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

What is it to be a human?

2 / 200

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

3 / 200

conscious
social
moral
essentialist

4 / 200

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.*

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

conscious
social
moral
essentialist

6 / 200

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.*

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

conscious
social
moral
essentialist

8 / 200

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 come from?

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

9 / 200

conscious
social
moral
essentialist

10 / 200

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.

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

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)

Agenda: Module 5

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

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?

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.

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.

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'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 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?

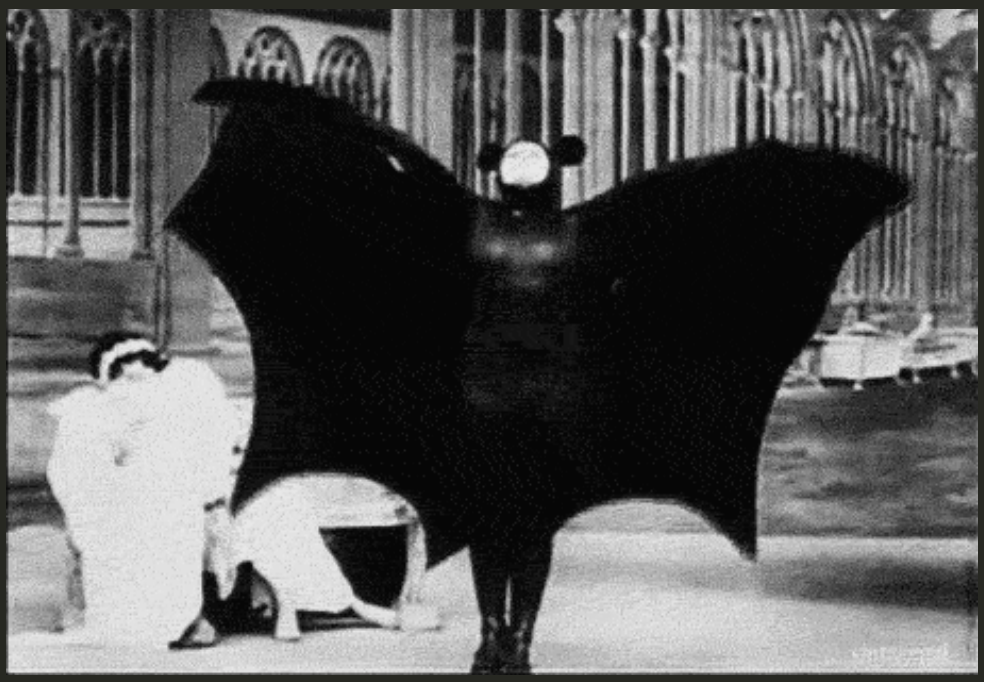
"Facts about what it is like to be an *X* are very peculiar, so peculiar that some may be inclined to doubt their reality, or the significance of claims about them." (437)

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*.



23 / 200

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)

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.

The "easy" problems

33 / 200

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

"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 were used to tickle with... choose the structure in which *with a red tailfeather* modifies *tickled the bird*
  - &c.
- Our sense 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

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

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

The shape of explanation for "easy"

1. Reportability
2. Accessibility
3. Wakefulness v sleep

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.

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"

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"

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."

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 representations used, and how the computation is done (i.e., the **function-in-intension**)
- **Implementational (neuro)**
  characterize the physical states that realize the algorithm

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.

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.

53 / 200

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.

54 / 200

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.

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

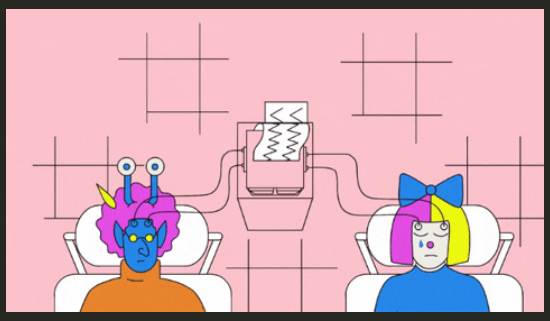
55 / 200

PHIL246: Foundations of Cognitive Science

Instructors Toby Mintz & Alexis Wellwood

TAs Megha Devraj, Rachel Keith & Nadja Winning

5 The conscious, social, moral, essentialist Day 2



56 / 200

Agenda

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

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.*

* Thomas Nagel (1986:19-20)

"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."*

* 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...."*

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

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.*

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

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, ...)

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."

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

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.

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)

Evidence for mental state recognition?

Evidence for mental state recognition?

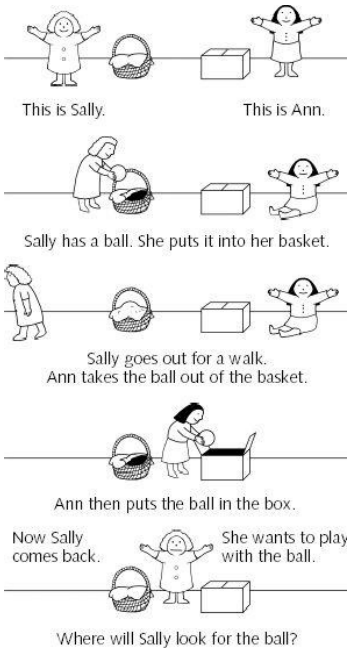
Garnering evidence sufficient to demonstrate mental state attributions is difficult ... Many behaviors that could potentially serve as indices of mental state attributions (e.g. pointing) can typically be interpreted in both mentalistic and non-mentalistic ways. This is particularly true for the attribution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

* Johnson (2000), p.22

Evidence for mental state recognition?

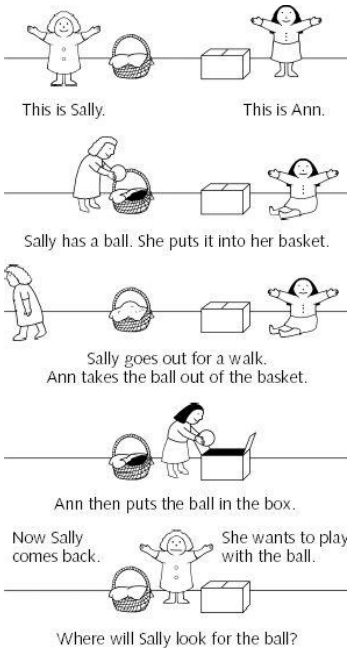
Garnering evidence sufficient to demonstrate mental state attributions is difficult ... Many behaviors that could potentially serve as indices of mental state attributions (e.g. pointing) can typically be interpreted in both mentalistic and non-mentalistic ways. This is particularly true for the attribution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

* Johnson (2000), p.22



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

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)

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*

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)

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)

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.

Case study in goal-directedness

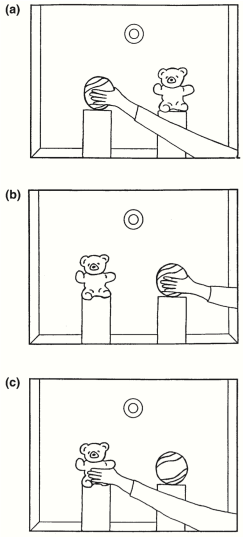
Woodward's [(1998)] study used the visual habituation method to test whether infants encode human actions in terms of the goals of the actor, or solely in terms of the spatiotemporal movements involved.... [Her] results suggest that even by five months of age, infants recognize that: (1) the behavior 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)

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.

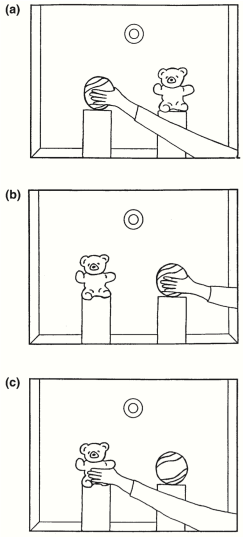


Case study in goal-directedness (Woodward 1998)

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.



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.

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)

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 interactions with other entities

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.

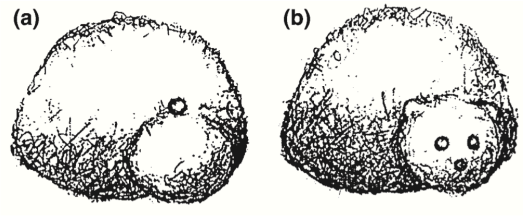


Fig. I. The novel object without a face (a) and with a face (b). (Reproduced, with permission, from Ref. a.)

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.

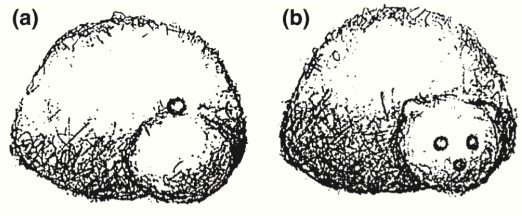


Fig. I. The novel object without a face (a) and with a face (b). (Reproduced, with permission, from Ref. a.)

Two factors tested:

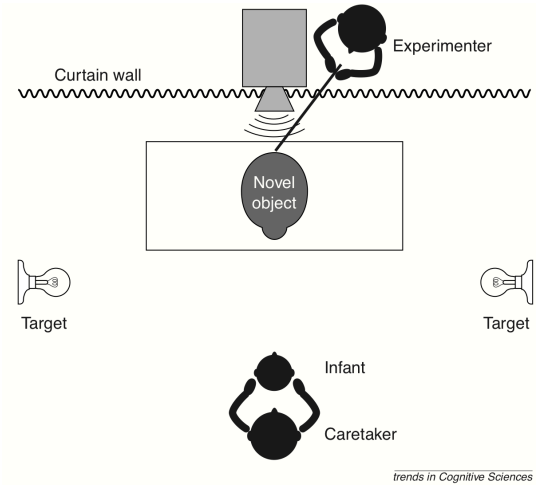
1. Has a face / is faceless
2. Responds contingently / responds non-contingently

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 / non-contingently



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

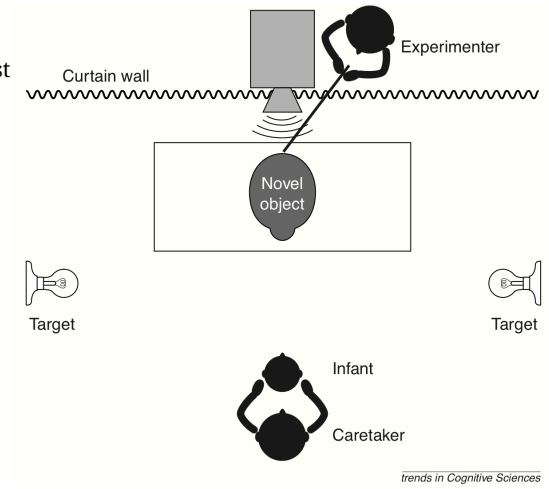
In the set-up, infants were familiarized with the novel entity for 60 seconds, during which the entity responded to the infant or randomly, as determined by the child's test condition below. There was one final "attention-getting" beep before the object would pivot to orient itself to one of two targets 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



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

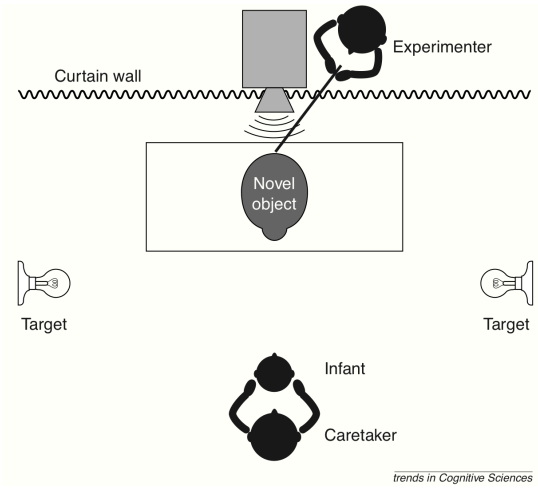
Test condition 1: Has a face, responds contingently

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.



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

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 ...

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)

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)

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-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. 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

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

[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?

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

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)

* 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)

* 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) 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)

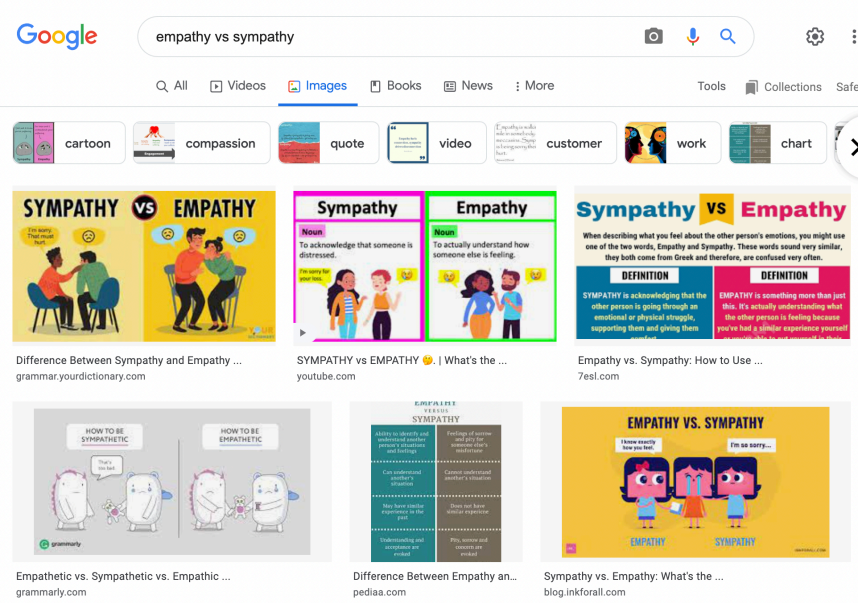
Tomasello: "shared intentionality"

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

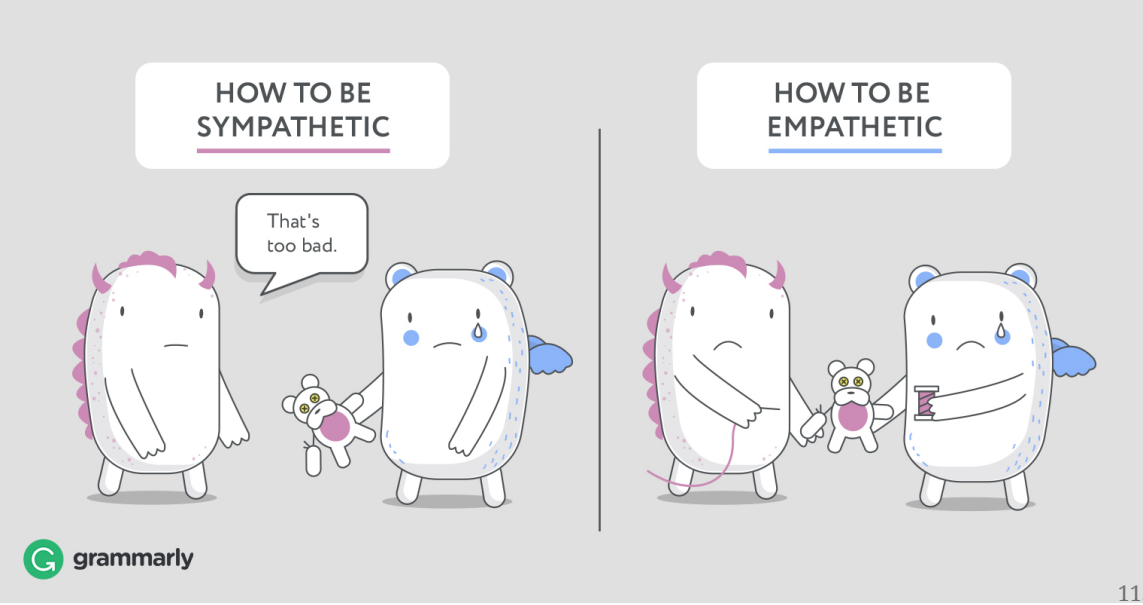
https://onlinelibrary.wiley.com › doi › full

Shared intentionality - Tomasello - 2007 - Wiley Online Library

Tomasello: "shared intentionality"



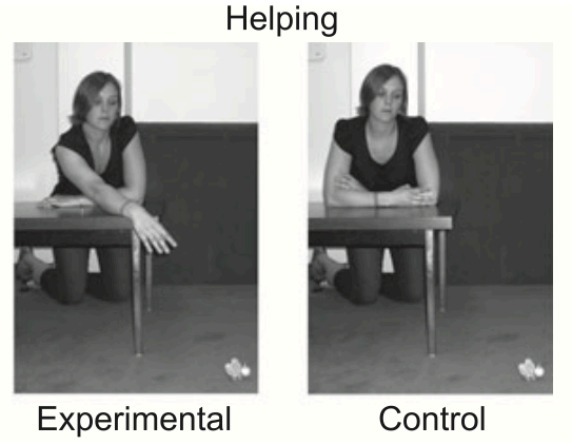
Tomasello: "shared intentionality"



115 / 200

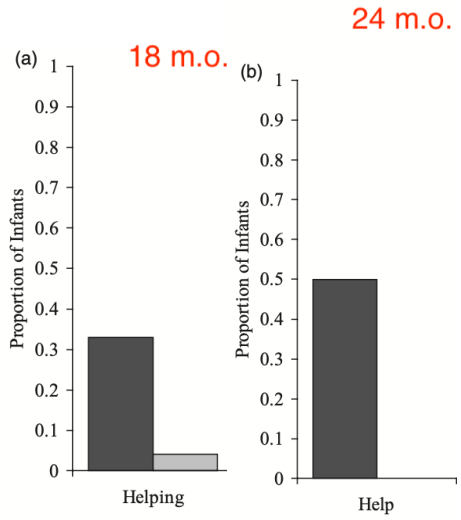
Case study in moral goodness

Case study in moral goodness



Dunfield et al (2011), "Examining the diversity of prosocial behavior: Helping, sharing, and comforting in infancy", *Infancy, 16(3)*, pp227-247.

Case study in moral goodness



(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)

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

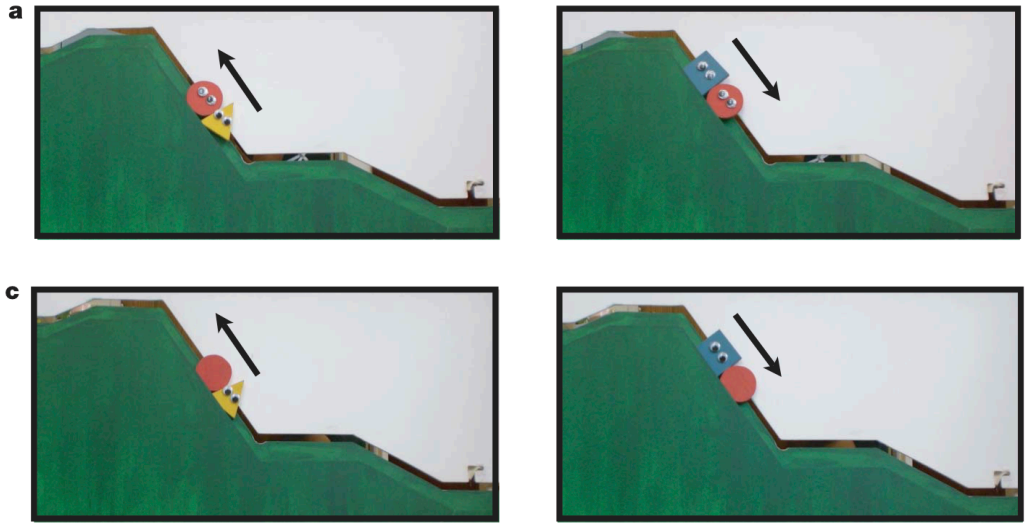
- **"Helper" facilitates** puppet P's goal (bump up the hill, open the box)
- **"Hinderer" prevents** P's goal (bump down the hill, slam the box closed)

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

Case study in prosociality (Hamlin et al 2007)

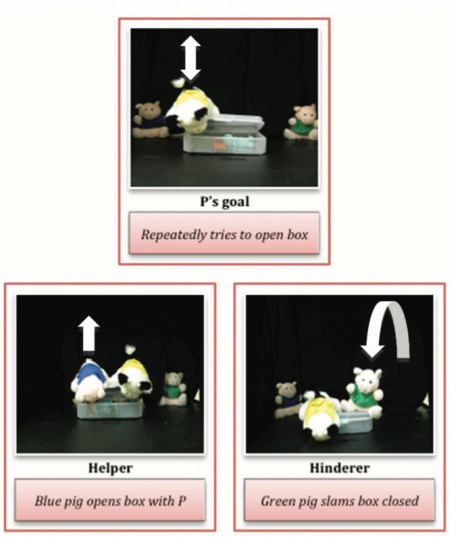
* 0:00-0:52, https://www.youtube.com/watch?v=anCaGBsBOxM ; from Hamlin, Wynn, & Bloom (2007), "Social evaluation by preverbal infants"

Case study in prosociality (Hamlin et al 2007)



Case study in prosociality (Hamlin & Wynn 2011)

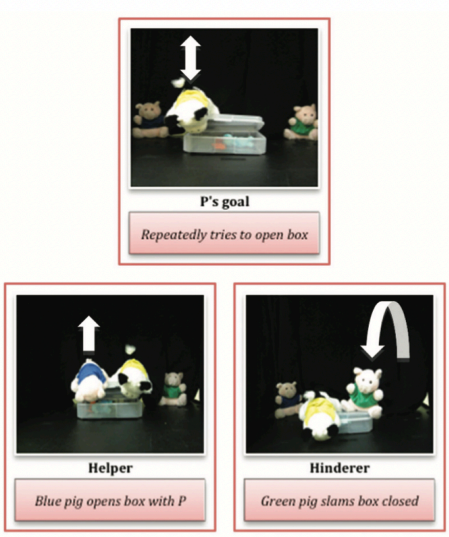
In a different follow-up discussed in the 2013 paper, Hamlin & Wynn found that infants preferred puppets who "helped" another puppet open a box, as opposed to puppets who "hindered" the puppet from opening the box, e.g. by slamming it shut.



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.

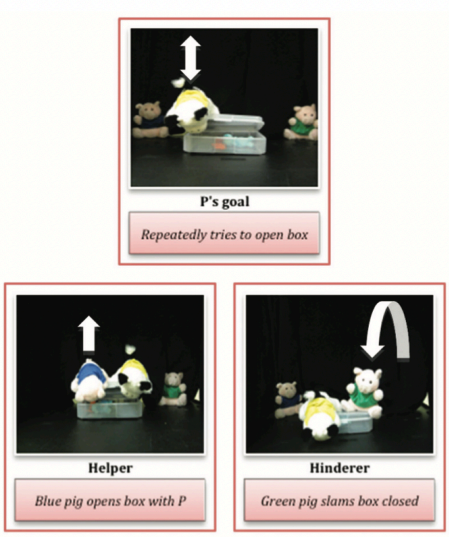
To ensure that perceptual cues could not underlie infants' preferences for characters in the box and ball scenarios, we showed age-matched infants 'control' plays, in which an inanimate claw, rather than an animate puppet, tugged on a box lid [...].... Identical puppets then directed the exact same actions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



Goals → mental states (?)

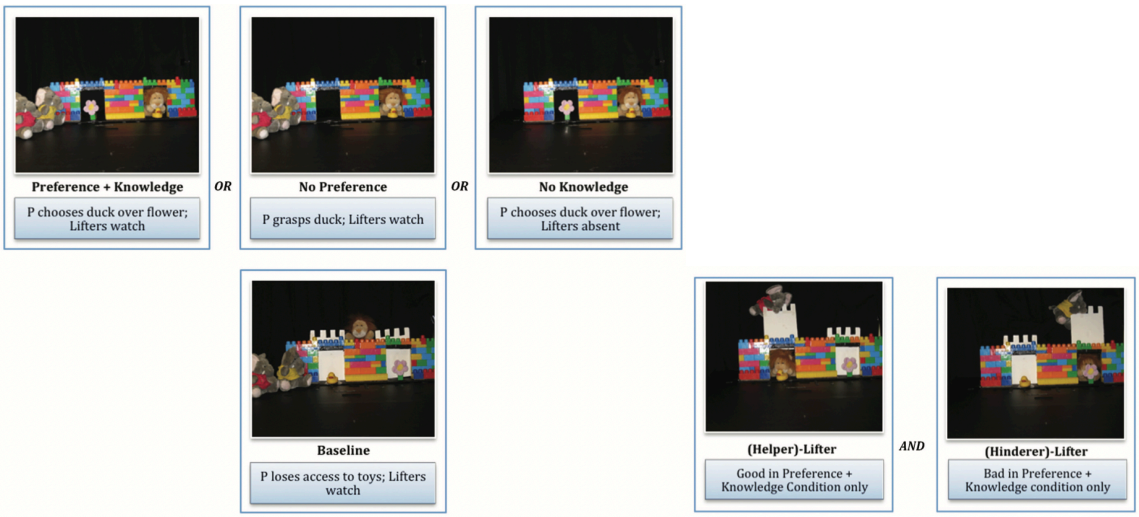
When determining moral responsibility, adults consider not only whether an individual facilitated or blocked a goal but the *mental states* that drove the action. Indeed, because humans sometimes act in error (e.g., unintentionally, naïvely), and because sometimes prosocial and antisocial behaviors are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

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)

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



130 / 200

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)

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)

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

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?

Cribbed from article introducing the *PNAS* paper: https://www.livescience.com/17204-babies-prefer-punishment.html, "Even Babies Think Crime Deserves Punishment".

Case study in moral retribution

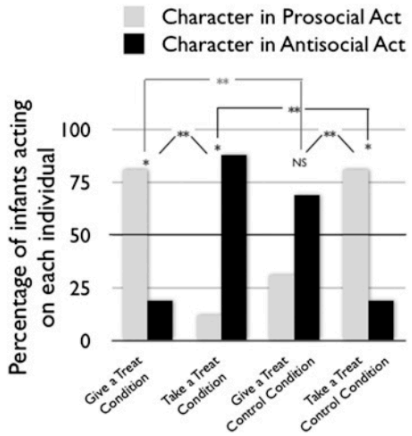


Fig. 2. Results in experiments 4 and 5. Percentage of toddlers directing Giving or Taking actions toward characters, who performed a Prosocial or an Antisocial act, or toward characters who received a Prosocial or an Antisocial act is shown. Asterisk, Binomial test, one-tailed $P < 0.05$; double asterisk,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

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?

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.

Bridge: essentializing and generalizing

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

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"!

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

Agenda

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

Agenda

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

"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.

"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.

"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'.

"The original sin of cognition"

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

"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.

"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.

"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.

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**)

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.

Generics

156 / 200

Generic sentences

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

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.

- Leslie holds that, instead, they express that instances of the noun category are **disposed** to have the property expressed by the predicate

  Being a mammal disposes one to have fur.

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)

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

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.

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.

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)

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

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

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"

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

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 ...
- Whether an individual **realizes** such dispositions is 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.

Essences: biological kinds vs artifacts*

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?

* 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"

* W. Ahn et al. (2001), "Why essences are essential 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)

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 someone from Ethiopia 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!

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)

Ks are terrorists

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

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)

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 of 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*).

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)

what is it like to be human?

199 / 200

what is it like to be human?

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

200 / 200