especially the sensory organs, it is surprising that we can consciously perceive things in the world that are far away from us. For example, I can open my eyes in the morning and see a cup of coffee waiting for me on the bedside table. There it is, a foot away, and I am not touching it, yet somehow it is making itself manifest to me. How does it happen that I see it? How does the visual system convey to my awareness or mind the image of the cup of coffee?

Source: https://thereader.mitpress.mit.edu/discovery-mind-body-problem/

The answer is not particularly simple. Very roughly, the physical story is that light enters my eyes from the cup of coffee, and this light impinges on the two retinas at the backs of the eyes. Then, as we have learned from physiological science, the two retinas send electrical signals past the optic chiasm down the optic nerve. These signals are conveyed to the so-called visual cortex at the back of the brain. And then there is a sort of a miracle. The visual cortex becomes active, and I see the coffee cup. I am *conscious* of the cup, we might even say, though it is not clear what this means and how it differs from saying that I see the cup.

.... okay so so far we have a relatively "easy" problem - it's not simple, but it's relatively easy -- how do we see things? The "hard" part is understanding what it means to be *conscious* of the cup, what it means to be you experiencing the cup:

Source: https://thereader.mitpress.mit.edu/discovery-mind-body-problem/

15 / 200

One minute there are just neurons firing away, and no image of the cup of coffee. The next, there it is; I see the cup of coffee, a foot away. How did my neurons contact me or my mind or consciousness, and stamp there the image of the cup of coffee for me?

It's a mystery. That mystery is the mind-body problem.

Source: https://thereader.mitpress.mit.edu/discovery-mind-body-problem/

Nagel's challenge for cognitive science

Many thinkers have attempted to reduce the mental to the physical, usually attended by claims like that our experience of ourselves is "epiphenomenal" to activity in the brain.

... However, even if we've given a complete (and *computational*) description of the brain along these lines, we need to understand that still, something will have been left out.

This will be true of brains, but not obviously so for other computing devices – once we have given a complete computational description of a Macbook, for example, we are likely to think that we are done saying WHAT IT IS; we do not normally furthermore ascribe CONSCIOUSNESS to the machine.

Nagel's challenge for cognitive science

Many thinkers have attempted to reduce the mental to the physical, usually attended by claims like that our experience of ourselves is "epiphenomenal" to activity in the brain.

Nagel:

- "I do not deny that conscious mental states and events cause behavior, nor that they may be given functional characterizations. I deny only that this kind of thing exhausts their analysis" (437)
- Why? "It is impossible to exclude the phenomenological features of experience from [such] a reduction in the same way that one [apparently legitimately] excludes the phenomenal features of an ordinary substance from a physical or chemical reduction of it—namely, by explaining them as effects on the minds of human observers. If physicalism [about the mind] is to be defended, the phenomenological features must themselves be given a physical account." (437)

The problem of consciousness — what Nagel characterizes as "the subjective character of experience" — is "what makes the mind-body problem unique" for humans and other creatures, and which typically "is ignored" in such attempts at reduction (435)

He writes,

...

This, for Nagel, is the challenge he raises for cognitive science: whether, and how, it is possible to give an objective or physical characterization of consciousness **at all**.

Let's unpack this.

"The subjective character of experience"

What is that?

- "no matter how the [physical] form may vary, the fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism" (436)
- A complete characterization of some organism's mind, then, must be able to characterize such facts

"The subjective character of experience"

What is that?

- "no matter how the [physical] form may vary, the fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism" (436)
- A complete characterization of some organism's mind, then, must be able to characterize such facts

Can we do that?

- "when we examine the[] subjective character [of experience] it seems that such a result is impossible. The reason is that every subjective phenomenon is essentially connected with a single point of view, and it seems inevitable that an objective, physical theory will [have to] abandon that point of view." (437)
- Nagel's extended case study of "what it's like" to be a bat illustrates the challenge:

What is it like to be a bat?

"Facts about what it is like to be an *X* are very peculiar, so peculiar that some may be inclined to doubt their reality, or the significance of claims about them." (437)

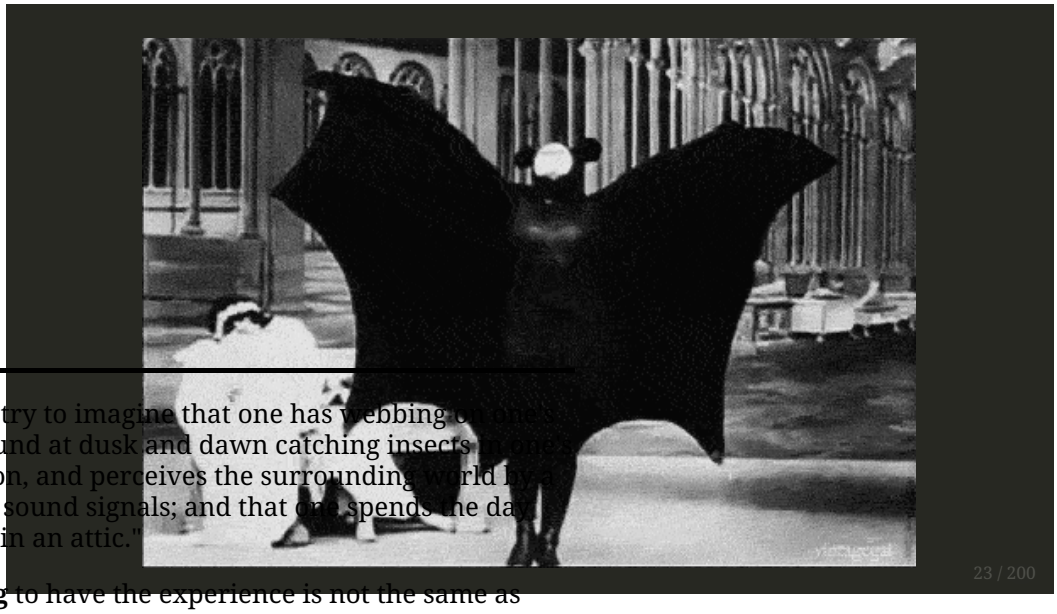
... Nagel is aware that his audience, like you, students, might be doubtful that there is real utility in the distinctions he's trying to draw. Let's see if we can get there.

What is it like to be a bat?

"Facts about what it is like to be an *X* are very peculiar, so peculiar that some may be inclined to doubt their reality, or the significance of claims about them." (p437)

Here is, roughly, the state of affairs as Nagel sees it (I'm freely mixed quoting here):

1. Bats have experience.
2. So, there is something it is like to be a bat.
3. Bat sonar, though clearly a form of perception, is not similar in its operation to any sense that we possess.
4. There is no reason to suppose that bat sonar is subjectively like anything we can experience or imagine.
5. If I try to imagine what it's like to be a bat, I am restricted to the resources of my own mind, and those resources are inadequate to the task. (Trying on the *supposition* that I possess the internal neurophysiological constitution of a bat is irrelevant.)
6. So, if extrapolation from our own case will be involved in extrapolating what it is like to be a bat, the extrapolation cannot be complete; we cannot form more than a schematic conception of what it's like to be a bat.
7. Therefore, there are facts about minds that will elude explanation, because they cannot even be properly *formulated*.



As Nagel puts it, "It will not help to try to imagine that one has webbing on one's arms, which enables one to fly around at dusk and dawn catching insects in one's mouth; that one has very poor vision, and perceives the surrounding world by a system of reflected high-frequency sound signals; and that one spends the day hanging upside down by one's feet in an attic."

And it will not, because **pretending** to have the experience is not the same as having it!

If we "try to imagine this", we are "restricted to the resources of our [own] mind, and those resources are [simply] inadequate"

What is it like to be someone else?

"Facts about what it is like to be an *X* are very peculiar, so peculiar that some may be inclined to doubt their reality, or the significance of claims about them." (p437)

The problem is not confined to exotic cases... for it exists between one person and another. The subjective character of the experience of a person deaf and blind from birth is not accessible to me, for example, nor presumably is mine to him. This does not prevent us from believing that the other's experience has such a subjective character" (p440)

What is it like to be us?

"Facts about what it is like to be an *X* are very peculiar, so peculiar that some may be inclined to doubt their reality, or the significance of claims about them." (p437)

in contemplating the bats we are in much the same position that intelligent bats or Martians would occupy if they tried to form a conception of what it was like to be us. The structure of their own minds might make it impossible for them to succeed, but we know they would be wrong to conclude that there is not anything precise that it is like to be us (p440)

"Subjective phenomenon"

"Facts about what it is like to be an *X* are very peculiar, so peculiar that some may be inclined to doubt their reality, or the significance of claims about them." (p437)

to deny the reality or logical significance of what we can never describe or understand is the crudest form of cognitive dissonance (p440-1)

The example of the bats' "bearing on the topic before us (namely, the mind-body problem) is that it enables us to make a general observation about the subjective character of experience. Whatever may be the status of facts about what it is like to be a human being, or a bat, or a Martian, these appear to be facts that embody a particular point of view." (p441)

This raises the question, of course, whether we can ever understand each other:

The point of view in question is not [necessarily] one accessible only to a single individual. ... There is a sense in which the phenomenological facts are perfectly objective: one person can know or say of another what the quality of the other's experience is. They are subjective, however, in the sense that even this objective ascription of experience is possible only for someone sufficiently similar to the object of ascription to be able to adopt his point of view—to understand the ascription in the first person as well as in the third, so to speak. The more different from oneself the other experiencer is, the less success one can expect with this enterprise. (p442)

27 / 200

What can be "objective"?

This bears directly on the mind-body problem. For if the facts of experience—facts about what it is like *for* the experiencing organism—are accessible only from one point of view, then it is a mystery how the true character of experience could be revealed in the physical operation of that organism. (p442)

What can be "objective"?

This bears directly on the mind-body problem. For if the facts of experience—facts about what it is like *for* the experiencing organism—are accessible only from one point of view, then it is a mystery how the true character of experience could be revealed in the physical operation of that organism. (p442)

- A Martian scientist could understand physical phenomena like a rainbow, but "he would never be able to understand the human concept[] of rainbow" (p443)
- "Lightning has an objective character that is not exhausted by its visual appearance, and this can be investigated by a Martian without vision. ... in understanding a phenomenon like lightning, it is legitimate to go as far away as one can from a strictly human viewpoint." (p443)
- But "what would be left of what it was like to be a bat if one removed the viewpoint of the bat?" (443). "The idea of moving from appearance to reality seems to make no sense here." (p444)

What can be "objective"?

Very little work has been done on the basic question (from which mention of the brain can be entirely omitted) whether any sense can be made of experiences' having an objective character at all. Does it make sense, in other words, to ask what my experiences are *really* like, as opposed to how they appear to me? We cannot genuinely understand the hypothesis that their nature is captured in a physical description unless we understand the more fundamental idea that they *have* an objective nature (p448)

What can be "objective"?

Very little work has been done on the basic question (from which mention of the brain can be entirely omitted) whether any sense can be made of experiences' having an objective character at all. Does it make sense, in other words, to ask what my experiences are *really* like, as opposed to how they appear to me? We cannot genuinely understand the hypothesis that their nature is captured in a physical description unless we understand the more fundamental idea that they *have* an objective nature (p448)

Strangely enough, we may have evidence for the truth of something we cannot really understand. (p447-8)

Precisifying the challenge (Chalmers)

What makes the hard problem hard and almost unique is that it goes *beyond* problems about the performance of functions.... even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience—perceptual discrimination, categorization, internal access, verbal report—there may still remain a further unanswered question: *Why is the performance of these functions accompanied by experience?*

* Chalmers (1995), "Facing up to the problems of consciousness"; in optional/supplemental readings for this module.

32 / 200

The "easy" problems

Now that we're well-warmed up, let's take a look at phenomena that we might, pre-theoretically, associate with consciousness, and see why these DON'T pose particularly grievous problems for our theorizing about the mind.

I'll follow Chalmers' discussion; he writes that, indeed, "If these phenomena were all there was to consciousness, then consciousness wouldn't be much of a problem"

Specific categories of "easy"

There are things that we are "conscious of", but it seems available to us to explain their direct presentation to to us in our experience using the standard toolkit of cognitive science as we've described it in this class.

That is, phenomena like these seem explainable (even if not yet explained) in terms of computational mechanisms/processes:

- Recognizing and categorizing objects
- Recognizing and resolving visual ambiguities
- Recognizing and resolving linguistic ambiguities (EXAMPLE)
- &c.

"Easy": Linguistic ambiguity resolution

I tickled the bird with a red tailfeather

"Easy": Linguistic ambiguity resolution

I tickled the bird with a red tailfeather

The process by which this ambiguous string is typically heard as if it only has a single intended meaning can be explained as follows:

- Language understanding involves computing a one-to-many function that maps (representations of) strings of sounds to structured representations analogous to tree diagrams

---

... ("syntactic theory", a part of linguistics, is thus a theory of which such functions are implemented by human minds)

"Easy": Linguistic ambiguity resolution

I tickled the bird with a red tailfeather

The process by which this ambiguous string is typically heard as if it only has a single intended meaning can be explained as follows:

- Language understanding involves computing a one-to-many function that maps (representations of) strings of sounds to structured representations analogous to tree diagrams
- This function can be characterized as sensitive to "top-down" contextual information to guide the selection of a specific structured representation among several:
  - If we're talking about which birds or animals we tickled... choose the structure in which *with a tailfeather* modifies *the bird*
  - If we're talking about what instruments were used to tickle with... choose the structure in which *with a red tailfeather* modifies *tickled the bird*
  - &c.

"Easy": Linguistic ambiguity resolution

I tickled the bird with a red tailfeather

The process by which this ambiguous string is typically heard as if it only has a single intended meaning can be explained as follows:

- Language understanding involves computing a one-to-many function that maps (representations of) strings of sounds to structured representations, analogous to tree diagrams
- This function can be characterized as sensitive to "top-down" contextual information to guide the selection of a specific structured representation among several:
  - If we're talking about which birds or animals we tickled... choose the structure in which *with a tailfeather* modifies *the bird*
  - If we're talking about what instruments we used to tickle with... choose the structure in which *with a red tailfeather* modifies *tickled the bird*
  - &c.
- Our sense of the *ambiguity* of the sentence's meaning is based on the output of such a process; we recognize the ambiguity only if the process fails to determine a uniquely suitable candidate in the context

38 / 200

... There's more to say than this, of course, it doesn't answer ALL questions, but it's a reasonable start to a scientific theory of a part of language understanding.

Nothing in here, though, says anything about what it's like to appreciate an English sentence (as opposed to, say, a sentence of Japanese).

But of course, there is a difference in our experience when we drive down a road lined with billboards written in English as opposed to Japanese; there is a difference in our experience when we're trying to read a book on the train and there's someone talking in English versus Japanese, etc.

Now let's look at some GENERAL categories

General categories of "easy"

More generally, there are phenomena that we associate with the notion of consciousness, and these too are not too hard to imagine explaining with the standard toolkit:

1. the reportability of mental states
2. the ability of a system to access its own internal states
3. the difference between wakefulness and sleep

... Chalmers, explicating these categories, writes that "For example, one sometimes SAYS that a mental state is conscious when it is verbally reportable, or when it is internally accessible. Sometimes a system is SAID to be conscious of some information when it has the ability to REACT on the basis of that information, or, more strongly, when it ATTENDS to that information, or when it can INTEGRATE that information and exploit it in the sophisticated control of behavior. We sometimes say that an ACTION is conscious precisely when it is deliberate. Often, we say that an organism is CONSCIOUS as another way of saying that it is awake", etc (p2, ACW caps)

General categories of "easy"

More generally, there are phenomena that we associate with the notion of consciousness, and these too are not too hard to imagine explaining with the standard toolkit:

1. the reportability of mental states
2. the ability of a system to access its own internal states
3. the difference between wakefulness and sleep

For these categories, no one really doubts that such phenomena can be explained scientifically; "we can at least state precisely *what would count* as an adequate explanation" in computational (or neural) terms

---

... let's look at how those explanations might go:

The shape of explanation for "easy"

1. Reportability
2. Accessibility
3. Wakefulness v sleep

We'll look at each in a little more detail, to get the sense of why Chalmers thinks they warrant the label "easy" at least

The shape of explanation for "easy"

1. Reportability
2. Accessibility
3. Wakefulness v sleep

It seems clear that, if we are able to report on our mental state, then we are conscious of that mental state

- e.g., to say (and mean) *I'm happy*, or, *I think she loves me*, is to report our awareness of particular desires or beliefs that we have.

... For Chalmers, this counts as an easy problem of consciousness just in case "an appropriate cognitive or neurophysiological model", specified in the **computational terms of domain, range, and mapping,** "can clearly do the explanatory work" -

The shape of explanation for "easy"

1. Reportability
2. Accessibility
3. Wakefulness v sleep

It seems clear that, if we are able to report on our mental state, then we are conscious of that mental state

- e.g., to say (and mean) *I'm happy*, or, *I think she loves me*, is to report our awareness of particular desires or beliefs that we have.

Theoretical challenge: "specify the mechanism by which information about internal states is... made available for verbal report"

Goal: "explain how a system could perform the function of producing reports on internal states"

... How might we do that? Well, we would establish some correspondence relation between the language faculty and such internal attitudes, which we know we need anyway to explain language production in general.

And to explain the eventual production, of course, we'll need some sense of how MOTIVATION figures in - that is, how our MOTIVATIONS to report our mental states leads to the productions that they do.

The shape of explanation for "easy"

1. Reportability
2. Accessibility
3. Wakefulness v sleep

It seems clear that, if we are able to react on the basis of some information, then we are conscious of that information

- e.g., you planned to turn right onto San Vicente Boulevard, but stopped suddenly after a pedestrian stepped in front of your car

The shape of explanation for "easy"

1. Reportability
2. Accessibility
3. Wakefulness v sleep

It seems clear that, if we are able to react on the basis of some information, then we are conscious of that information

- e.g., you planned to turn right onto San Vicente Boulevard, but stopped suddenly after a pedestrian stepped in front of your car

Theoretical challenge: "specify the mechanism by which information about internal states is retrieved" and fed into action planning systems

Goal: "explain how a system could ... use information about [its internal states] in directing later [computational] processes"

... Similarly here: we perceive something, and we are conscious of that perception, and that leads us to act in ways that we would not have without that awareness.

But the **general** problem here is relatively unproblematic: cognitive science will specify how perception works, in general, and tell us how what we perceive influences our actions.

The shape of explanation for "easy"

1. Reportability
2. Accessibility
3. Wakefulness v sleep

Sometimes we say that a creature is "conscious" when we just mean that they're awake

- "All it means for a system to be awake is for it to be appropriately receptive to information from the environment and for it to be able to use this information in directing behavior in an appropriate way."

Kind of a different sense of "conscious", but it is not irrelevant

The shape of explanation for "easy"

1. Reportability
2. Accessibility
3. Wakefulness v sleep

Sometimes we say that a creature is "conscious" when we just mean that they're awake

- "All it means for a system to be awake is for it to be appropriately receptive to information from the environment and for it to be able to use this information in directing behavior in an appropriate way."

Theoretical challenge: provide "an appropriate neurophysiological account of the processes responsible for organisms' contrasting behavior in those states [wakefulness vs sleep]"

Goal: ~~explain how a system with~~ those neurophysiological properties can have two states that correspond to the relevant experiences

Why think these are "easy"?

Why think these are "easy"?

As anticipated by Nagel, and echoed by Chalmers: these are problems about the explanation of **functions**

- These forms of "consciousness" can be explained by specifying a **mechanism** or algorithm that implements the appropriate function
- Extant cognitive (and neuroscientific) modeling techniques are designed for (exactly) such purposes - e.g. this is what the computational theory of mind (CTM) excels at

Why think these are "easy"?

As anticipated by Nagel, and echoed by Chalmers: these are problems about the explanation of **functions**

- These forms of "consciousness" can be explained by specifying a **mechanism** or algorithm that implements the appropriate function
- Extant cognitive (and neuroscientific) modeling techniques are designed for (exactly) such purposes - e.g. this is what the computational theory of mind (CTM) excels at

Recall Marr's levels. For a given cognitive process...

- **Computational (cognitive)**
  characterize the function in the abstract (i.e., the **function-in-extension**)
- **Algorithmic (cog/neuro)**
  specify the operations by which the computation is done (i.e., the **function-in-intension**)
- **Implementational (neuro)**
  characterize the physical states that realize the algorithm

...In Chalmer's terms: the easy problems are easy because, once we have specified the mechanism that implements the function, the bulk of our work is done

And this seems right; and it finds many analogues across the sciences:

An "easy" analogy from biology

To explain the gene, for instance, we needed to specify the mechanism that stores and transmits hereditary information from one generation to the next. It turns out that DNA performs this function; once we explain how [DNA does this], we have explained the gene.

An "easy" analogy from biology

To explain the gene, for instance, we needed to specify the mechanism that stores and transmits hereditary information from one generation to the next. It turns out that DNA performs this function; once we explain how [DNA does this], we have explained the gene.

A "hard" disanalogy from biology

There is no analoguous further question in the explanation of genes.... If someone says "I can see that you have explained how DNA stores and transmits hereditary information from one generation to the next, but you have not explained how it is a *gene*", they are making a conceptual mistake.

... "all it means to be a gene is to be an entity that performs the relevant storage and transmission function"

But when we're thinking about minds...

To explain learning, we need to explain the way in which a system's behavioral capacities are modified in light of environmental information, and the way in which new information can be brought to bear in adapting a system's actions to its environment. If we show how a neural or computational mechanism does the job, we have explained learning.

... the same sort of explanatory structure can be expected for many areas of cognitive science, namely those that correspond to the "easy" problems, e.g. explaining perception, memory, and language.

But when we're thinking about minds...

To explain learning, we need to explain the way in which a system's behavioral capacities are modified in light of environmental information, and the way in which new information can be brought to bear in adapting a system's actions to its environment. If we show how a neural or computational mechanism does the job, we have explained learning.

But if someone says "I can see that you have explained how information is discriminated, integrated, and reported, but you have not explained how it is *experienced*", they are not making a conceptual mistake. This is a nontrivial further question.

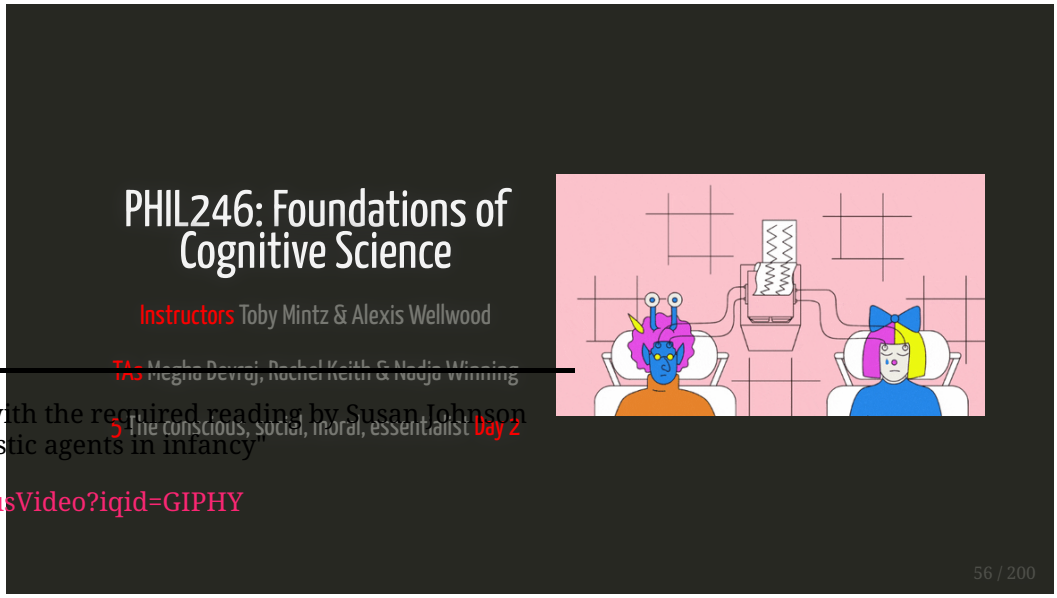
And, we'll give Chalmers the last word, echoing Nagel:

Why doesn't all this information-processing go on "in the dark", free of any inner feel? Why is it that when electromagnetic waveforms impinge on a retina and are discriminated and categorized by a visual system, this discrimination and categorization is experienced as a sensation of vivid red? We know that conscious experience *does* arise when these functions are performed, but the very fact that it arises is the central mystery.

... For Tuesday, read Susan Johnson (2000). Keep on the look out for SA4 instructions - you'll have two weeks to complete it. Submit late SA2 by Tuesday midnight if you want.

* David Chalmers (1995), "Facing up to the problems of consciousness", *Journal of Consciousness Studies, 2(3).*

55 / 200



Today we discuss BEING SOCIAL, with the required reading by Susan Johnson (2000), "The recognition of mentalistic agents in infancy"

GIF Source: http://smarturl.it/GeniusVideo?iqid=GIPHY

Agenda

1. Consciousness
2. (Rudiments of) Sociality
   - Representing others' minds
   - Representing others as agents
   - What makes a (kind of) agent?
3. Morality
4. Essentialism

Today we talk about one of the basic elements of being the kind of social animals we are - the ability to represent others as having minds, just as we do, and in particular to represent others as AGENTS, as we take ourselves to be.

It is possible to frame one version of the problem of consciousness that we discussed last week as the problem of explaining just what it's LIKE to be an agent. We have no doubt, each of us, that we have experiences, and we are well aware that are constantly performing intentional, goal-directed actions.

And, we routinely attribute the same sorts of experiences and activities to those around us. Questions we might ask are: how do we come to do this? And when? And why?

For Nagel,

(Apart from this intro, following overview in Johnson 2000, "the recognition of mentalistic agents in infancy")

The interesting problem of other minds is not the epistemological problem... It is the conceptual problem, how I *understand* the attribution of mental states to others.*

... Now, the epistemological problem may be put: "how do I know (or how can I justify the belief) that others have similar mental states to myself?" (SEP)

Nagel invites us to focus our attention on a slightly different problem, the conceptual one, which may be put: "**why do I so much as think that there are other thinking, feeling beings**?" (SEP)

Thomas Nagel (1986:19-20)

58 / 200

According to the philosopher Peter Strawson, we do this because we **must**:

[Quoted in the SEP article on the problem of other minds: https://plato.stanford.edu/entries/other-minds/]

"It is a necessary condition of one's ascribing states of consciousness, experiences, to oneself, in the way that one does, that one should also ascribe them, or be prepared to ascribe them, to others who are not oneself."*

... Why does he think that?, Well—

* P.F. Strawson (1959:99)

"It is a necessary condition of one's ascribing states of consciousness, experiences, to oneself, in the way that one does, that one should also ascribe them, or be prepared to ascribe them, to others who are not oneself."

it follows from the "primitiveness of the concept of a person", the concept of "a type of entity that *both* predicates ascribing states of consciousness *and* predicates ascribing corporeal characteristics...are equally applicable...."*

...The idea is that, if we take ourselves to be a person, then we thereby understand ourselves to be the kind of thing that is both conscious, and corporeal -- a thing with BOTH mind, and body. Understanding anyone ELSE as a person, then, should mean that we are equally well prepared to ascribe the same sorts of properties to them.

* Quoted from SEP + P.F. Strawson (1959:101-2)

Here, "the idea is that we [...] understand what one experiences when one sees another's behavior as itself requiring mental attributions ...." In other words, we see physical activity, and from it we infer mental activity -- we see physical behavior as indicative of the mental behavior for PERSONS.

We all know that we make such attributions to people, attributions of their presumed physical and mental properties; and we know that we interpret people's behavior in this way, inferring their mental states from the evidence of their perceptible behavior.

Today, we're discussing Johnson's (2000) overview paper of relevant research in psychology, which addresses the question of how or when we come to have expectations like this -- this sort of understanding of other minds. That project can be construed as providing empirical scope and limits on our understanding of the conceptual problem that Nagel drew our attention to.

Indeed, Susan Johnson echoes that problem, writing that

Quotes from the SEP article on the problem of other minds:
https://plato.stanford.edu/entries/other-minds/

The ability to construe ourselves and others as agents with minds having mental states such as perceptions, attention, desires and beliefs, is critical to humans' social, linguistic and cognitive competence. When and how this ability becomes available to us during development is therefore of particular theoretical importance.*

... Her focus is on the concept of "AGENT", and she holds that THIS -- rather than, for example, our concept of PERSON as Strawson had it -- is fundamental to many of our cognitive capacities and abilities. This conceptual grasp underpins our ability and willingness to communicate referentially with others, to predict and explain the behavior of ourselves and of others, and to manipulate complex mental states for the purposes of problem solving or deception.

* Susan C. Johnson (2000), "The recognition of mentalistic agents in infancy", Trends in Cognitive Sciences, 4(1).

61 / 200

For Johnson, Strawson's appeal to the concept of a PERSON might be a more demanding conceptual construct than what is needed to get the game of mental state attribution going, at least in the earliest stages of life.

Let's begin where Johnson begins, with

The challenge

Mental states are special in that, if one didn't **expect** to find them in the world, one might not find them very easily

The challenge

Mental states are special in that, if one didn't **expect** to find them in the world, one might not find them very easily

- They are unobservable/internal states
  - If we perceive them, we don't perceive them in the way that we perceive rocks or cows.
  - That's okay, of course; there are all sorts of unobservables that we nonetheless believe exist (e.g., atoms, prime numbers, ...)

... So, what is intentionality, again?

The challenge

Mental states are special in that, if one didn't **expect** to find them in the world, one might not find them very easily

- They are unobservable/internal states
  - If we perceive them, we don't perceive them in the way that we perceive rocks or cows.
  - That's okay, of course; there are all sorts of unobservables that we nonetheless believe exist (e.g., atoms, prime numbers, ...)
- They are a special kind thereof
  - They bear a specific kind of relationship with the world; "they are *directed at* the world; they are *about* things; they have *content*"; as philosophers put it, they are **intentional*** objects
  - Many other sorts of unobservables (e.g., atoms, prime numbers, ...) indeed might exist in the world, but they are not **about** the world.

* I'm told that, next to consciousness, the problem of intentionality is the hardest problem in cognitive science.

Intentionality*

The primary evolutionary role of the mind is to relate us in certain ways to the environment, and especially to other people. My subjective states relate me to the rest of the world, and the general name of that relationship is "intentionality." These subjective states include beliefs and desires, intentions and perceptions, as well as loves and hates, fears and hopes. "Intentionality," to repeat, is the general term for all the various forms by which the mind can be directed at, or be about, or of, objects and states of affairs in the world."

Let us turn to this quote from John Searle, philosopher of langage, mind, and social philosophy. As an aside, he is no longer professor emeritus at UC, Berkeley, as of 2019, due to violations of that university's sexual harassment policies; he had taught there since 1959)

* John R. Searle (1999), *Mind, Language and Society: Philosophy in the Real World.*

... And as an instance of what we can do with the presumption that others' have intentional mental states like we do, we can apply a straightforward pattern of reasoning to **predict** what people are likely to do:

The intentional stance*

Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs.

This is how Dan Dennett, philosopher of mind, science, and biology, puts the reasoning pattern:

The intentional stance*

Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do.

* Daniel Dennett (1996). *The intentional stance* (ACW underlining)

...

In other words, your presumptions about what others will do are based on the same sorts of calculations we use to figure out what WE should do: given our beliefs and desires, it would be rational to do X, so, when trying to figure out what someone else would do, treat them first as a rational agent like yourself, make an educated guess about what they BELIEVE and what they WANT, and this will reveal what the agent ought to do -- again, presuming that the other is rational, you predict that they will do what they ought to.

Crucially, the empirical landscape that Johnson surveys is targeted at the first step of such practical reasoning -- the presumption that the other is a rational agent -- and **when** that presumption arises.

So let us see: What evidence do we have for early recognition of anyone as a rational agent, possessed of their own internal subjective mental states?

Evidence for mental state recognition?

...

We have all of the evidence of our experience to tell us that we are the bearers of intentional states, and we have a strong belief that our fellow creatures are similarly endowed.

Because mental states are a special kind of internal state, though, it can be hard to know THAT someone recognizes the mental states of others. And it is correspondingly MORE difficult to figure out WHEN an individual might come to have that awareness.

So how would we do it?

Evidence for mental state recognition?

Garnering evidence sufficient to demonstrate mental state attributions is difficult ... Many behaviors that could potentially serve as indices of mental state attributions (e.g. pointing) can typically be interpreted in both mentalistic and non-mentalistic ways. This is particularly true for the attribution of mental states that are correlated with reality (e.g. perception or goals.)*

* Johnson (2000) p22

... How so? What does she mean here?

How would you frame a mentalistic account of pointing? E.g., you observe someone pointing at an object. What does this action signify to you on a mentalistic interpretation of the action? (It is *communicative*, it is *intentional*, more specifically it indicates an internal state of *wanting* of some sort, or...)

What would a non-mentalistic account look like? (On p25, Johnson notes that "e.g. some authors argue that 12-month-olds' produce points and requests because of their **instrumental effectiveness** long before they are actually understood for their communicative nature", based on "extensive experience interacting with their caretakers" that "may condition infants" towards certain responses to such stimuli.

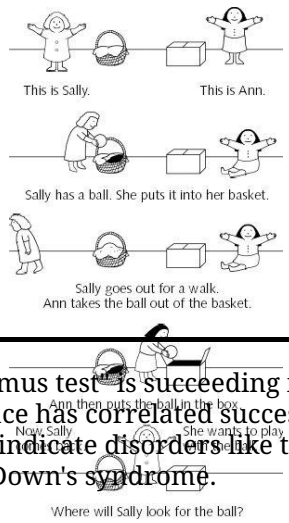
Evidence for mental state recognition?

Garnering evidence sufficient to demonstrate mental state attributions is difficult ... Many behaviors that could potentially serve as indices of mental state attributions (e.g. pointing) can typically be interpreted in both mentalistic and non-mentalistic ways. This is particularly true for the attribution of mental states that are correlated with reality (e.g. perception or goals.) The recognition of this problem led Dennett to suggest that the only convincing evidence for the attribution of mental states must come from successful reasoning about *false* beliefs, because false beliefs drive behaviors not otherwise predicted by reality.*

... she goes on to note that "Dennett's insight motivated a generation of research (typically called 'Theory of Mind' or ToM) on the development of successful false-belief reasoning in children between the ages of two and four." (22)

This is a quite important literature for cognitive science, so let us review what the best test case for a child's mental state attribution looks like, according to Dan Dennett.

* Intuition (2020) p. 22

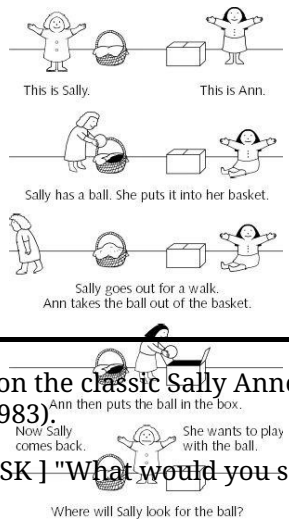


What Dennett in 1978 called "the real litmus test" is succeeding in a task like this one. And indeed, diagnostic research since has correlated success or failure in such a task at certain ages to selectively indicate disorders like those on the autism spectrum, but not, for example, Down's syndrome.

(Read more about the false belief task: https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-1698-3_91)

The Sally Anne task

Original (Baron-Cohen, Leslie, and Frith 1985) used dolls; live action version in 1988. Here Wimmer & Perner (1983).



The Sally Anne task

Original (Baron-Cohen, Leslie, and Frith 1985) used dolls; live action version in 1988. Here Wimmer & Perner (1983).

72 / 200

What I'm showing you here is a variant on the classic Sally Anne task, implemented by Wimmer and Perner (1983).

"Let's demo!" [DESCRIPTION OF THE TASK] "What would you say -- will she look in the BASKET? or the BOX?"

To succeed in this task, participants must answer the question at the end of the sequence as an adult would - and of course, answering that question correctly requires correctly representing and acting on the idea that the individual in the story -- in this case Sally -- has a false belief.

Typically-developing children succeed in this task around about the age of 4.

No indeed, while this task will certainly get at someone's representing false beliefs, but there may be reasons to think that it's not the only way. As the literature has asked, and as Johnson emphasizes, it is possible that individuals might fail a task like this for reasons other than their ability to represent individuals as rational agents—as she puts it,

(Image source: https://www.spectrumnews.org/opinion/viewpoint/1985-paper-on-the-theory-of-mind/)

Too stringent requirement?

the absence of successful false-belief reasoning before the age of four, does not mean that younger children and infants do not attribute *any* mental states to agents; only that sufficient evidence for such a claim is difficult to generate. (22)

... And indeed, suggestive evidence from the developmental psychologist Renee Baillargeon's lab finds some evidence that 2 year olds can succeed in false-belief type reasoning. The linguist Shevaun Lewis' dissertation argues that failures between 2 and 4 might just reflect issues with linguistic **performance**, rather than conceptual **competence**.

(The review article by Baillargeon discussing her work in the context of this wider literature can be found in the supplementary materials for this Module)

Too stringent requirement?

the absence of successful false-belief reasoning before the age of four, does not mean that younger children and infants do not attribute *any* mental states to agents; only that sufficient evidence for such a claim is difficult to generate. (p22)

1. between 9 and 12 mos., infants begin reliably producing and comprehending gestures like pointing
2. at 18 mos., use speaker's gaze (rather than their own) to identify the referent of a novel word
3. at 18 mos., fulfill adults' food requests based on the adults' previous emotional reaction to the food
4. at least by 2 years, explicitly mentalistic vocabulary like *want* and *see*

Here are some samples of the evidence that she considers. ...

... So the question becomes: What if children show robust evidence of representing and reasoning about others' goals, or representing people as the kinds of things that *can have* intentional goals? Would that be sufficient?

A robust sort of case study looks at how infants react to (apparently) goal-directed action,...

Goal-directedness

Other possible evidence that infants attribute mental states to people is based on infants' tendency to increase their visual attention to novel objects or events that violate their expectations.... Using the violations-of-expectations method, Woodward and colleagues found ... that seven-month-olds looked longer when a moving person collided with another person, than when inanimate objects were involved in the same collisions. (p23)

For example: ...

... Johnson wants to suggest that we can infer representation of mental states by looking at children's **expectations of agents' actions**, just like Dennett described. If a child behaves as if they are predicting others goals, we can infer that they have engaged in the kind of belief-desire representation that Dennett was talking about.

What is the reasoning that suggests THIS particular finding is evidence that 7-month-olds represent people as intentional agents?

Another example:

Goal-directedness

.redpPhillips and colleagues] ... reported evidence that 12-month-olds understand that desires (as indexed by facial expressions and direction of gaze) predict actions. They showed that 12-month-olds looked longer at a person who smiled at (expressed desire at) one object but then picked up a different object than at a person who smiled at and picked up the same object. (p23)

...

What is the reasoning that suggests THIS particular finding is evidence that 12-month-olds represent people as intentional agents?

Let's look at a study that Johnson considers in some detail.

Case study in goal-directedness

Woodward's [(1998)] study used the visual habituation method to test whether infants encode human actions in terms of the goals of the actor, or solely in terms of the spatiotemporal movements involved.... [Her] results suggest that even by five months of age, infants recognize that: (1) the behavior of some (but not all) entities is directed at the world; and (2) the identity of the entity's target is relevant, i.e. the content of the relationship is represented.

Amanda Woodward is a cognitive and developmental psychologist at the University of Chicago.

... Given these findings, Johnson suggests:

Case study in goal-directedness

Woodward's [(1998)] study used the visual habituation method to test whether infants encode human actions in terms of the goals of the actor, or solely in terms of the spatiotemporal movements involved.... [Her] results suggest that even by five months of age, infants recognize that: (1) the behavior of some (but not all) entities is directed at the world; and (2) the identity of the entity's target is relevant, i.e. the content of the relationship is represented. We can therefore say that infants attribute an intentional relationship between the object and the world (Box 1, p24)

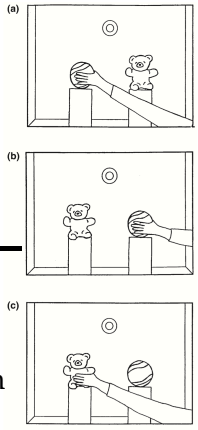
... Let's look at this study in some detail, and see whether we agree.

Case study in goal-directedness (Woodward 1998)

Habituation phase
Children are shown scenes of a hand approaching (the same of) one of two toys, until they become bored.

Test event type 1: PERCEPTUAL
The spatio-temporal path of the novel event is different from that of the habituation phase.

Test event type 2: GOAL-DIRECTED
The object targeted by the hand in the novel event is different from that in the habituation phase.



We've seen this sort of paradigm before. There are two phases in a habituation study.

In the first, the "habituation phase", infants are exposed to instances of a certain type of scene until they become bored. Their boredom is determined by measuring differences in how long they spend attending to the scene as opposed to other things in the environment. It is expected that as a given type of display is repeatedly shown, infants' looking times decrease, and will recover when shown something new just to the extent that it is interestingly different from what they've been bored by.

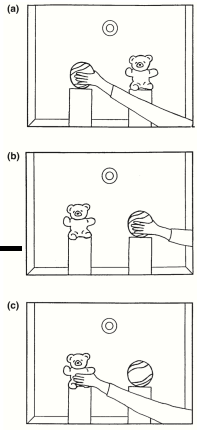
In the "test phase", then, infants are presented with something new. Johnson writes, "In studies of infant cognition, researchers are typically interested in testing whether infants interpret some given phenomena in terms of either surface, perceptual features or deeper, structural relations. Thus one test event will be perceptually more similar but structurally more different to the habituation event than the other, and vice versa. If infants interpret the habituation event in terms of its deep [or "conceptual"--ACW] structure, they are expected to recover more interest... even if [the structurally more different display] is more perceptually similar to the original." (24)

Case study in goal-directedness (Woodward 1998)

Habituation phase
Children are shown scenes of a hand approaching (the same of) one of two toys, until they become bored.

Test event type 1: PERCEPTUAL
The spatio-temporal path of the novel event is different from that of the habituation phase.

Test event type 2: GOAL-DIRECTED
The object targeted by the hand in the novel event is different from that in the habituation phase.



Now, what would you expect to happen in this study if the children represented the hand as belonging to an agent?

Suppose that the children dishabituated more to the scene where the hand was reaching for a different object (the test event type labeled "GOAL-DIRECTED"). What reasoning would you lead you to suppose that this finding is explainable in terms of children representing the hand as belong to an agent?

What do you think would happen if the hand was replaced with a metal rod, but everything else was the same?

(You might recall from the paper, that there was no difference in dishabituation between the two test event types when a metal rod was involved. How is this CONVERGENT evidence that children represented the hand as belonging to an agent? What can we say about their beliefs about the rod?

... It would be tempting to conclude from this that the notion of a PERSON performing the action--or perhaps a more general notion of something that CAN be an AGENT--is critical to representing and reasoning about GOALS.

However,

Case study in recognizing non-human agents

A number of researchers have pointed out that knowledge domains that entail domain-specific reasoning (as the mentalistic domain seems to) may also entail domain-specific object-identification processes. The object-identification processes leading to the recognition of a mentalistic agent could be isomorphic with the representation of people, but they need not be.

... The suggestion here is that the early notion, "mentalistic agent", might not be synonymous with the concept "people". In other words, it may not be that infants have the concept PERSON as Strawson said was requisite to ascribe properties both conscious and corporeal to the self and to others.

Instead, they may have domain-specific object-recognition capacities that trigger the assignment of mentalistic properties to objects that meet certain domain-specific conditions.

Case study in recognizing non-human agents

A number of researchers have pointed out that knowledge domains that entail domain-specific reasoning (as the mentalistic domain seems to) may also entail domain-specific object-identification processes. The object-identification processes leading to the recognition of a mentalistic agent could be isomorphic with the representation of people, but they need not be. (25)

- "Several theoretical proposals have been offered about the cues that lay thinkers, infants and adults alike, might use to identify the presence of mentalistic agents .... However, the degree to which any of these features might serve specific functions in infants' reasoning in the mentalistic domain, as opposed to other domains (e.g. the social or biological domains) is still largely unknown" (25)

Now here she's said "domain-specific", but of course, we can't be sure, at the current stage of inquiry, that that's what we're dealing with ...

I will say that, in general, we are still very much in the early stages of understanding the taxonomy of the mind.

Still, let's take a look at some evidence that Johnson considers about apparently specific cues that might be invoked in order to attribute mental states to things in the environment.

Case study in recognizing non-human agents

Possibly domain-specific, object-recognition cues to **possible** agents:

- possession of face and eyes
- asymmetry along one axis
- non-rigid transformation
- self-propulsion
- ability to engage in contingent and reciprocal interaction with other entities

Importantly, Johnson notes that "the ability of infants to detect these features [in the first place] goes uncontested" (25) -- what is in question is whether these specific features might cue the recognition not of mere entities, but of entities with **intentions**.

...

83 / 200

In Johnson's case study, the faetures highlighted in red are manipulated:

(GIF source giphy.com; "The Ministry of Silly Walks" is a sketch from the Monty Python comedy troupe's television show Monty Python's Flying Circus, season 2, episode 1)

Case study in non-human agents (Johnson et al 1998)

Johnson and colleagues tested a subset of these potential cues and found that both the presence of a face, and contingently interacting behavior even in the absence of a face, proved powerful ... cues. They tested the willingness of 12-month-olds to follow the 'gaze' of novel objects as a function of the cues embodied in the object.

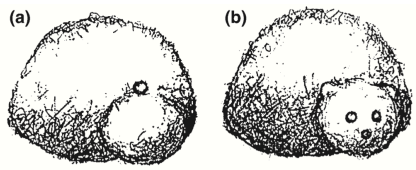


Fig. I. The novel object without a face (a) and with a face (b). (Reproduced, with permission, from Ref. a.)

Case study in non-human agents (Johnson et al 1998)

Johnson and colleagues tested a subset of these potential cues and found that both the presence of a face, and contingently interacting behavior even in the absence of a face, proved powerful ... cues. They tested the willingness of 12-month-olds to follow the 'gaze' of novel objects as a function of the cues embodied in the object.

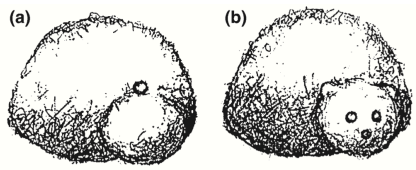


Fig. I. The novel object without a face (a) and with a face (b). (Reproduced, with permission, from Ref. a.)

Two factors tested:

1. Has a face / is faceless
2. Possesses contingently interacting behavior / not contingently

Their novel entity embodied some of the cues for mentalistic agents without being person-like: it was "The size of a small beach ball, it was made of natural-looking fuzzy brown fur and had a naturalistic shape that was symmetrical along only one axis with a small cone-shaped bulge at one end".

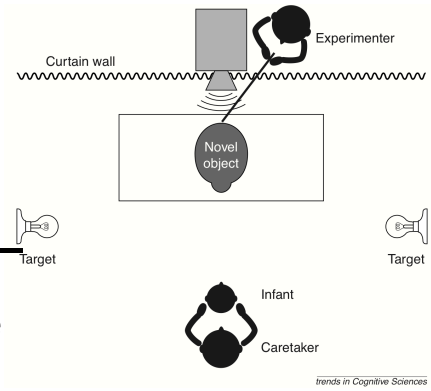
So much remained constant; but, between children tested, they manipulated whether or not the novel entity had a face, as on the right here, versus not, as on the left.

Case study in non-human agents (Johnson et al 1998)

Johnson and colleagues tested a subset of these potential cues and found that both the presence of a face, and contingently interacting behavior even in the absence of a face, proved powerful ... cues. They tested the willingness of 12-month-olds to follow the 'gaze' of novel objects as a function of the cues embodied in the object.

Two factors tested:

1. Has a face / is faceless
2. Responds contingently / not contingently



86 / 200

"[The] 'behavior' [of the novel entity] was generated via a small remote-controlled beeper and incandescent light hidden inside" - for some children, the novel entity responded contingently, such that "when the infant babbled, the object babbled back and when the infant moved, the internal light flashed in response". For others, the novel entity acted randomly, flashing its light and beeping whenever, regardless of what the child was doing.

These two factors - whether the entity has a face, and whether it responds contingently to the child's actions - led to four experimental conditions, each of which was tested on a different group of children:

Case study in non-human agents (Johnson et al 1998)

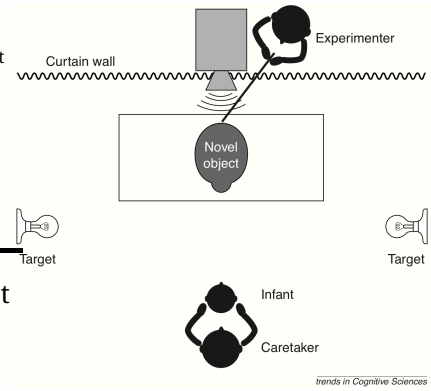
In the set-up, infants were familiarized with the novel entity for 60 seconds, during which the entity responded to the infant or randomly, as determined by the child's test condition below. There was one final "attention-getting" beep before the object would pivot to orient itself to one of two targets on either end of the scene. Infants' gaze direction and duration at this point were measured.

Test condition 1: Has a face, responds contingently

Test condition 2: Has a face, responds non-contingently

Test condition 3: Faceless, responds contingently

Test condition 4: Faceless, responds non-contingently



87 / 200

Let's break down what we might expect across these conditions, considering just infants GAZE DIRECTION -- where they look upon the final beep and the entity shifting to "face" one of the two targets.

If having a face were sufficient to cue children to the presence of an intentional agent, what would you expect across these conditions?

If responding contingently to the child's own actions were sufficient to cue children to the presence of an intentional agent, what would you expect?

What if the child would only use both cues in combination? What would you expect then?

And of course, what did she find?

Case study in non-human agents (Johnson et al 1998)

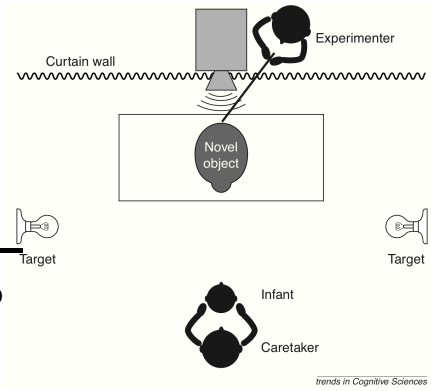
Test condition 1: Has a face, responds contingently

Test condition 2: Has a face, responds non-contingently

Test condition 3: Faceless, responds contingently

Test condition 4: Faceless, responds non-contingently

Johnson et al found that infants followed the gaze of the novel object whenever it had a face, and whenever its behavior was responsive to the infants' own behavior. Infants only failed to follow the gaze of the object when it **both** lacked a face **and** when its responses were uncorrelated with the infants' own behavior.



... Johnson et al, like good mentalists, concluded that 12 month olds seem able to attribute mental states such as perception and attention to novel entities other than people.

[DESCRIBE RESULTS]

This study thereby demonstrates BOTH intentional attribution, AND a potentially wider scope for the understanding of "agency" in infancy.

She writes,

No difference human/non-human (Johnson et al 1998)

Importantly, the object in the non-contingent, faceless condition embodied the same shape and movement cues as it did in the other conditions, but infants showed no reliable sign of following its 'gaze'. A further comparison condition with unfamiliar adults taking the place of the object suggested that infants were no more likely to follow the gaze of a contingently interacting person (with a face) than a contingently interacting fuzzy brown object with a face. Taken together, these results seem to show that infants use quite selective cues to decide when an object does or does not have a mind to perceive or attend with, specifically the presence of a face, or the propensity to interact contingently. (p26, Box 2)

Test condition 1: Has a face, responds contingently

Test condition 2: Has a face, responds non-contingently

Test condition 3: Faceless, responds contingently

Test condition 4: Faceless, responds non-contingently

This is quite compelling research, but

Caution is warranted (as always)

The existence of specialized object-identification processes by which we determine goal-directedness would not in and of itself conclusively show that infants have a *mentalistic interpretation* of those objects.

Non-mentalistic accounts could predict that infants' competency would vary in the ways that Woodward and Johnson tested (as Johnson discusses).

However, **convergent evidence** could make the mentalistic interpretation yet more plausible, even in the absence of success in false-belief tasks:

- convertent results across a wide variety of methodological paradigms
- in different labs
- with different materials ...

... Johnson is confident that such convergent evidence exists, and that more can be found; if so, this would constitute good evidence for the mentalistic approach, since the non-mentalistic explanations will have to vary along with each case:

Seeking convergent evidence

the mechanisms underlying infants' recognition of, and interaction with, agents and what meaning, if any, infants attach to them, are not fully understood. Methodological progress on how to satisfy the evidentiary criteria for the attribution of mental states such as perception, attention, or goals has been slow. One point is clear, however: non-mentalistic accounts... give rise to the prediction that the infant's competency in this domain will vary across different behavioral contexts (e.g. gaze-following, communication, and imitation). (27)

Why does she say that?

Seeking convergent evidence

the mechanisms underlying infants' recognition of, and interaction with, agents and what meaning, if any, infants attach to them, are not fully understood. Methodological progress on how to satisfy the evidentiary criteria for the attribution of mental states such as perception, attention, or goals has been slow. One point is clear, however: non-mentalistic accounts... give rise to the prediction that the infant's competency in this domain will vary across different behavioral contexts (e.g. gaze-following, communication, and imitation). ... Conversely, the degree to which putatively mentalistic attributions are manifested across a variety of diverse behaviors and contexts is evidence against non-mentalistic interpretations. (27)

... Specifically,

Seeking convergent evidence

the mechanisms underlying infants' recognition of, and interaction with, agents and what meaning, if any, infants attach to them, are not fully understood. Methodological progress on how to satisfy the evidentiary criteria for the attribution of mental states such as perception, attention, or goals has been slow. One point is clear, however: non-mentalistic accounts... give rise to the prediction that the infant's competency in this domain will vary across different behavioral contexts (e.g. gaze-following, communication, and imitation). ... Conversely, the degree to which putatively mentalistic attributions are manifested across a variety of diverse behaviors and contexts is evidence against non-mentalistic interpretations. (27)

if the selective set of object features that elicit gaze-following in Johnson et al.'s study do so because they invoke an intermediary representation ([i.e., one in between] perception and action*) of mentalistic being, that representation should also be available to support other behaviors thought to be based on mentalistic attributions.

* A "concept", anyone?

Seeking convergent evidence

the mechanisms underlying infants' recognition of, and interaction with, agents and what meaning, if any, infants attach to them, are not fully understood. Methodological progress on how to satisfy the evidentiary criteria for the attribution of mental states such as perception, attention, or goals has been slow. One point is clear, however: non-mentalistic accounts... give rise to the prediction that the infant's competency in this domain will vary across different behavioral contexts (e.g. gaze-following, communication, and imitation). ... Conversely, the degree to which putatively mentalistic attributions are manifested across a variety of diverse behaviors and contexts is evidence against non-mentalistic interpretations. (27)

~~if the selective set of object features that elicit gaze-follow~~ing in Johnson et al.'s study do so because they invoke an intermediary representation ([i.e., one in between] perception and action) of mentalistic being, that representation should also be available to support other behaviors thought to be based on mentalistic attributions. If results consistent with mentalistic interpretations are obtained in divergent methodological paradigms, objections based on local interpretive issues of individual methods will no longer obtain.

PHIL246: Foundations of Cognitive Science

Instructors Toby Mintz & Alexis Wellwood

TAs Megha Devraj, Rachel Keith & Nadja Winning

5 The conscious, social, moral, essentialist Day 3

95 / 200

Bridge: sociality and morality

We saw last time that representing another as a RATIONAL AGENT figures into a very practical form of reasoning that plausibly all of us use all the time: first, you assume that someone is a rational agent (just as you are); next, you assume that they will act in accord with their beliefs and desires, just as you would; and then, assuming that you have a reasonably accurate sense of their beliefs and desires, you can predict with reasonable accuracy what they ought to do.

We furthermore saw evidence that we engage in this type of reasoning as early as *infancy*, and some reasons for thinking that what counts as an "agent" (at least at this stage of development) is rather more general than what counts as a "person". For example, even a fuzzy brown thing can count as an agent, so long as it either has a face, or interacts with you in a manner contingent on your own behavior.

All of this -- these processes of predicting what people will do, and of deciding whether they are worth predicting -- depend on the ability to represent that class of odd and special unobservables -- **intentional states**.

The focus on intentional states is similarly critical for thinking about morality, our topic for today:

Bridge: sociality and morality

The question is not simply, "When do children understand others' actions?" but rather, "When do children understand others' actions in the light of the actors' beliefs, desires, and intentions?" [This] focus on psychological states is ... fundamental in the domain of moral judgment.

* Baird & Astington 2004, "The role of mental state understanding in the development of moral cognition and moral action."

Bridge: sociality and morality

The question is not simply, "When do children understand others' actions?" but rather, "When do children understand others' actions in the light of the actors' beliefs, desires, and intentions?" [The] focus on psychological states is ... fundamental in the domain of moral judgment.

As Kant ... remarked, "When moral worth is in question, it is not a matter of actions which one sees but of their inner principles which one does not see"

* from Kant's *Foundations of the Metaphysics of Morals*, cited in Baird & Astington 2004

98 / 200

Indeed, as Immanuel Kant, a central figure in the Enlightenment period in Western philosophy, put it: ...

... To see this:

Imagine a court of law. Someone has died, and someone else appears to have been critically involved in the causal chain leading up to their death. The prosection is seeking a murder charge. What is an important and relevant thing **in this context** that would need to be shown in order to successfully prosecute the defendant for murder?

Today, focusing on Kiley Hamlin's (2013) overview, we consider that...

[The] tendency to see certain actions and individuals as right, good, and deserving of reward, and others as wrong, bad, and deserving of punishment—generally referred to as a *moral sense*—is often considered a defining feature of what it means to be human. ... all normally developing adults in all cultures have the basic notion that some things are right and others are wrong (Brown, 1991). Where does this moral sense come from?

... We will follow Hamlin's framing,

Kiley Hamlin (2013), "Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core". *Current Directions in Psychological Science*.

99 / 200

Agenda

1. Consciousness
2. Sociality
3. Morality
   - Moral goodness
   - Moral evaluation
   - Moral retribution
4. Essentialism

---

looking for critical components of the human "moral sense" in infancy: the three components she identifies for this sense is, first, that one *be* morally good (and knows which actions are "right" as opposed to "wrong"); second, of *evaluating* the moral valence of others' actions (whether they are "good" or "bad"); and finally, of *rewarding* or *punishing* others for "immoral" actions.

We'll begin where she does, with

The traditional view in psychology*

Psychologists posited "that infants start out as fundamentally different moral creatures from adults—either *amoral* (possessing no moral sense) or *immoral* (possessing a moral sense that opposes adults', perhaps because of selfishness or cognitive limitations)—and acquire a mature moral sense over time through various developmental processes." (186)

... Sound familiar?

* According to Hamlin (2013).

The traditional view in psychology*

Psychologists posited "that infants start out as fundamentally different moral creatures from adults—either *amoral* (possessing no moral sense) or *immoral* (possessing a moral sense that opposes adults', perhaps because of selfishness or cognitive limitations)—and acquire a mature moral sense over time through various developmental processes." (186)

Though there is moral *development*, and *differences* in the moral states attained across cultures, "development and cultural variation do not themselves preclude the existence of innate capacities. As has been argued for humans' 'core' understanding of certain conceptual domains—including objects, numbers, geometry, agents, and groups (see Spelke, 1994; Spelke & Kinzler, 2007)—there may be aspects of morality that emerge in the absence of experiences" (186)

... How about now?

* According to Hamlin (2013).

Empiricism v. nativism redux

What experiences *could* do it?

- "being helped or harmed in particular ways/situations, observing others be similarly helped or harmed, or being explicitly taught which acts are right and which are wrong." (186-7)

Empiricism v. nativism redux

What experiences *could* do it?

- "being helped or harmed in particular ways/situations, observing others be similarly helped or harmed, or being explicitly taught which acts are right and which are wrong." (186-7)

Always the prior question:

- how would such acts, experiences, and teachings be **encoded** as moral *as such*, without prior access to the appropriate moral *concepts*?

(Is this an *a priori* challenge for the traditional picture in psychology already?)

Why should an innate moral core have developed?

Hamlin suggests, following researchers like Michael Tomasello, that the moral sense is the result of a "co-evolution" between cooperation and morality.

Why should an innate moral core have developed?

Hamlin suggests, following researchers like Michael Tomasello, that the moral sense is the result of a "co-evolution" between cooperation and morality.

> Evolutionary biologists, anthropologists, and primatologists are increasingly exploring the integral nature of morality to successful group living [...]. Though the details vary, these functional accounts generally posit that the moral sense evolved to sustain collective action and cooperation—which lead to great mutual gain but sometimes require personal sacrifice—within groups of unrelated individuals. (187)

What would it consist in/be responsive to?

Cooperative systems require (at least) three abilities/propensities, all relevant to the moral sense:

What would it consist in/be responsive to?

Cooperative systems require (at least) three abilities/propensities, all relevant to the moral sense: (a) moral *goodness*: the tendency to feel concern for and help others despite personal costs, presumably rooted in empathetic processes;

What would it consist in/be responsive to?

Cooperative systems require (at least) three abilities/propensities, all relevant to the moral sense: (a) moral *goodness*: the tendency to feel concern for and help others despite personal costs, presumably rooted in empathetic processes; (b) moral *evaluation*: the ability to identify and dislike those who are uncooperative/unempathetic/unhelpful or who are likely to be in the future, requiring an ability to analyze others' social behaviors;

What would it consist in/be responsive to?

Cooperative systems require (at least) three abilities/propensities, all relevant to the moral sense: (a) moral *goodness*: the tendency to feel concern for and help others despite personal costs, presumably rooted in empathetic processes; (b) moral *evaluation*: the ability to identify and dislike those who are uncooperative/unempathetic/unhelpful or who are likely to be in the future, requiring an ability to analyze others' social behaviors; and (c) moral *retribution*: the tendency to carry out, or to support, punishment of those who misbehave, which could involve emotional processes, behavioral analyses, or both. (p187)

... A lot of the linking between these features and the Tomasello idea that they evolved for the purposes of cooperative behavior.

For example, why should we think that each of (a)-(c) important for cooperative behavior?

What do you suppose the world would look like if most people lacked the capacity for (a)-(c)? Why?

Hamlin considers each of these abilities/propensities in turn, discussing the empirical evidence for each at the earliest stages of human life. We'll follow her presentation, looking at selected studies in somewhat more detail.

(a) Moral Goodness

IF there is an innate moral sense,
THEN there should be evidence for moral goodness early in human infancy

(a) Moral Goodness

IF there is an innate moral sense,
THEN there should be evidence for moral goodness early in human infancy

A growing body of developmental data suggests [it] is present in infancy and is rooted in empathy. From birth, newborns show rudimentary emotional reactions to other's suffering [...]. As soon as they are physically able, infants begin to supplement those emotional responses with a variety of active prosocial behaviors, including comforting those in distress, helping others achieve goals, informing others of things they should know, and sharing their own resources (p187)

... What does all this mean? Well, it appears that infants are aware, on some level, of what counts as morally good behavior, and they act appropriately.

What does it mean to say it is "rooted in empathy?"

I think she has in mind a big hypothesis by Michael Tomasello, who she cites here but doesn't elaborate; here's a clue according to him: "Evolutionarily, the key difference [between humans and our otherwise social primate cousins] is that humans have evolved not only social-cognitive skills geared toward competition, but also social-cognitive skills and motivations geared toward complex forms of cooperation—**what we call skills and motivations for shared intentionality**" https://www.eva.mpg.de/documents/Sage/Tomasello_Ape_CurrDirPsychScience_2010_1552616.pdf

Tomasello: "shared intentionality"

Shared intentionality, sometimes called 'we' intentionality, refers to **collaborative interactions in which participants share psychological states with one another** (Gilbert, 1989; Searle, 1995; Tuomela, 1995). Dec 20, 2006

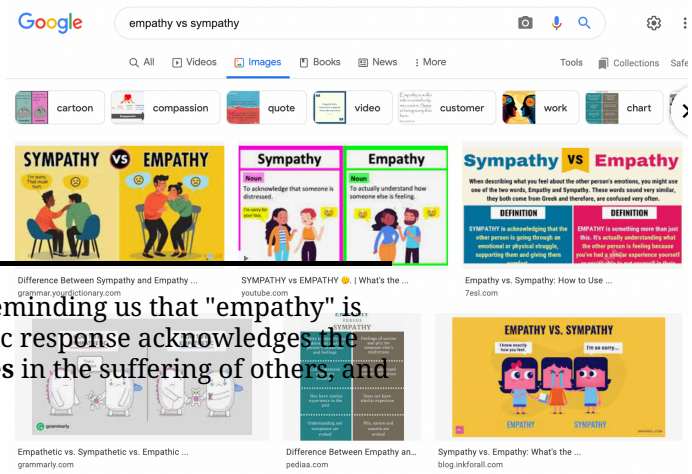
https://onlinelibrary.wiley.com › doi › full

Shared intentionality - Tomasello - 2007 - Wiley Online Library

... "As a specification of [his] cultural intelligence hypothesis", another way of putting it is that the capacity for "shared intentionality" is "the ability to participate with others in collaborative activities with shared goals and intentions." Source: https://www.pnas.org/content/115/34/8466

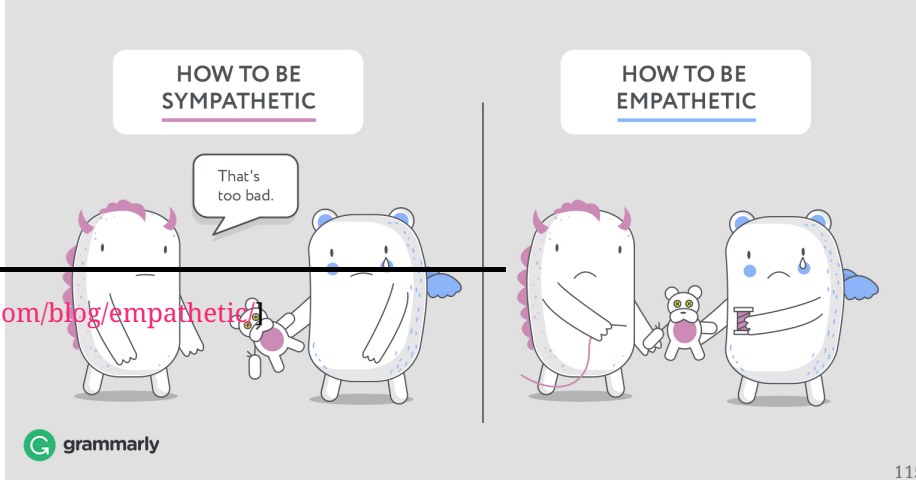
113 / 200

Tomasello: "shared intentionality"



Now, the internet and other sources are fond of reminding us that "empathy" is critically different from "sympathy": a sympathetic response acknowledges the suffering of others, an empathetic response **shares** in the suffering of others, and in the search for solutions to the suffering. E.g.,

Tomasello: "shared intentionality"



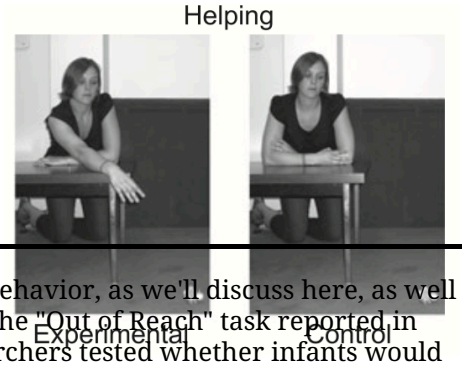
[Image source: https://www.grammarly.com/blog/empathetic/]

115 / 200

Case study in moral goodness

... Now let us consider a case study in moral goodness, from Dunfield et al (2011). The question is, do infants show empathetic responses to people in need?

Case study in moral goodness



Dunfield K.A. (2013), "Examining the diversity of prosocial behavior: Helping, sharing, and comforting in infancy", *Infancy 16(2)*, pp227-247.

These researchers examined "Helping" behavior, as we'll discuss here, as well as others (sharing, and comforting). Using the "Out of Reach" task reported in Warneken & Tomasello (2006), the researchers tested whether infants would engage in helping behavior under some conditions but not others. In the experimental condition, the experimenter demonstrated a need, and in the control condition, she did not demonstrate a need:

- In the experimental condition, the experimenter walks a small plastic toy across a table and then drops it off the edge, saying "Oops!". Then, she reaches for the toy, gazing at it for 5 seconds, followed by alternating gaze between the toy and the infant until the infant responds, or a total of 10 seconds have elapsed.

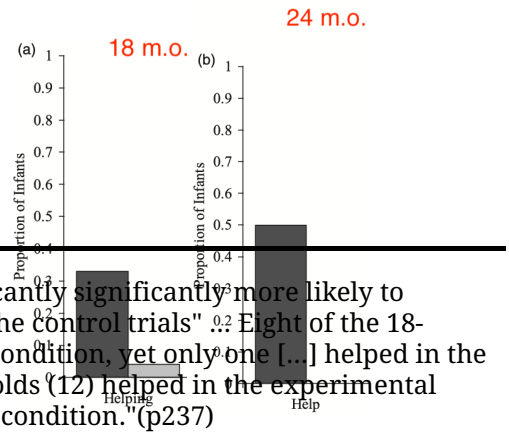
- In the "control" variant, instead of dropping the toy she places it and says "There!", folding her arms at the edge of the table, looking towards the infant for 10 seconds, before moving on to the next task.

Each infant saw both the experimental and control conditions, interspersed with other tasks. The trials were extremely similar apart from the "need" demonstrated: "the participant observed the experimenter engage in very similar actions, over the same period of time, with a similar amount of eye contact and attention toward the infant" (p232).

Here, 'helping' behavior was coded as retrieving the toy from the ground and giving it to the experimenter. Non-helping behaviors "included ignoring the toy, picking up the toy and playing with it, or explicitly refusing to help (e.g., shaking head or verbally saying “No”)." (p.236)

Here's what they found:

Case study in moral goodness



Both 18 and 24 month olds were "significantly significantly more likely to provide aid in the experimental versus the control trials" ... Eight of the 18-month-olds helped in the experimental condition, yet only one [...] helped in the control condition. Half of the 24-month-olds (12) helped in the experimental condition, but none did so in the control condition."(p237)

These results "suggest[] that[,...] by at least 18 months of age, infants are able to recognize [...] needs and are beginning to show the ability to spontaneously engage in" prosocial behavior. "This is the case even when the recipient of aid is an adult whom they had just met and the behavior provides no immediate benefit to themselves."

And importantly, even though not all infants engaged in helping behavior, those that *did* did so more often when there was a demonstrated need, as opposed to when there wasn't. (p241)

Now, the authors say that this behavior is "spontaneous" because (i) no one asked for help, (ii) there was no guidance as to what counted as "correct" behavior, (iii) the parent didn't encourage the infant to help, and (iv) there was no reward offered. "Further," the authors write, "the majority of the infants began to respond within the first 5 sec of the trial, before the experimenter directed her gaze to the infant." (p243)

Questions about this?

(b) Moral Evaluation

IF there is an innate moral sense,
THEN there should be evidence for moral evaluation early in human infancy

(b) Moral Evaluation

IF there is an innate moral sense,
THEN there should be evidence for moral evaluation early in human infancy

a growing body of work suggests that preverbal infants understand several types of cooperative behaviors that occur between third parties. By the end of their 1st year, infants grasp that agents work together toward shared goals..., categorize goal-helping as positive and goal-hindering as negative..., and understand how being helped and hindered influences one's social preferences. (p187)

Case study in prosociality (Hamlin et al 2007)

"In addition to understanding cooperatively relevant behavior, infants positively evaluate those who cooperate and negatively evaluate those who do not." She cites herself with colleagues Karen Wynn and Paul Bloom as providing the first studies of this nature.

Case study in prosociality (Hamlin et al 2007)

The "morality play" paradigm testing "goal-failure scenarios"

- **Hill**: a wooden puppet trying to reach the top of a hill
- **Box**: an animal puppet trying to open a box
- ...

Children exposed either to "Helper" or "Hinderer" scenario

- **"Helper" facilitates** puppet P's goal (bump up the hill, open the box)
- **"Hinderer" prevents** P's goal (bump down the hill, slam the box closed)

Infants' preference for the individual Helper or Hinderer determined by "reach behavior"

... The challenge for the empiricist is always to demonstrate how one **could** learn to represent others as mentalistic agents.

Hamlin is asking: how do we learn how we should work together as intentional agents? When and how do we learn to be *prosocial*, if it is in fact learned at all.

And now, as advertised, one of my favorite experiments of all time; but we'll hear it from Kiley herself.

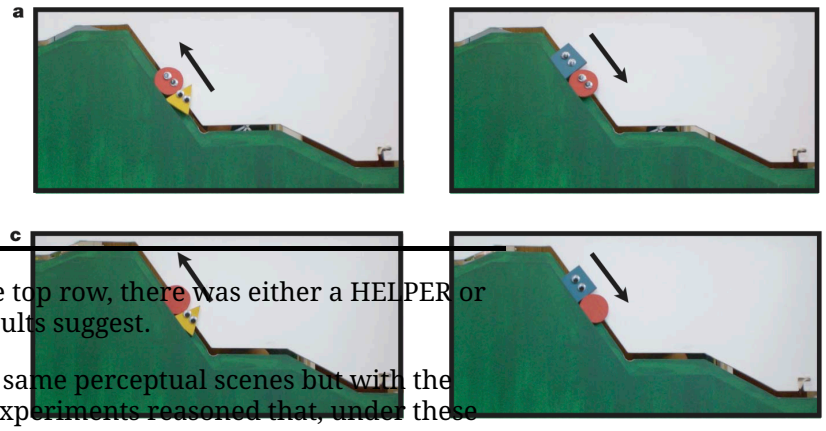
Case study in prosociality (Hamlin et al 2007)

... in subsequent studies, they observed the same pattern with 6 and 10 month old infants - each group showed strong evidence of *representing* and showing *positive preference for* individuals that engage in helping behaviors.

Importantly, here too we see evidence for a hard-wiring interaction with lower level systems of object recognition, as we did in the Johnson study:

* 0:00-0:52, https://www.youtube.com/watch?v=anCaGBsBOxM ; from Hamlin, Wynn, & Bloom (2007), "Social evaluation by preverbal infants"

Case study in prosociality (Hamlin et al 2007)



In the experiments we just saw, as on the top row, there was either a HELPER or a HINDERER. At least, that's what the results suggest.

In the next experiments, infants saw the same perceptual scenes but with the eyes on the yellow circle removed. The experiments reasoned that, under these circumstances, the object should appear as if it were not undergoing any **self-propelled motion**, and so should feel less inclined to view it as an AGENT.

If the object is not viewed as an AGENT, then all of the social categories that are deployed for such entities should not be deployed, including any assignment of certain social or moral significance to what they're doing or what is done to them.

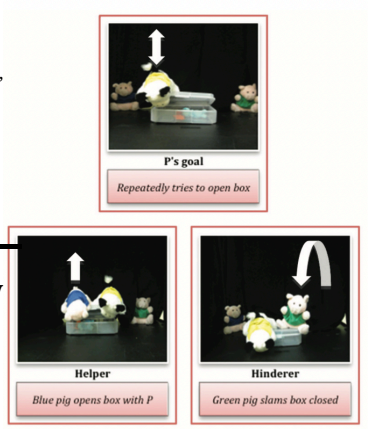
As Hamlin puts it in the 2007 paper, "Although the two characters' physical trajectories and respective effects on the pushed object were similar to those of the helper and hinderer" before, "these events are not social interactions and cannot be viewed as instances of 'helping' and 'hindering'. If perceptual preferences, not social evaluations, drove infants' choices in" the previous experiment, then "similar preferences should be obtained here:" - in other words, infants should simply prefer the "pusher-up character" or the yellow triangle, and disprefer the "pusher-down character", i.e. the blue square.

But neither age group did so: Infants at 6 months chose at a rate of roughly 50/50 between the yellow triangle and the blue square.

From Hamlin et al 2007

Case study in prosociality (Hamlin & Wynn 2011)

In a different follow-up discussed in the 2013 paper, Hamlin & Wynn found that infants preferred puppets who "helped" another puppet open a box, as opposed to puppets who "hindered" the puppet from opening the box, e.g. by slamming it shut.

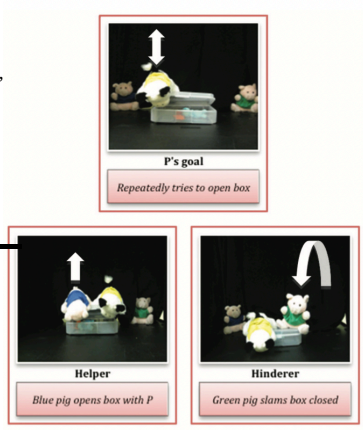


What they did next should remind you of what happened in the metal rod study we discussed last time:

Case study in prosociality (Hamlin & Wynn 2011)

In a different follow-up discussed in the 2013 paper, Hamlin & Wynn found that infants preferred puppets who "helped" another puppet open a box, as opposed to puppets who "hindered" the puppet from opening the box, e.g. by slamming it shut.

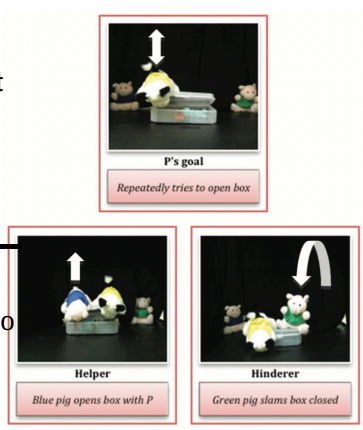
To ensure that perceptual cues could not underlie infants' preferences for characters in the box and ball scenarios, we showed age-matched infants 'control' plays, in which an inanimate claw, rather than an animate puppet, tugged on a box lid [...]. Identical puppets then directed the exact same actions toward the claw as they had directed toward P during 'test' plays. (p188)



Case study in prosociality (Hamlin & Wynn 2011)

Yet, despite the actions at test being identical to those of the previous study, here

infants did not prefer box-opening... puppets. This result is consistent with empirical work suggesting that infants do not attribute unfulfilled goals to nonagents, such as claws [...]. Because only agents have goals, puppets' actions toward nonagents are neither prosocial nor antisocial. Like adults do when making moral judgments, then, infants evaluate Helpers and Hinderers for their goal-relevant, social behaviors, rather than for performing some specific physical act or causing some outcome. (p189)



... In other words, infants know something about not only what counts as the morally right thing to do, they evaluate others for the same traits, and respond to them accordingly.

When we recognize things as *agents* with *goals*, we know that it is good to help them fulfill their goals; people (or puppets) who fail to do so should be avoided.

Let's loop this back to mental states:

Goals → mental states (?)

When determining moral responsibility, adults consider not only whether an individual facilitated or blocked a goal but the *mental states* that drove the action. Indeed, because humans sometimes act in error (e.g., unintentionally, naïvely), and because sometimes prosocial and antisocial behaviors are physically identical, mental states are critical for accurately understanding what individuals are really like and how they will likely behave in the future (p189)

Recall Kant...

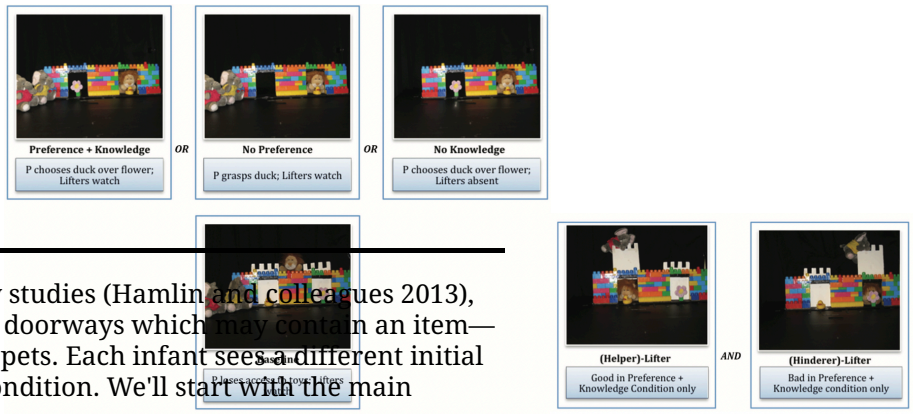
Goals → mental states (?)

When determining moral responsibility, adults consider not only whether an individual facilitated or blocked a goal but the *mental states* that drove the action. Indeed, because humans sometimes act in error (e.g., unintentionally, naïvely), and because sometimes prosocial and antisocial behaviors are physically identical, mental states are critical for accurately understanding what individuals are really like and how they will likely behave in the future (p189)

To accurately evaluate a giving behavior, then, one must read the mental states of all parties: the giver's intention (Mental State 1) toward the recipient, which is based on the giver's knowledge (Mental State 2) of the recipient's preference (Mental State 3). (p189)

... What evidence do we have that infants in fact do this? Here's an example:

Case study in "mind-reading" (Hamlin et al 2013)



The basic set-up in this series of morality studies (Hamlin and colleagues 2013), there is a protagonist—here, a lion—two doorways which may contain an item—a flower or a duck—and two "Lifter" puppets. Each infant sees a different initial set-up depending on the experimental condition. We'll start with the main condition, "Preference + Knowledge"

Here, the lion displays a clear preference for one of two items, with the Lifters onlooking, implying that they are aware of the lion's preference.

Next, two little fences come down, blocking the lion's access to the items. The researchers were interested in seeing whether infants' would respond differently to the two Lifter puppets depending on what they do next.

- One of the Lifters acts to lift the door that will allow P to reach the toy he had chosen -- that Lifter might be interpreted as a Helper, and thereby good.
- The other Lifter acts to lift the door that will allow P to reach the toy he had avoided -- that Lifter might be interpreted as a Hinderer, and thereby bad.

And the researchers found that, despite their physically identical actions, infants strongly preferred the Helper-Lifter to the Hinderer-Lifter.

Two additional conditions systematically removed either the lion's, or the Lifter's, (189) mental states:

- In the "No Preference" condition, there is just one item that the lion can go to, so when he does, the Lifters should have no cause to think that the lion's selection expresses a **preference** for that object.
- In the "No Knowledge" condition, the set-up is the same as in initial "Preference + Knowledge" condition, except that the Lifters are not present, so they have no knowledge of the lion's preference.

And everything else was the same. Here, the reserachers found that infants chose the two Lifters equally in each of these conditions, "indicating that in the preference + knowledge condition, they did not simply prefer the puppet who happened to facilitate P's goal or who allowed P to grasp the toy he had grasped

before: Infants evaluated Lifters on the basis of their prosocial and antisocial mental states" (p190)

Case study in "mind-reading" (Hamlin et al 2013)

Critically, these results demonstrate that infants' evaluations are not limited to simple heuristics like 'causing-someone-to-do-X = good,' 'former-action-facilitating = good,' or even 'intention-facilitating = good'; these heuristics applied to conditions in which infants failed to distinguish characters. Instead, like adults, infants evaluate others as good and bad mentalistically: Good guys are those who knowingly and intentionally facilitate a third party's goal; bad guys are those who knowingly and intentionally block it. (p190)

(c) Moral Retribution

IF there is an innate moral sense,
THEN there should be evidence for moral retribution early in human infancy

(c) Moral Retribution

IF there is an innate moral sense,
THEN there should be evidence for moral retribution early in human infancy

moral evaluations need 'teeth'. ... In some cases, intentionally and knowingly blocking others' goals is considered the 'right' thing to do, whether it be to punish bad behavior, keep vulnerable individuals safe, be loyal to one's allies, and so forth. Thus, a core system for moral goodness should allow for goal-preventing behaviors, and a core moral evaluation tendency should excuse, or even celebrate, those who perform justified goal-blocking actions intentionally. (p191)

... In the case study we'll consider, Hamlin and colleagues find that, before the age of 2, toddlers will take resources from a hinderer but they will not take resources from a helper (Hamlin, Wynn, Bloom, & Mahajan, 2011).

As we have seen...

Case study in moral retribution

Infants in their first year of life will approach individuals who have acted positively toward others and avoid those who have acted negatively. Infants also expect others to respond in this manner—to approach those who have helped them and avoid those who have harmed them. ... Infants' understanding might extend beyond these simple heuristics, however. For one thing, the intentions of an agent may inform even our earliest judgments." (p19931)

... Consider the pattern with adults: "Under at least some circumstances, [adults] are motivated to approach individuals who have been intentionally harmful in the past, to punish them. [...] In other instances, however, adults approach harmful individuals not to punish them, but because the harmful acts and individuals have been positively evaluated. ... First, harms directed toward those who are themselves wrongdoers may be evaluated positively, as deserved acts of punishment. ... Alternatively, attraction toward those who harm others might be based on social alliances. [...] For adults[,] then, social evaluation often goes beyond analyzing the immediate, local valence of a behavior enacted toward another. We consider our previous evaluations of [someone] to determine the overall, what we might call [] global valence of the action" (p19931)

Hamlin et al (2011). "How infants and toddlers react to antisocial others". *PNAS, 108(50)*, pp19931–19936.

134 / 200

Case study in moral retribution

In Hamlin's (and others') work, we learn that babies prefer a ball-giving puppet to one that takes a ball away.

- 5 month olds always liked a "ball-giver" (over a "ball-taker"), no matter whether the puppet that had dropped the ball had been mean or helpful in the previous scene; 5 month olds simply like puppets to be nice in the moment **("local" positive valence)**
- 8 month olds only liked the "ball-giver" when they gave the ball back to a previously helpful puppet, otherwise they preferred the "ball-taker"; 8 month olds liked puppets who appropriately rewarded/punished based on past behavior **("global" positive valence)**

Would that behavior translate into the infants' own behavior, in the role of rewarder/punisher?

... "The researchers [followed-up] with 32 toddlers ages 19 months to 23 months, this time adding a twist. The toddlers got to watch puppets being nice or mean to each other and then got to play the role of rewarder or punisher. Some toddlers were shown one nice puppet and one mean puppet and asked which they'd like to share a treat with. Others were shown a nice puppet and a mean puppet, both with treats, and were asked to take a treat away from one."
https://www.livescience.com/17204-babies-prefer-punishment.html

"Nice" and "mean", here, were established by a prior series of 3 vignettes like that of the "box slamming" or opening/"blocking/facilitating access" scenes that we've seen before.

In the puppet rewarder/punisher study the FOAS paper: https://www.livescience.com/17204-babies-prefer-punishment.html, "Even Babies Think Crime Deserves Punishment".

Case study in moral retribution

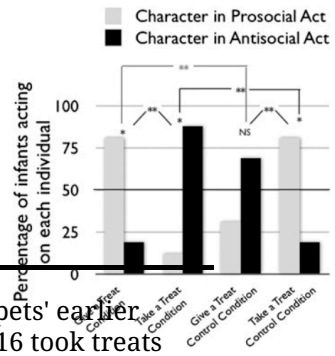


Fig. 2. Results in experiments 4 and 5. Percentage of toddlers directing Giving or Taking actions toward characters, who performed a Prosocial or an Antisocial act, or toward characters who received a Prosocial or an Antisocial act is shown. Asterisk, Binomial test, one-tailed $P < 0.05$; double asterisk, Fisher's Exact test, $P < 0.05$.

"In all cases, the toddlers meted out justice according to the puppets' earlier actions. Thirteen of 16 gave a treat to a nice puppet, while 14 of 16 took treats away from a mean puppet." https://www.livescience.com/17204-babies-prefer-punishment.html

"Give a Treat" and "Take a Treat" on the left -- gave or took from the character who PERFORMED the prosocial or antisocial act.

"Give a Treat (control)" and "Take a Treat (control)" on the right -- gave or took from the character who RECEIVED the prosocial or antisocial act.

What mechanism(s) of mind could produce this?

Why do infants prefer those who harm Hinderers?

- Do they believe that Hinderers **deserve** punishment?
- Do they believe that **the enemy of their enemy** is a friend?

Hamlin leaves us with the question of mechanisms - the cognitive processes that underwrite these kinds of preferences from a very young age.

What mechanism(s) of mind could produce this?

Why do infants prefer those who harm Hinderers?

- Do they believe that Hinderers **deserve** punishment?
- Do they believe that **the enemy of their enemy** is a friend?

Some reasons given to prefer the "enemy" mechanism:

- "It can be applied to anyone who is likely to cause harm, including Hinderers, but also to outgroup members, friends of enemies, and so forth"
- infants also prefer those who hinder (vs. help) puppets who are different from them

Last word

these early tendencies are far from shallow, mechanical predispositions to behave well or knee-jerk reactions to particular states of the world: Infants' moral inclinations are sophisticated, flexible, and surprisingly consistent with adults' moral inclinations, incorporating aspects of moral goodness, evaluation, and retaliation. This research... suggests that morality is a core aspect of human nature.

...Where here she understands "human nature" as "what we're born with", as nativists would have it.

Bridge: essentializing and generalizing

"infants also prefer those who hinder (vs. help) puppets who are different from them" !!!

At the end of the Hamlin paper she raises a certain interaction effect in early moral cognition that relates to a large and growing body of work suggesting important "ingroup" and "outgroup" mechanisms, which will play a prominent role in the last topic of this module.

That is, the infant data we've looked at suggests that humans, from very early on, are able to determine inner motivations, and are drawn to individuals that show more virtuous behavior.

But it ain't all roses...

Bridge: essentializing and generalizing

Seeing an individual engage in negative behavior (say, the **hinderer** from Hamlin et al 2007) can lead us to **generalize** from that behavior to an expectation about a salient social group of which that individual is a member.

Bridge: essentializing and generalizing

Seeing an individual engage in negative behavior (say, the **hinderer** from Hamlin et al 2007) can lead us to **generalize** from that behavior to an expectation about a salient social group of which that individual is a member.

Sarah Jane Leslie's original sin of cognition:

- we **essentialize** groups of people (apparently we can't help ourselves!)
- we **generalize** from individual members of those groups to the group as a whole

Bridge: essentializing and generalizing

Seeing an individual engage in negative behavior (say, the **hinderer** from Hamlin et al 2007) can lead us to **generalize** from that behavior to an expectation about a salient social group of which that individual is a member.

Sarah Jane Leslie's original sin of cognition:

- we **essentialize** groups of people (apparently we can't help ourselves!)
- we **generalize** from individual members of those groups to the group as a whole
- ... but just when the behavior is "strikingly negative"!

... According to Leslie, this cognitive bias is an important part of the story of prejudice and bigotry in human societies, and thus is quite relevant for us in this time.

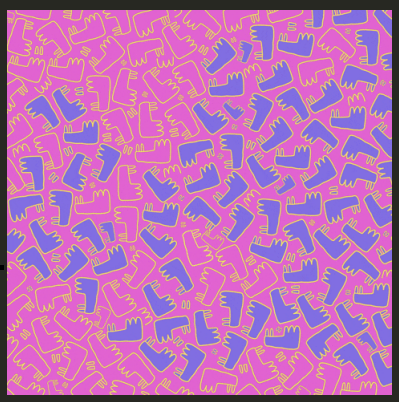
That's where we go next. Read Sarah Jane's paper, or at least pages 393-412, for next Tuesday.

PHIL246: Foundations of Cognitive Science

Instructors Toby Mintz & Alexis Wellwood

TAs Megha Devraj, Rachel Keith & Nadja Winning

5 The conscious, social, moral, essentialist Day 4



144 / 200

Following primarily Sarah Jane Leslie, "The original sin of cognition"

GIF Source: 5-essentialism.gif Source: http://gelheureux.com/

Agenda

1. Consciousness
2. Sociality
3. Morality
4. Essentialism

A contemporary strand of psychological theorizing concerns "essentialist reasoning" -- the human propensity for positing that things have ESSENCES which explain why they are the way they are -- and the CONSEQUENCES of "essentializing" groups of people.

We're discussing the Sarah Jane Leslie paper which concerns a corner of research into essentialism that intersects with moral considerations - it concerns how we tend to think about *groups* of people, and how we might be inclined to think about *individuals* in those groups as a consequence solely of their perceived group-membership.

We might put the central tension of Leslie's discussion in the context of this class, as follows:

If people are, factually, very much made of the same stuff (biologically, mentally, socially, and morally),

and if all people are, thereby, equally capable of virtuous behavior,

and if we are born recognizing (and being attracted to) virtuous behavior,

then WHY do we often assume that *some people* are inherently less virtuous than we are?

Leslie's paper aims to give us some sense of an answer to this question, based in factual considerations of how we apparently extend reasoning about NATURAL or BIOLOGICAL kinds of things, sort of PATHOLOGICALLY to "SOCIAL KINDS" of things.

[References are to Leslie 2017 unless otherwise specified.]

Agenda

1. Consciousness
2. Sociality
3. Morality
4. Essentialism
   - Generics
   - Generalizing about kinds
   - Essences & essentialized kinds

In this lecture, we'll first focus on Leslie's discussion of the connection between so-called "generics" or generic statements and essentialist thinking. Then, we'll talk about the link between thinking of things as kinds, and essentializing inferences. Finally, we'll review what Leslie thinks has gone wrong: we've taken thought of essences out of its ORIGINAL or PROPER domain, into a new domain to which it doesn't quite fit.

Let's begin, then, with what Leslie means by "the original sin", she writes:

"The original sin of cognition"

Long before we learn to talk, our expectations concerning novel members of a category are shaped by our experience with already-encountered members.

... Here, the ORIGINAL part of "the original sin" is meant to evoke the idea that the relevant sorts of reasoning arise early on in human development, and is reasoning that is about CATEGORIES, or things falling under specific CONCEPTS.

Leslie points out that early cognitive development research supports early expectations like the following:

"The original sin of cognition"

Long before we learn to talk, our expectations concerning novel members of a category are shaped by our experience with already-encountered members. We expect, for example, that objects that share obvious perceptible qualities will also share dispositional properties. If a given item rattles when shaken, nine-month-olds expect that other items that share the same perceptible profile will rattle when shaken.

... This is an example of the ORIGINAL domain for the relevant type of reasoning -- in reasoning about artifactual or biological kinds. It seems reasonable, to be sure, to assume that if one object (like a toy, or a tool) shares perceptible qualities with another object, then those two may very well share "dispositional properties" -- roughly, properties indicating that the two things should behave similarly under similar circumstances.

But it goes further than mere perceptible qualities, and fast:

"The original sin of cognition"

Long before we learn to talk, our expectations concerning novel members of a category are shaped by our experience with already-encountered members. We expect, for example, that objects that share obvious perceptible qualities will also share dispositional properties. If a given item rattles when shaken, nine-month-olds expect that other items that share the same perceptible profile will rattle when shaken. By our first birthday, these inductive inferences are guided by language; we expect that even superficially dissimilar objects will share their hidden properties if they are identified by the same common noun—if, for example, each is introduced as 'a blicket'.

... here, "inductive inference" is just a fancy way of saying that one infers, from specific cases (that is, finite experience), some general principle that applies to boundlessly many potential cases.

This pattern of early inference, Leslie suggests, indicates that

"The original sin of cognition"

There is, presumably, some early developing cognitive mechanism that is responsible for these infant inductive generalizations.

... With this, Leslie formulates a **hypothesis**: given the question, why do infants make the inductive inferences that they do? She believes that those inferences are the result of some underlying causal mechanism - some mechanism of mind that DISPOSES or COMPELS the infant to make inferences just like this.

She furthermore suggests a way in which LANGUAGE may be a "window into the mind", revealing the inner aspects of this mechanism:

"The original sin of cognition"

There is, presumably, some early developing cognitive mechanism that is responsible for these infant inductive generalizations. In earlier papers, I argued that generics—sentences such as 'ravens are black' and 'tigers are striped'—express the generalizations that are delivered by this cognitive mechanism. If this is so, then generics provide a window onto the workings of the mechanism.

... With this, Leslie refers to a secondary hypothesis: given a question like, how can we peer into the hypothesized mechanism? She proposes that we can do so through careful consideration of the cognitive signficance of particular linguistic forms that are often called GENERICS - sentences of the form "Ravens are black".

Now,

"The original sin of cognition"

There is, presumably, some early developing cognitive mechanism that is responsible for these infant inductive generalizations. In earlier papers, I argued that generics—sentences such as 'ravens are black' and 'tigers are striped'—express the generalizations that are delivered by this cognitive mechanism. If this is so, then generics provide a window onto the workings of the mechanism. In this paper, I am concerned with a particular aspect of this mechanism, namely, the route by which we reach general conclusions regarding dangerous or harmful features.

Her study in the present paper, then, building on her earlier work, is on the influence of the TYPE of feature that may be attributed in a generic sentence -- namely, striking or dangerous ones -- and what role that TYPE plays in one's disposition to form one of these cognitively basic inductive generalizations in the first place.

These pieces - generic language, and the specific role of dangerous or harmful features there expressed - will primarily occupy our attention today.

But of course, in the end, all of this is put to a further and indeed loftier goal:

"The original sin of cognition"

There is, presumably, some early developing cognitive mechanism that is responsible for these infant inductive generalizations. In earlier papers, I argued that generics—sentences such as 'ravens are black' and 'tigers are striped'—express the generalizations that are delivered by this cognitive mechanism. If this is so, then generics provide a window onto the workings of the mechanism. In this paper, I am concerned with a particular aspect of this mechanism, namely, the route by which we reach general conclusions regarding dangerous or harmful features. My aim is to identify and discuss a cognitive bias that has contributed to certain virulent forms of prejudice.

... Leslie believes that there is an important connection between our early cognitive biases, and the language that gives them expression, to the propagation of PREJUDICE -- dislike, hostility, or unjust behavior deriving from unfounded opinions -- and some PRESCRIPTIONS for mitigating its spread.

So, why "the original sin"?

A theme that emerges from the paper is that...

- there may be good reasons that evolution has bestowed upon us an early-emerging inductive inference mechanism like that Leslie describes (**original domain**)
- but we also seem to be wired to extend the application of that mechanism in ways that are plainly inappropriate/harmful (**actual domain**)

... So to answer the question, and to play out the Christian metaphor: one of those people in the Garden of Eden may have had good reasons to take a bit of the apple, and invite her friend to do so; however, so the story goes, the implications of this action were in fact very negative.

That's my take anyway.

Now, Leslie uses the platform of this paper to offer some indications of how study in this vein can be used to COMBAT prejudice, with some important caveats:

From the conclusion

The perspective offered here is not intended as the final word on [the matter of prejudice]—nothing could be further from my intent. My aim has rather been to take a small slice of a huge phenomenon and identify some of the contributing cognitive factors. Identifying these contributing cognitive factors leads us to pose the question of how best to alter those factors and thereby combat those aspects of prejudice that arise from "the original sin" of cognition, namely, its primitive tendency to generalize strikingly negative information across the members of highly essentialized kinds.

... Now, as I indicated to you before, we will not be discussing her prescriptions (and I have left those sections of the paper open as "optional" reading); instead, we will spend our time working to understand the last point -- that this mechanism is sensitive to the degree of essentialization of a group, and sensitive to the TYPE of property instantiated by a small number of its members.

Generics

We'll begin with generics

156 / 200

Generic sentences

Basketball players are tall.
A mammal has fur.
Mosquitoes carry the West Nile virus.
Tigers eat people.

These are the type of sentences that Leslie suggests express "cognitive fundamental" generalizations.

It is widely agreed that such sentences are the primary vehicle by which we communicate information about *kinds of things*.

Generic sentences

Form

- Typically **Bare plural count noun** + **present tense predicate**

  Basketball players are tall.

- Also some **Singular indefinite count noun** + **present tense predicate**

  A mammal has fur.

Generic sentences

Form

- Typically **Bare plural noun** + **present tense predicate**
- Also some **Singular indefinite noun** + **present tense predicate**

Meaning

- Some authors: generics just express that *most* or *all* instances of the category have the property

Most basketball players are tall.

All mammals have fur.

- Slim proposal: just that the members of the noun category are **disposed** to have the property expressed by the predicate

Being a mammal disposes one to have fur.

... These claims are HYPOTHESES about what the relevant class of sentences mean. Leslie's intention is that understanding the MEANING of "A mammal has fur" involves thinking about DISPOSITIONS of mammals. This is not the same as a meaning like "ALL mammals have fur"; for such an "explicitly quantified" sentence, it could very well be that the property holds of mammals by accident.

Why does she want to talk about dispositions instead of "how many" instances of the category have the property?

Well, the proposal that the sentences are equivalent to *most* or *all* sentences is just too easy to falsify:

Dispositions vs quantity

['Mosquitoes carry the West Nile virus' is] intuitively true, even though very few members of the kind in question possess the predicated property. As it happens, less than one percent of mosquitoes carry the West Nile virus, and yet we are quick to assent to 'mosquitoes carry the West Nile virus', even after learning this statistical fact. (Conversely, 'mosquitoes do not carry the West Nile virus' remains patently false, even though 99% of mosquitoes do not carry the virus.) (395)

... Let's probe this a bit further, looping in the particular TYPE of predicate that Leslie is honing in on; and we can ask:

Striking property generalizations

How many members of the class have to have the property for us to suppose the generic statement is true?

Basketball players are tall.
A mammal has fur.
Mosquitoes carry the West Nile virus.
Tigers eat people.

Striking property generalizations

How many members of the class have to have the property for us to suppose the generic statement is true?

Basketball players are tall.
A mammal has fur.
Mosquitoes carry the West Nile virus.
Tigers eat people.

The answer is: **it depends** on the sort of property

- **positive** or **neutral**
  - generic sentences with this type of predicate are considered true only on the basis of pretty solid evidence that an arbitrary member of the group has the property
  - i.e., the property occurring with **high prevalence** in the group correlates with high acceptance of the relevant generic
- **striking** or **dangerous**
  - generic sentences with this type of predicate can be considered true on the basis of very little evidence that any arbitrary member of the group has the property
  - i.e., the property occurring with **low prevalence** in the group does not preclude high acceptance of the relevant generic

162 / 200

...

For example, the facts surrounding the sentences at the top: it is probably the case that close to 100% of basketball players are tall, and close to 100% of mammals have fur. For the mosquitoes sentence, it's less than 1%. For the last one I'm not sure (and I'm also not sure that I'd endorse it...)

Dispositions (not *how many*)

If even just a few members of the kind possess a property that is harmful or dangerous, then a generic that attributes that property to the kind is likely to be judged true.

Mosquitoes carry West Nile virus.
→ `Mosquitoes are disposed to carry West Nile virus'

Appeal to **dispositions** "provide[s] for the possibility of some further specific state or behaviour"; the idea that generic sentences attribute dispositions to a class is that they tell you about a *latent property* that members of the class share, even if it's not realized by (many or even most) specific individuals.

... Accepting a generic, then, on Leslie's view, does not really depend on knowing anything about "how many" members of a given class have the specified property.

- Generics do not tell you that you can reasonably expect, for an arbitrary member of the class, whether they have the property.
- Instead they just tell you that *being a member of the class* **disposes** one to have the property.

As with the Hamlin paper, we are treated to some evolutionary psychology to suggest how this might have arisen:

Why?

Since we are working under the hypothesis that generics give voice to psychologically fundamental generalizations, [our observations imply] that our basic way of dealing with dangerous or harmful information involves rapidly generalizing this information to the salient kind or category.

Why? Evolutionary psychology...

Since we are working under the hypothesis that generics give voice to psychologically fundamental generalizations, [our observations imply] that our basic way of dealing with dangerous or harmful information involves rapidly generalizing this information to the salient kind or category. We do not wait around to see what percentage of tigers eat people before drawing a general conclusion—even a single instance may be enough for us to conclude that tigers eat people. It is not hard to imagine the evolutionary benefits... since the cost of undergeneralizing such information are potentially huge.

... in other words, if we see our friend getting attacked by a tiger, it would not be advantageous for us to conclude, "WE'D BETTER AVOID THAT BEAST IN THE FUTURE!"

Instead, we conclude, "WE'D BETTER AVOID *THAT KIND OF THING* IN THE FUTURE!" -- we **assume**, inductively, that other things which are just like the thing attacking our friend, is DISPOSED TO ATTACK PEOPLE and so are better avoided.

Why? Evolutionary psychology...

Since we are working under the hypothesis that generics give voice to psychologically fundamental generalizations, [our observations imply] that our basic way of dealing with dangerous or harmful information involves rapidly generalizing this information to the salient kind or category. We do not wait around to see what percentage of tigers eat people before drawing a general conclusion—even a single instance may be enough for us to conclude that tigers eat people. It is not hard to imagine the evolutionary benefits... since the cost of undergeneralizing such information are potentially huge. Our ancestors were far better off jumping to conclusions, as it were, than taking the time to judiciously determine the precise likelihood of their being eaten. (396)

...

And so, what we have is a suggestion from evolutionary advantage, that there is a "proper domain" of application for this cognitive mechanism -- for one that infers from a specific instance of an individual with a strikingly negative property, TO the dispositions of the group of which that individual is a member.

To say that this is the "ORIGINAL or PROPER" domain is to suggest that, indeed, there are areas of life or periods of time in which it might be beneficial to draw such inductive inferences.

But again, it ain't all roses. To appreciate this in full, let us talk about how Leslie thinks about kinds.

First, questions?

Generalizing about *kinds*

Leslie's theory

One basic way that we represent things in the world is in terms of *kinds K* and their *properties F*.

- Generic statements of the form '*K*s are *F*' don't express that all, some, or even most of the members of a kind *K* are *F*.
- Rather, they express that *there is something about K*s that *dispose* them towards being *F*...

Leslie's theory

One basic way that we represent things in the world is in terms of *kinds K* and their *properties F*.

- Generic statements of the form '*K*s are *F*' don't express that all, some, or even most of the members of a kind *K* are *F*.
- Rather, they express that *there is something about K*s that *dispose* them towards being *F*...

When we observe an individual member of *K* display a property *F* that is *dangerous* or potentially *harmful*, we're predisposed to attribute *F* to *K* itself, **because, evolutionarily, that was the safest thing for us to do**

Leslie's theory

One basic way that we represent things in the world is in terms of *kinds K* and their *properties F*.

- Generic statements of the form '*K*s are *F*' don't express that all, some, or even most of the members of a kind *K* are *F*.
- Rather, they express that *there is something about K*s that *dispose* them towards being *F*...

When we observe an individual member of *K* display a property *F* that is *dangerous* or potentially *harmful*, we're predisposed to attribute *F* to *K* itself, **because, evolutionarily, that was the safest thing for us to do**

A partial "**parts list**" for the early-emerging cognitive mechanism, then:

- ~~representing things in terms of *K* membership~~
- recognizing *F* as *striking/dangerous* or not
- upon observing a member of *K* that is *F*,
  infer that *there is something about* K*s that disposes them to be* F *(later: a* K-**essence**)

Generalizations, dispositions, and predictors

Leslie's claim: The same cognitive patterning is responsible for a mind that endorses both (1) and (2); but the generics differ in the **sort** of kind they're generalizing about.

(1) Mosquitoes carry the West Nile Virus. ⇝ BIOLOGICAL KIND
(2) Muslims are terrorists. ⇝ SOCIAL KIND

... Now, what can we say about (1) and (2)? They are both generics; they are both KIND-involving; they both have strikingly negative or harmful predicates; and they are about different SORTS of kinds.

So, Leslie wants to say, this last factor affects their TRUTH VALUE:

Generalizations, dispositions, and predictors

Leslie's claim: The same cognitive patterning is responsible for a mind that endorses both (1) and (2); but the generics differ in the **sort** of kind they're generalizing about.

(1) Mosquitoes carry the West Nile Virus. ⇝ BIOLOGICAL KIND
(2) Muslims are terrorists. ⇝ SOCIAL KIND

Even if the prevalence rates were the same (>1%), we can accept (1) as true, but reject (2) as false:

- (1) says something TRUE about mosquitoes - unlike other insect species, mosquitoes are indeed disposed to carry WNV
- (2) says something FALSE about Muslims - Muslim people are not any more disposed to be terrorists than any other group of people

[Leslie footnotes for discussion of dispositions: M. Fara "Dispositions and habituals" (2005), and D. Lewis, "Finkish dispositions", (1997).]

What striking generic sentences mean

This pattern makes sense, if you think striking generic sentences have the following truth conditions:

'*K*s are F_{striking}' is true iff some members of *K* are *F*, and members of *K* are disposed to be *F*

What striking generic sentences mean

This pattern makes sense, if you think striking generic sentences have the following truth conditions:

'*K*s are $F_{striking}$' is true iff some members of *K* are *F*, and members of *K* are disposed to be *F*

For Leslie, part of what makes a member of *K* disposed to be *F* is for *F* to be a good predictor of *K*-hood.

- For *F* to be a **good predictor** of *K* requires "a kind that is inclusive enough to aid us in avoiding the property, but not so inclusive as to needlessly hamper our activities"

---

re: "a kind that is inclusive enough", she writes- "it is easy enough to see an evolutionary rationale behind generalizing striking properties only so far up the taxonomic hierarchy. If our ancestors had undertaken to avoid all mammals after seeing a tiger eating one of their companions, the costs of doing so may very well have outweighed the benefits. (One could waste a lot of time running from small, harmless creatures.) Someone who avoided all animals, big or small, after witnessing a lion maul his companion would be at a significant disadvantage relative to a more sophisticated competitor who limited his conclusions to lions alone."

The reason people can DISAGREE about whether the relevant generic statement is true is because DISPOSITIONS, like mental states, are not directly observable. !!

What striking generic sentences do mean

'*K*s are F_{striking}' is true iff some members of *K* are *F*, and members of *K* are disposed to be *F*

What makes *K* disposed to be *F* is for *F* to be a good predictor of *K*-hood.

- For *F* to be a **good predictor** of *K* requires that *K* "is inclusive enough to aid us in avoiding the property, but not so inclusive as to needlessly hamper our activities"
- To assume that members of *K* are **disposed** to be F requires a tacit belief about the shared natures and dispositions of members of certain kinds — for example, it requires **essentialized** kinds

---

So, let us talk about essences and essentialized kinds.

Essences & essentialized kinds

Essentialized kinds

Even without relevant scientific knowledge, we have **tacit** beliefs that certain kinds of things have shared natures and dispositions

- A kind is **essentialized** whenever we have "the (tacit) belief that there is some hidden, nonobvious, and persistent property or underlying nature shared by members of that kind that causally grounds their common properties and dispositions."

Essentialized kinds

Even without relevant scientific knowledge, we have **tacit** beliefs that certain kinds of things have shared natures and dispositions

- A kind is **essentialized** whenever we have "the (tacit) belief that there is some hidden, nonobvious, and persistent property or underlying nature shared by members of that kind that causally grounds their common properties and dispositions."
- "There's just something about Ks..."

Essentialized kinds

Even without relevant scientific knowledge, we have **tacit** beliefs that certain kinds of things have shared natures and dispositions

- A kind is **essentialized** whenever we have "the (tacit) belief that there is some hidden, nonobvious, and persistent property or underlying nature shared by members of that kind that causally grounds their common properties and dispositions."

Understanding an essence to **causally ground** *K* means believing that an essence plays an important **explanatory role** in Ks being what they are, in having the appearance they have, or engaging in the behaviors they engage in, etc.

Essentialized kinds

Even without relevant scientific knowledge, we have **tacit** beliefs that certain kinds of things have shared natures and dispositions

- A kind is **essentialized** whenever we have "the (tacit) belief that there is some hidden, nonobvious, and persistent property or underlying nature shared by members of that kind that causally grounds their common properties and dispositions."

Understanding an essence to **causally ground** *K* means believing that an essence plays an important **explanatory role** in Ks being what they are, in having the appearance they have, or engaging in the behaviors they engage in, etc.

- A tiger is a kind of thing that, because it is the kind of thing it is, has stripes, has a long tail, lives in the jungle, eats people ...
- ... but none of these features are not criterial for membership in *K*: a tiger can lack stripes, lack a tail, live in the zoo, refrain from certain dietary habits ... while still being a tiger.

... Now, fascinatingly, even without tapping into talk of social kinds we can see that there are differences in how we understand what is possible for different kinds of things:

Essences: biological kinds vs artifacts*

The assumptions that children make about the causal role of essences can differ depending on the category, as for example when thinking about natural kinds like *water*, biological kinds like *animals*, or artifactual kinds like *wrench*.

Essences: biological kinds vs artifacts*

Walker (1999) tested under what circumstances children in North America and Nigeria would believe that changes to the appearance of things would fundamentally alter their identity: "rural Yoruba children..., like North American children, believe that superficial changes transform the identity of artifacts, but not living kinds"

Key/nail. Dele, a blacksmith, took this key and heated the metal so that it became very soft. Then he carefully rolled it, made the end very sharp and made the top flat. When he finished, was it a key or a nail?

Spider/scorpion. A boy took a spider, cut off two of its legs, and added a long, curved tail and two claws. In the tip of the tail he added a bit of poison. Now the spider looks like this. Do you think it's a spider or a scorpion?

What would you judge? Let's go through them.

... Now, with particular respect to biological kinds:

* quoted in W. Ahn et al. (2001), "Why essences are essential in the psychology of concepts"; stimuli from Walker

Essences: biological kinds*

"people (including children) have some, albeit not complete, ideas about the nature of the mechanism mediating between category identity and properties" (60) - they "expect innate potential to prevail over competing environmental influences" (61)

Essences: biological kinds*

"people (including children) have some, albeit not complete, ideas about the nature of the mechanism mediating between category identity and properties" (60) - they "expect innate potential to prevail over competing environmental influences" (61)

Gelman and Wellman (1991): (a) insides matter, (b) independently of language

- (a) "Children who were told, for example, about a baby cow raised with pigs typically reported that the cow would nonetheless display cow-like properties (mooing, having a straight tail) as an adult" (60)
- (b) "children were told, for example, that a seed was taken out of an apple and planted with flowers in a flowerpot. When asked what would pop out of the ground, 4-year-olds typically predicted that an apple tree, not flowers, would emerge, even when the seed was never described as either an apple tree or an apple seed"

... According to Leslie, considering matters of biological and natural kinds is the "original domain" for the cognitive mechanism that generates the relevant inductive inferences about kinds.

The problem arises when that reasoning applies in cases where there is no biological or natural kind - that is, with social kinds.

* Wolfe et al. (2003): "Why does a natural kind matter in the psychology of concepts"

"Original domain" vs "actual domain"

Original domain

Leslie's "cognitively fundamental generalizations" that we express using generics likely evolved in the context of reasoning about the biological world; it was *for* tracking the properties of biological and natural kinds.

"Original domain" vs "actual domain"

Original domain

Leslie's "cognitively fundamental generalizations" that we express using generics likely evolved in the context of reasoning about the biological world; it was *for* tracking the properties of biological and natural kinds.

Actual domain

Leslie thinks that part of the story of prejudice is that we extend this basic form of reasoning and representation to categories of people—**social kinds**.

"Original domain" vs "actual domain"

Original domain: Leslie's "cognitively fundamental generalizations" that we express using generics likely evolved in the context of reasoning about the biological world; it was *for* tracking the properties of biological and natural kinds.

Actual domain: Leslie thinks that part of the story of prejudice is that we extend this basic form of reasoning and representation to categories of people—**social kinds**.

The basic idea is simple: just as it takes but a few instances of sharks attacking bathers, or of mosquitoes carrying the West Nile virus, for us to make the corresponding category-wide generalization, so also a strikingly negative action of a few members of a racial, ethnic, or religious minority may lead others to a general belief concerning their entire group. (Leslie 2017, 399)

... Moreover, "If reasoning of this sort really is a pervasive cognitive disposition, then we should find many examples of it in whatever historical period we examine" (Leslie 2017, p400), and she gives several historical examples that buttress the supposition.

Now, about social kinds:

Social kinds and essence

Recently social psychologists have argued that we view certain social kinds as essentialized in much the same way as we do biological kinds.

- "Such social kinds, we may suppose, will surely include racial, ethnic, and religious groups" (410)

Social kinds and essence

Recently social psychologists have argued that we view certain social kinds as essentialized in much the same way as we do biological kinds.

- "Such social kinds, we may suppose, will surely include racial, ethnic, and religious groups" (410)
- (Just in case:) Racial kinds **are not biological**; there is no genetic ground for dividing people along racial lines

Racial != biological kinds

A young woman's tooth found in Russia revealed a new species of humans—the Denisovans.

PHOTOGRAPH BY ROBERT CLARK, NAT GEO IMAGE COLLECTION

In many ways, genetics makes a mockery of race. The characteristics of normal human variation we use to determine broad social categories of race—such as black, Asian, or white—are mostly things like skin color, morphological features, or hair texture, and those are all biologically encoded.

But when we look at the full genomes from people all over the world, those differences represent a tiny fraction of the differences between people. There is, for instance, more genetic diversity *within* Africa than in the rest of the world put together. If you take a Yoruba from Nigeria and someone from the Sudan, they are more likely to be *more* genetically different from each other than either one of those people is to anyone else on the planet!

... still, people posit essences, in many cases there needn't even be a supposition of a shared biological nature:

[Source: "Why race is not a thing, according to genetics", in the National Geographic - https://www.nationalgeographic.com/science/article/genetics-history-race-neanderthal-rutherford#:~:text=In%20many%20ways%2C%20genetics%20makes,those%20are%20all%20biologically%20encoded.]

Social kinds and essence

Recently social psychologists have argued that we view certain social kinds as essentialized in much the same way as we do biological kinds.

- "Such social kinds, we may suppose, will surely include racial, ethnic, and religious groups" (410)
- (Just in case:) Racial kinds **are not biological**; there is no genetic ground for dividing people along racial lines

consider attitudes towards Muslims—a group that is rarely if ever taken to be *biologically* defined. Here, the absence of belief in a biological essence does not preclude belief in an essence more broadly understood, as is illustrated by persisting beliefs that Muslims are *fundamentally different* from other people, while also being *fundamentally all alike*. (411)

And there are consequences

(on "excusing" individuals that dont have the properties; see Anthony Appiah, "Race, Culture, Identity: Misunderstood connections", 1994: https://philpapers.org/archive/apprci.pdf)

Ks are terrorists

"After the events of September 11th, 2001, hate crimes against Muslims rose **more than 1600%**".

A hate crime is defined as one which is motivated not by personal animus, but by the desire to harm members of a particular group BECAUSE they are members of that group. Such crimes carry a heavier penalty because of the sense of TERROR they impose on all members of that group.

Ks are terrorists

"After the events of September 11th, 2001, hate crimes against Muslims rose **more than 1600%**".

Many of these crimes were committed against Muslim children; the perpetrators surely were not under the impression that their victims were themselves involved in or personally responsible for the 9/11 bombings. It was sufficient that the victims appeared to be Muslims. (Leslie 2017, 400)

...

Now, an important point to notice here is that we, individuals, essentialize OTHERS, not ourselves. Those who would endorse "Ks are terrorists" isn't likely a member of that K.

Check it: do you think members of your racial, ethnic, or religious group are all alike?

Or, if you happen to be similar to the sorts of individuals who typically carry out mass shootings in America, ask yourself: would you endorse the generic statement, "White boys/men are terrorists"? No!

Back to the original domain

The psychologically privileged level at which an essence is taken to ground a wide range of properties is known as the "basic level".

- Given a taxonomic hierarchy like: *Bengal tigers, Siberian tigers* > *Tigers* > *Mammals*...
- ... *Tigers* lie the "basic level"

Different levels of such taxonomic hierarchies are essentialized to different degrees.

- Some are seen as possessing highly distinctive essences that ground a large number of shared features (Tiger), and others not (Mammal, Bengal Tiger)

Back to the original domain

The psychologically privileged level at which an essence is taken to ground a wide range of properties is known as the "basic level".

- Given a taxonomic hierarchy like: *Bengal tigers, Siberian tigers* > *Tigers* > *Mammals*...
- ... *Tigers* lie the "basic level"

Different levels of such taxonomic hierarchies are essentialized to different degrees.

- Some are seen as possessing highly distinctive essences that ground a large number of shared features (Tiger), and others not (Mammal, Bengal Tiger)

~~For Leslie, this is because the basic level is more predictive o~~f more distinctive properties:

- Mammals include creatures as diverse as tigers, whales, mice, and humans, yet these sub-kinds differ widely in their properties in important ways (cp. whales and mice)
- Conversely, Bengal tigers share many properties in common with each other, but this shared nature doesn't usefully distinguish them from Siberian tigers

Leslie's understanding of the actual domain

A kind or group is **essentialized** just in case its members are viewed as sharing a fundamental nature that causally grounds a substantial number of their outwardly observable properties.

- This nature needn't be biologically grounded, nor immutable, nor strictly necessary.

The most likely social kinds to be essentialized are at the "basic level"

- A "basic-level" social kind *K* is most likely to be attributed such an essence; such a *K* will be distinctive (not too large a group) and usefully predictive (not too small a group)
- The perception that *K* membership is distinctive and predictive produces willingness to generalize across the group, but not more inclusively (e.g., not up to the category *religious person* or *human being*).

...

Thus, such *K* groups are the typical targets of striking property generalizations.

Last words

The perception that members of a given group share highly predictive natures supports a willingness to generalize across the group, and the perception that these natures are distinctive produces reluctance to generalize to more inclusive categories—to the category human being, for example. As with animal kinds, these highly essentialized social kinds are the typical targets of striking property generalizations. (p393)

Last words

The perception that members of a given group share highly predictive natures supports a willingness to generalize across the group, and the perception that these natures are distinctive produces reluctance to generalize to more inclusive categories—to the category human being, for example. As with animal kinds, these highly essentialized social kinds are the typical targets of striking property generalizations. (p393)

The intriguing upshot is that our very choice of words to describe racial, ethnic, and religious kinds may subtly communicate to children that these kinds are to be essentialized. We need not say anything negative about these groups—the use of generics or even simply labels may communicate that these are essentializable groups, and so open the door to prejudice. (p418)

... Knowing what we know now, all of us should be on guard, in our thought and in our talk, for essentialist reasoning, and over-generalization.

And, this concludes our "grab-bag module", in which we started out by asking:

what is it like to be human?

and rephrased to focus on some more specific properties we instantiate, asking:

199 / 200

what is it like to be human?

what is it like to be a conscious, social, moral, essentialist?

Thursay we start on a 3-day lecture on MINDS & BRAINS, with Professor Mintz;
your Short assignment 4 and optional Extra credit work is due Friday!

200 / 200