

**Learning and Doing:
Toward a Unified Account of Rationality in Belief,
Desire, and Action**

John Locke Lectures 2018
Dedicated to Derek Parfit (1942-2017)

Lecture 6:
“Moral Learning, Moral Evolution, and Moral Realism”

Peter Railton
(University of Michigan)
Oxford, May 2018

Discussion seminar tomorrow morning

- 9:00 am, Ryle Room, Radcliffe Humanities Building
- All welcome!

Anon.

- “Myths are invented, but morality is discovered.”

(I) Let's briefly review

A good theory

- A good theory should be *descriptively adequate*.
 - That is, it should fit the data—or perhaps, the least controversial data—reasonably well.
- It should promise to be *explanatorily adequate*.
 - That is, it should offer a plausible explanation of the data in terms that are systematic and can integrate well with other, well-confirmed theories.
- And it should help us make headway with problems that are *independently* seen as serious. Sometimes this means *dissolving* the problem by showing how it arises from a set of assumptions it throws into question, and to which it offers plausible alternatives.

The “orthodox belief-desire model” of action

- **belief** + **desire** → **action**
- *representational*
- *inert*
- *mind-to-world*
- *T/F*
- *cognitive*

- *potentially rational* + →

The “orthodox belief-desire model” of action

- **belief** + **desire** → **action**
- *representational* *non-representational*
- *inert* *motivating*
- *mind-to-world* *world-to-mind*
- *T/F* *not T/F*
- *cognitive* *non-cognitive*

- *potentially rational* + *?????* →

The “orthodox belief-desire model” of action

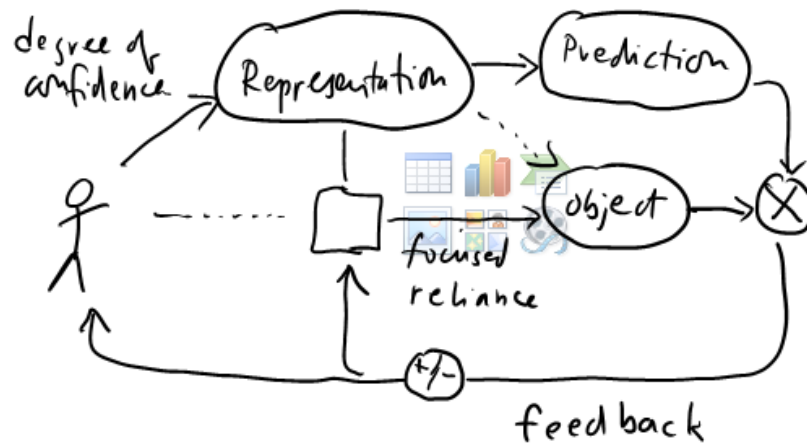
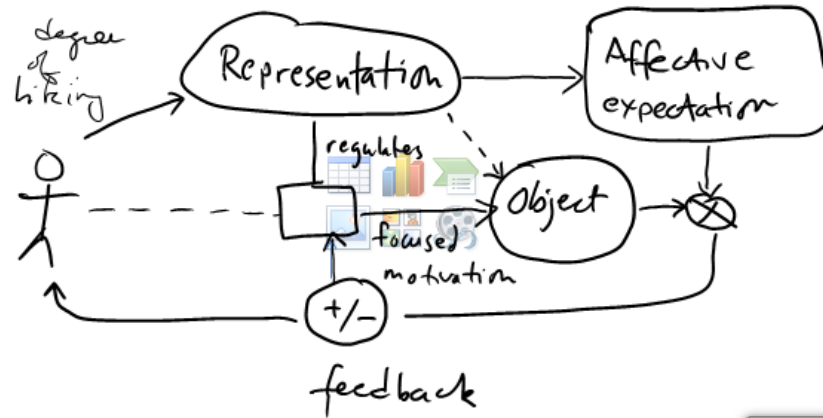
- **belief** + **desire** → **action**
- *representational* *non-representational*
- *inert* *motivating*
- *mind-to-world* *world-to-mind*
- *T/F* *not T/F*
- *cognitive* *non-cognitive*

- *potentially rational* + *?????* → *potentially rational*

Beyond the “orthodox belief-desire model” of action

- Moreover, it was unclear how such different kinds of states as desires and beliefs could come together to cause action.
 - Attempts to explain this by introducing the agent tended to launch a regress—internal acts were posited to explain external acts.
- Perhaps the problem is that we hadn’t looked carefully enough at the nature of desire and belief as such.
 - We assembled “field notes” on desire and belief, and used these to develop novel models:

Desire and belief as affective, representational, regulative, and action-guiding



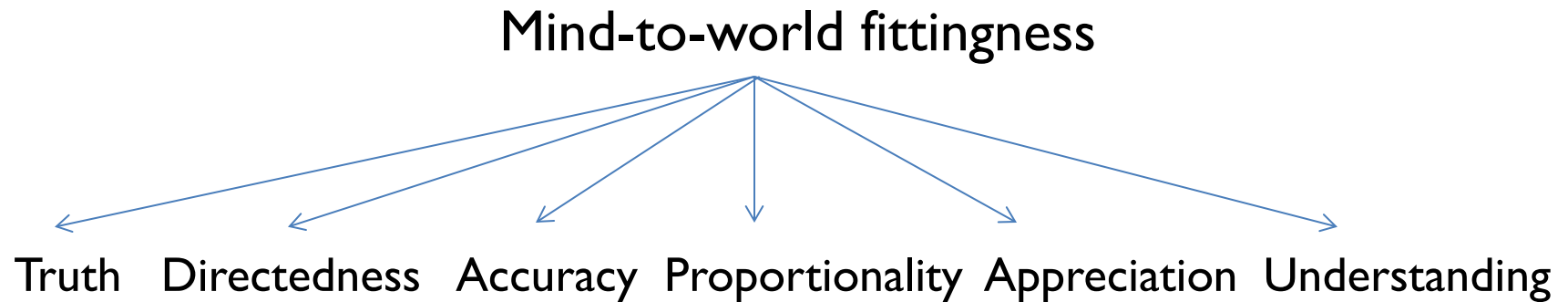
These models could explain ...

- ... a range of otherwise puzzling features of desire and belief:
 - Involuntariness
 - Projective, generalizing, content-adding
 - Provide weights both generate and guide action, that spontaneously revise in response to experience
 - Two kinds of strength of desire or belief
 - Various dysfunctions or dysregulations of desire and belief
 - How emotions can enter on all fours with desire and belief to shape thought and action tendencies

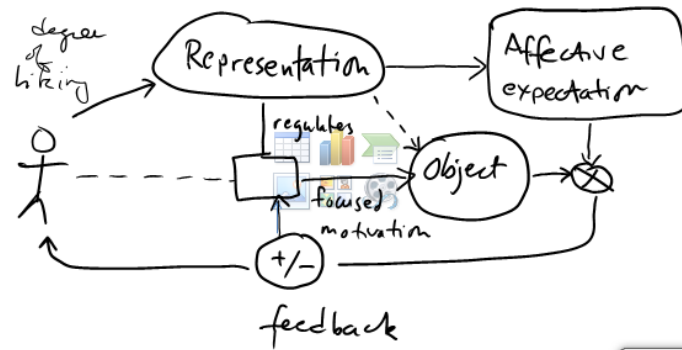
Resulting revisionist belief-desire model

- **belief** + **desire** → **action**
- *representational* *representational*
- *affective+act-guiding* *affective+act-guiding*
- *mind-to-world* *m-t-w as well as w-t-m*
- *accuracy of predict.* *accuracy of evaluative predict.*
- *representation*
regulates reliance *regulates motivation*
with learning *with learning*
from discrepancy *from discrepancy*
- *potentially rational* + *potentially rational* → *potentially rational*

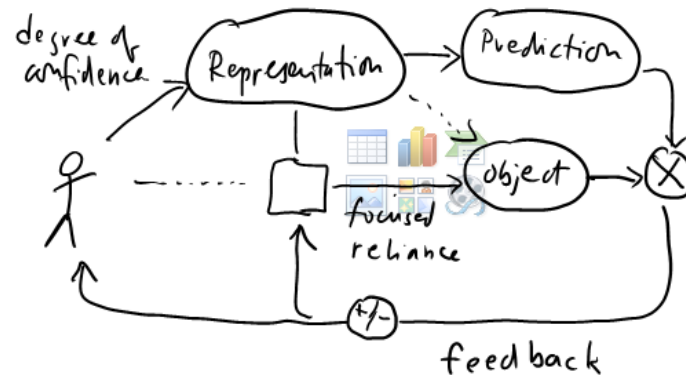
There are multiple dimensions of mind-to-world fit



Desire and belief remain distinct, constitutively and functionally – there is no need to introduce conceptually or empirically problematic “besires”, etc.



Find



(2) Some explanations

Belief and desire as dispositional states

- Robert Stalnaker: “Belief and desire ... are correlative dispositional states of a potentially rational agent.
- “To desire that p is to be disposed to act in ways that would tend to bring it about that p in a world in which one’s beliefs, whatever they are, were true.
- “To believe that p is to be disposed to act in ways that would tend to satisfy one’s desires, whatever they are, in a world in which p (together with one’s other beliefs) were true.”
[Stalnaker (1984), 15]

Dispositions

- There are, of course, many difficulties with the idea of dispositions, but if Stalnaker is right, then potentially rational individuals are *spontaneously disposed* to act in these ways—they seek out opportunities to act and action does not require some further mental action on the part of the agent.
 - E.g., we should see explicit, self-conscious deliberation and decision as *one of the things we are disposed to do*, not as something we need to *add* to beliefs and desires in order for our action to make us potentially rational.
- Belief and desire would not be doing their job in helping us to be rational—and to avoid regress—if their did not in this way help us to become *skilled with reasons and reasoning*.

For example: acting intentionally

- It has been a puzzle how to characterize *acting intentionally* if this is not a matter of forming and following a self-conscious intention.
 - And if regress is to be avoided, this must be possible.
- The dynamic, regulative model of desire and belief provides an explanation of how acting intentionally can come about and have many of the distinguishing features of acting via an explicit intention—minus the explicit intention.

Acting intentionally

- For example, when one is *A-ing* intentionally:
 - One is *A-ing* “under an idea”—one has some *representation* of what one is doing
 - This representation presents *A-ing* as having some “desirability characteristic”, such that the *A-ing* is *intelligible*.
 - The representation also gives a *satisfaction condition* for one’s acting, and orchestrates over time the deployment of one’s attention, perception, memory, inference, motivation, and behavior *for the sake of this end, held in view*.
 - Thus, the behavior is *teleologically organized*, if only implicitly.

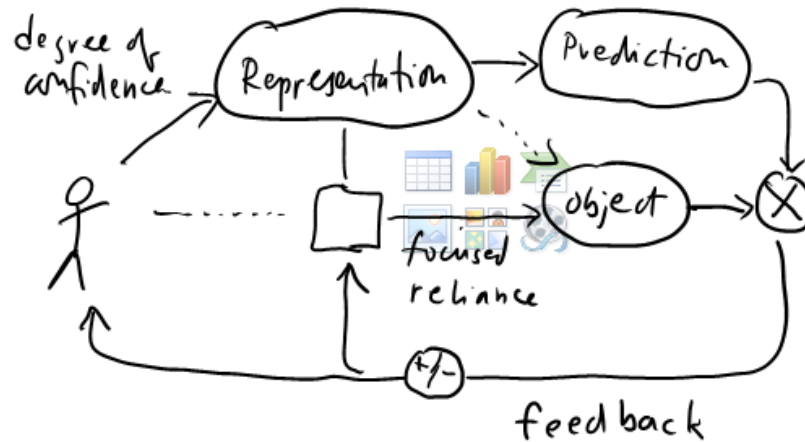
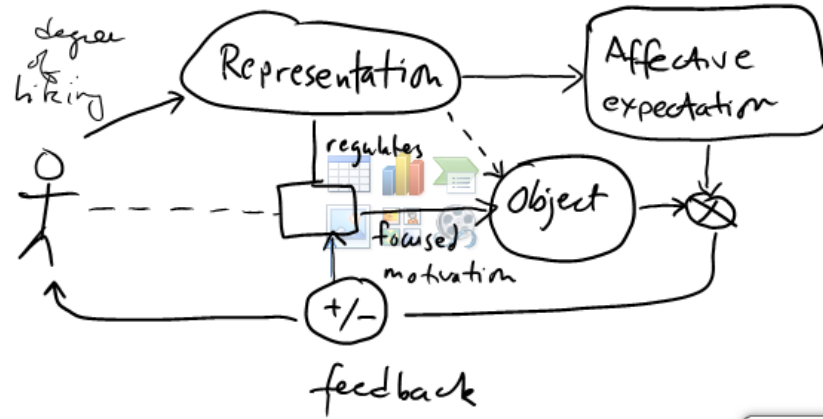
Acting intentionally

- The organizing idea thus affords the agent an answer to the question, “*What are you doing?*” that renders the behavior intelligible. (Though this answer will not always be immediately accessible to the agent.)
- Given the *regulative role* of the agent’s representation of this idea, and the inherent *learning* dynamic of desire and belief, what makes the act intelligible also explains and *guides* it.

Acting intentionally

- Thus there is a difference between behavior *caused* by one's beliefs and desires, and action that is an intentional expression of them.
 - This, we saw, could then be used to explain the possibility of forming explicit intentions, or following a rule, as an apt response to reasons, without regress.

Desire and belief



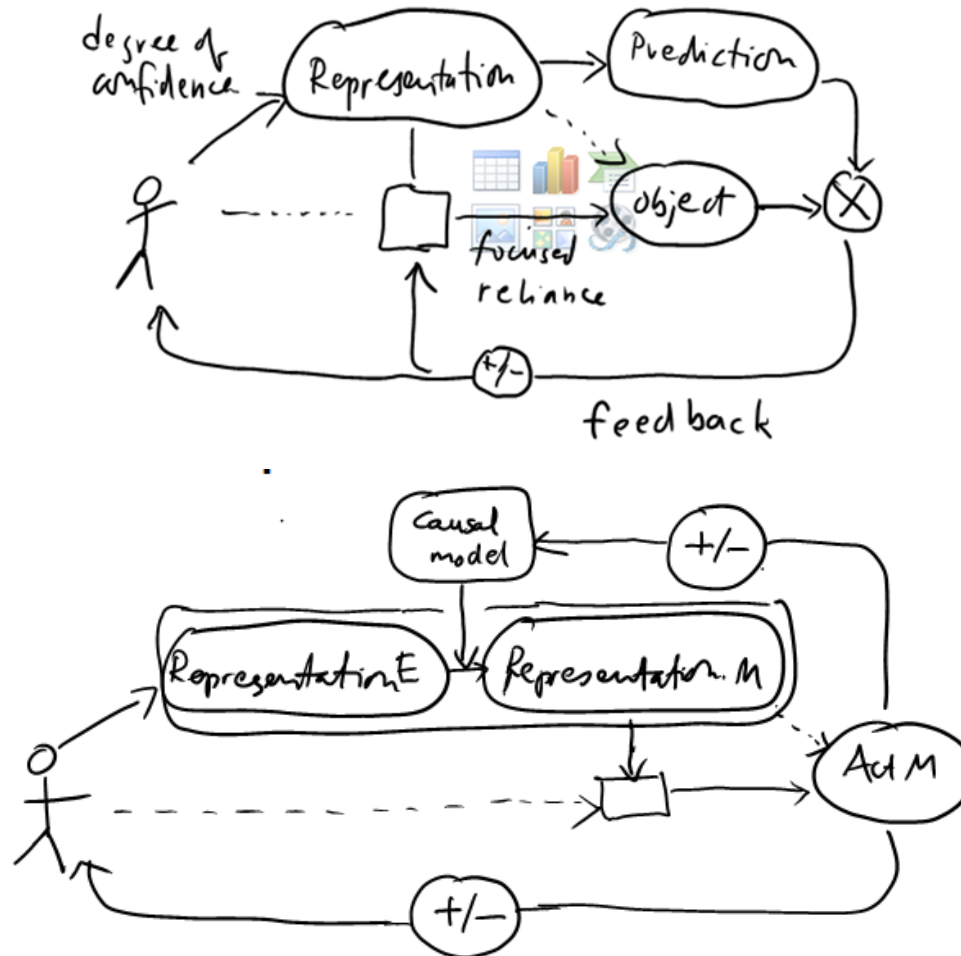
Aristotle

- “... it is always the object of desire which produces movement, but this is either the good or the apparent good” [DA 433a26]
- “... the object of desire is the starting-point for the practical intellect, and the final step is the starting-point for action.” [DA 433a9]
- “Now the origin of action (the efficient, not the final cause) is choice, and the origin of choice is appetition and purposive reasoning. ... Hence choice is either appetitive intellect or intellectual appetition; and man is a principle of this kind.” [NE I 139a32-b5]

Learning and doing

- Though Aristotle may not have intended this, the result is a “continuous process” picture of agency in the world, where one is at any time potentially acting from many desires, which are shaping different aspects of what one is doing, without needing to attend to all.
 - It is also extended over time—guidance of action not only immediately, but in on-going, monitoring way.
 - With the forming expectations and their comparison with actual outcomes, *action* takes the form of *experimentation*:
 - Learning and doing ... and doing and learning.

Deliberative appetite or appetitive deliberation



“Inference to the best explanation”

- While the view presented here is congenial to epistemologies that operate in terms of degrees of belief, we saw that it also has a place for something like outright belief at the conscious level.
- The inherent learning dynamic of belief and desire, combined with their regulative role, pushes toward the development of more accurate and powerful models.
 - We thus have here a picture of why our response to experience spontaneously and implicitly could be described as taking the form of “inference to the best explanation”—rather than confirmation-seeking or content-conservative.

Model-based intuitive understanding

- We saw evidence that intelligent animals with excellent foraging skills form generative causal-evaluative models:
 - Models that map out space in non-perspectival as well as perspectival ways (Moser *et al.*, 2008).
 - Models that are substantially independent of current sensation and can be used for imaginative simulation and learning without external reinforcement (Ji & Wilson, 2007).
 - Models that appear to figure directly in action-guidance and to permit “implicit deliberation” in the form of simulating and evaluating alternative possible pathways (Redish, 2016).

Given this dual role of models

- ... in not only initiating action, but also generating expectations and setting us up for the kind of adjustment needed to control and monitor action in a context- and information-sensitive way, we can see them as “practical modes of presentation”.
 - This would be a representational and information-intensive kind of understanding, but also practical.

(3) Affect and evaluation

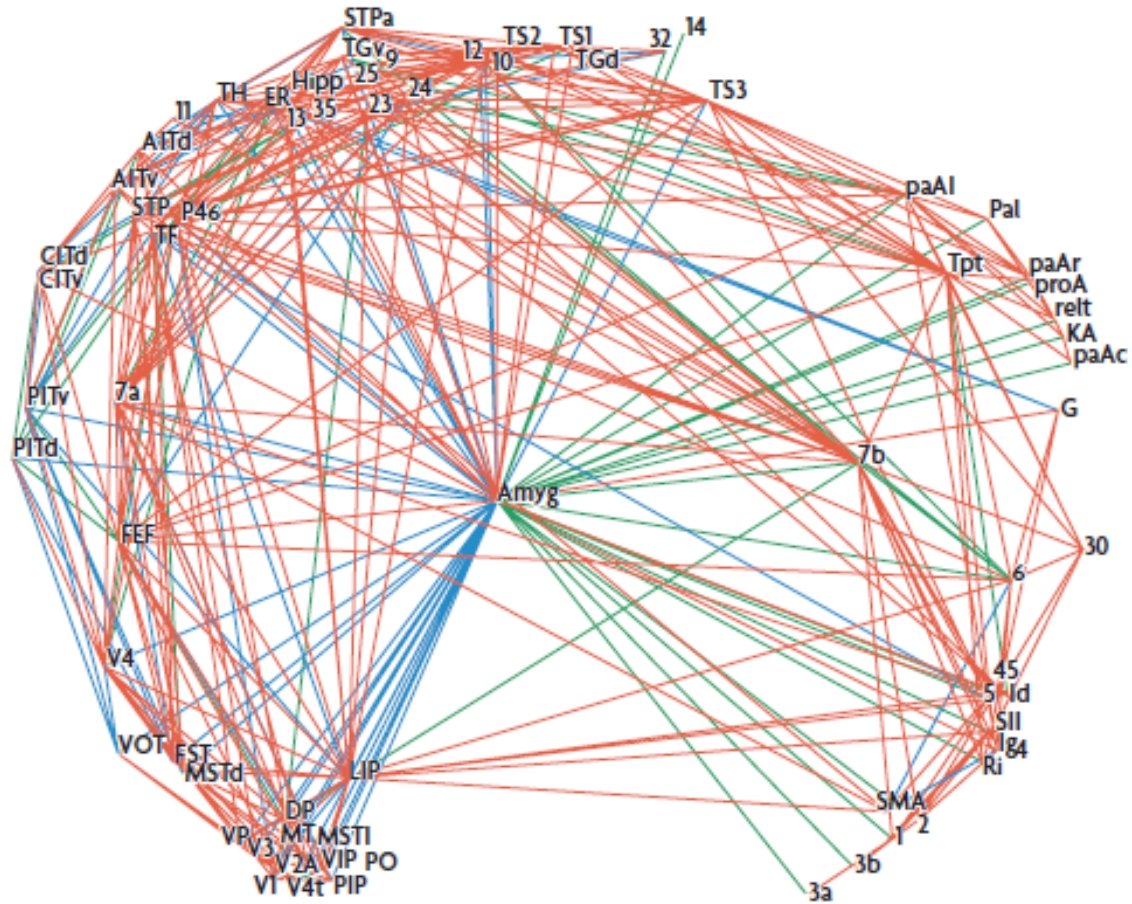
- A distinguishing feature of the account is the central role it assigns to affect. Why?

Affect and evaluation

- Different species of affect, aroused or default—fear vs. confidence, surprise vs. assurance, anger vs. affection, disappointment vs. satisfaction, and so on—correspond to different dimensions of information and value important to the regulation of thought and action in intelligent, social creatures.
 - The affective system permits a *common economy* of valuation, directly shaping attention, perception, memory, thought, feeling, motivation, and decision.
- The affective character of these responses is itself apt—affect delivers an appropriate *phenomenology* for the appreciation of value—e.g., *fear* for risk, *trust* for reliability and loyalty, etc.

Brain connectivity graph, amygdala

(Pessoa, 2008)

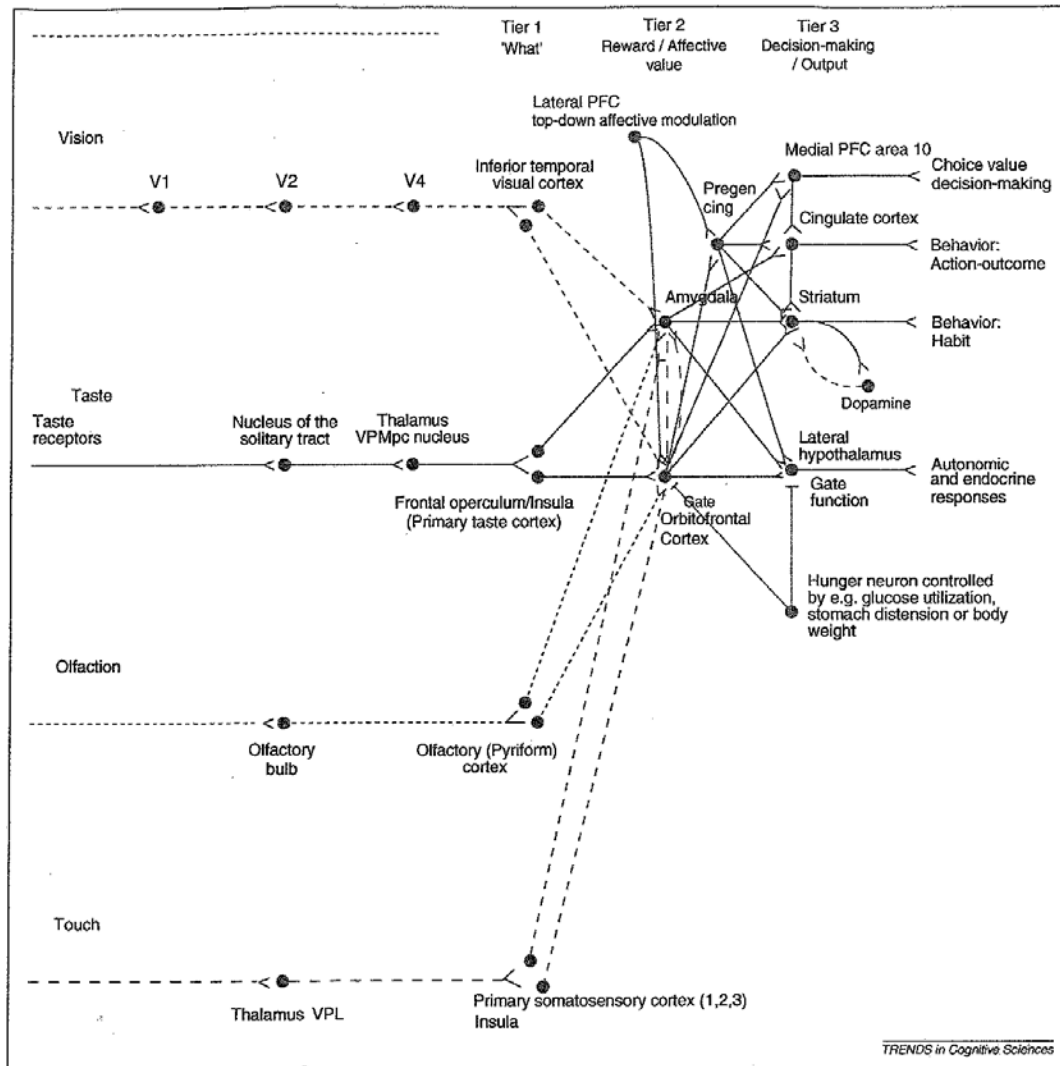


Affect and apt response

- We saw evidence from direct, dynamic recording of neurons that the affective system independently represents magnitudes of uncertainty and value, and forms corresponding expected value predictions that shape choice and behavior.
- Affect enters early in the visual stream to provide an assessment of new information that then informs more declarative reasoning.

Neural processing for valuation and decision

(Grabenhorst & Rolls, 2011)



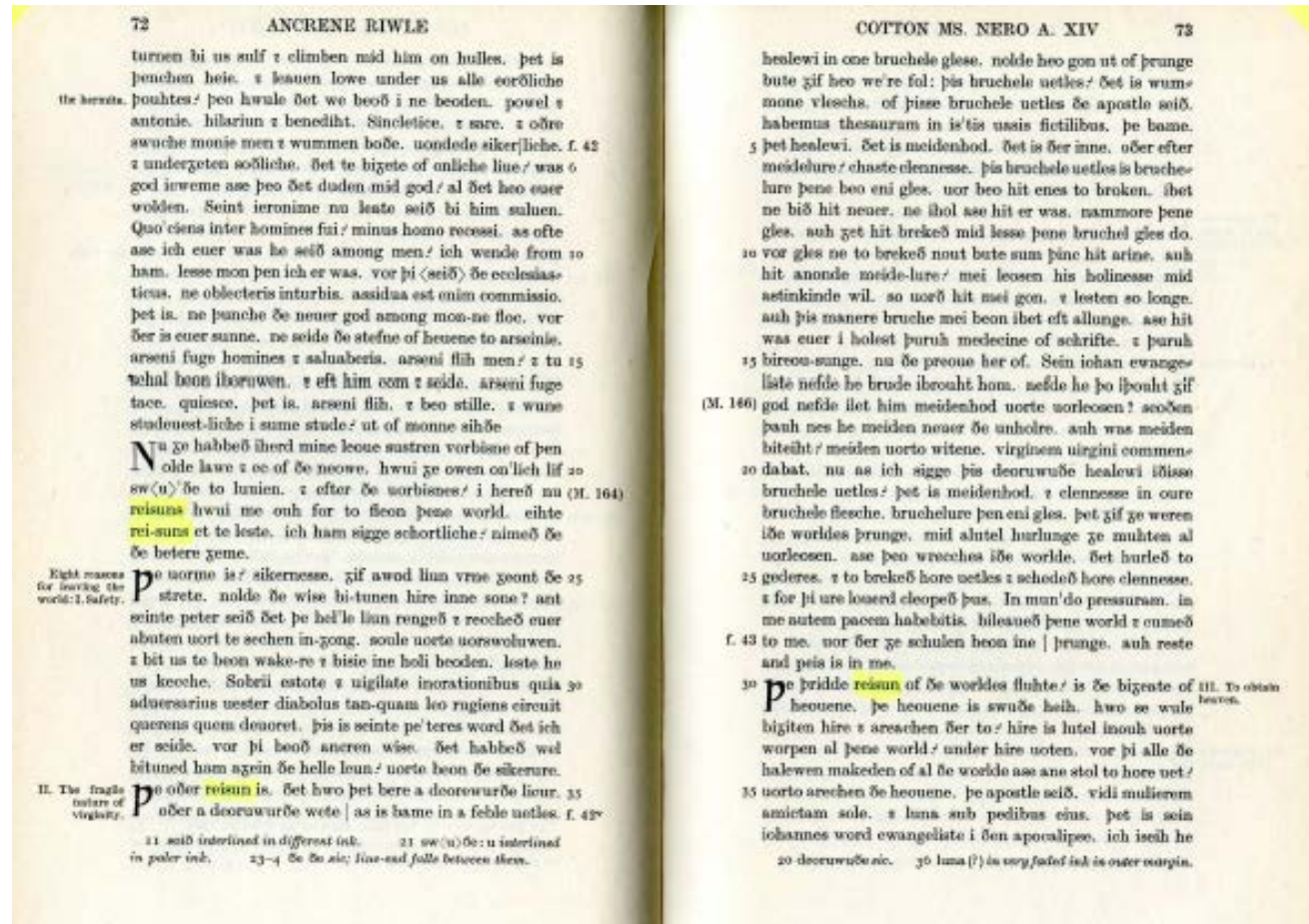
Evaluative perception

- This provides a working model of *evaluative perception* without requiring self-conscious deliberation and application of normative concepts.
- Evaluative perception of this kind, Aristotle argued, is central to how our practical understanding can engage particular situations and actions, and avoid deliberative regress.
 - Of the particulars in practical intellect he writes “these are matters of perception, and if we keep on deliberating at each stage we shall go on without end.” [NE I I 13a]
 - “We must therefore have perception of these particulars, and this perception is understanding.” [NE I I 43b]

Evaluative perception and intuitive understanding

- The learning dynamic inherent in belief and desire, we argued, can supply such understanding.
 - Aristotle: “... these states [perceptual understanding] actually seem to grow naturally, so that ... people seem to have natural consideration, comprehension, and judgment.” [NE I 143b]
- Such “consideration, comprehension, and judgment” grow with age, so that “we must attend to the undemonstrated remarks and beliefs of experienced and older or of prudent people, no less than to demonstrations. For these people see correctly because experience has given them their eye.” [NE I 443b]

(4) “Wittes skile”



Intelligence and skill with reasons and reasoning

- We should not be asking self-conscious deliberation to do what it cannot do by way of enabling us to count as aptly responsive to reasons.
 - We modeled how the regulative and learning dynamics of desire and belief could yield intuitive intelligence and skills with reasons and reasoning, that is, skill at being aptly responsive to reasons, “having reason” in the broad sense:
 - “reisun, thet is, wittes skile” (*Ancrene Riwe*, 1225—with thanks to John Broome).
- What might such “wittes skile” look like? It look like—and be of a piece with—other context- and goal-sensitive skills, and skills of this kind appear to be based upon models.

Intuition, understanding, and rationality in the broad sense

- At least since Aristotle, it's been clear that there must be forms of apt non-deliberative responsiveness to reasons—forms of responsiveness that put us in touch with.
- And at least since Aristotle, it has been common to call such non-deliberative capacities *intuition*, partly on the model of how perception puts us in touch with reasons.
 - But in addition to intuition, philosophers have spoken of *understanding*—which constitutes a form of *knowledge* represented in forms capable of being applied non-deliberatively to guide thought and action.

Intuition, understanding, and tacit knowledge

- Think, for example, of the complex but largely tacit body of information that constitutes our understanding of a language and of how to use language conversationally.
- Or our largely implicit understanding of physical and social dynamics that enables us to interact fluently with the physical and social world.
 - On the present account, we can understand such a bodies of tacit knowledge as involving casual-evaluative models, which appear to play a role in the direct guidance of skilled action, and to explain how such action can be flexible and effective in a complex and changing environment (Todorov & Jordan, 2002; Yarrow *et al.*, 2011).

Moral intuition and understanding

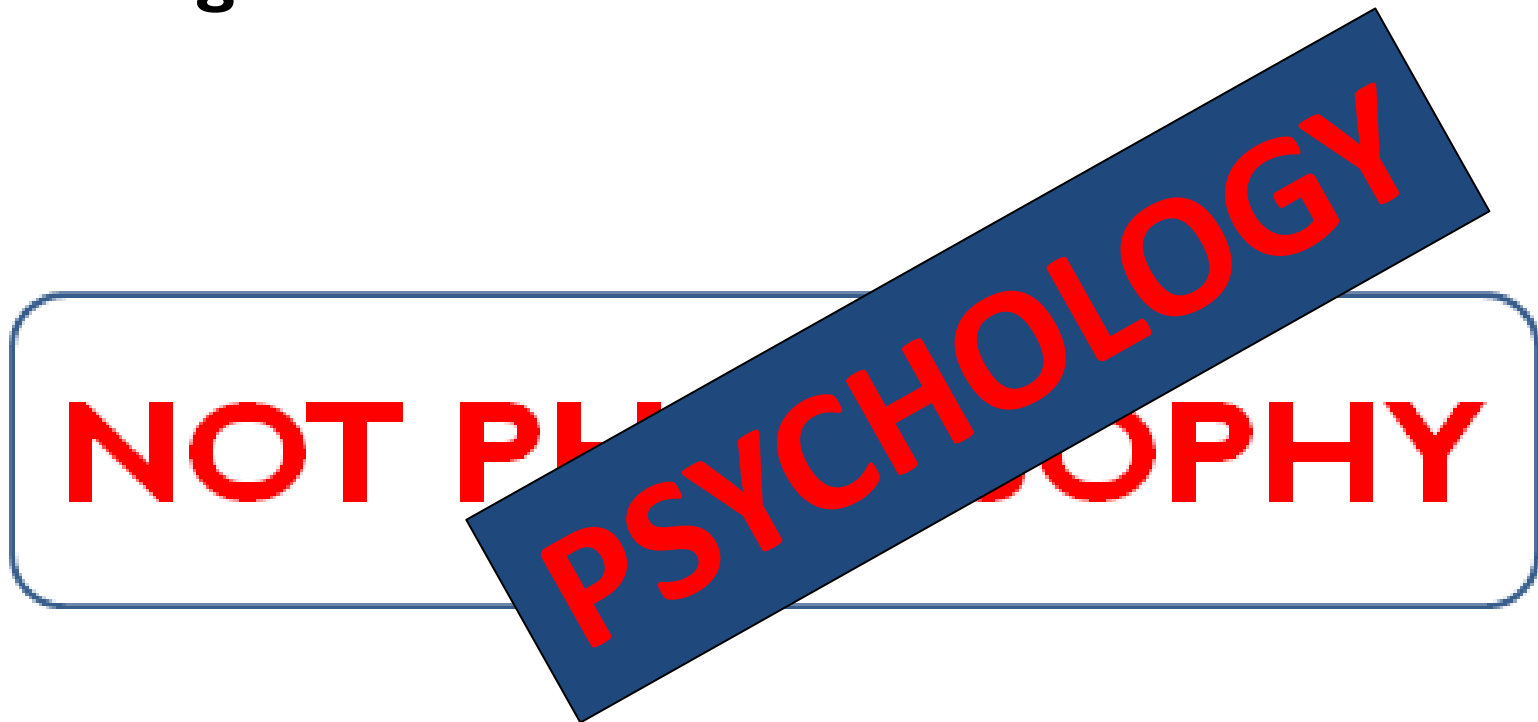
- Can this picture help us explain moral intuition and understanding,
- ... and provide the beginnings of an answer to questions about the nature and potential epistemic status of moral intuition and understanding?

(5) Moral intuition, revisited

Warning:

NOT PHILOSOPHY

Warning:



Last time

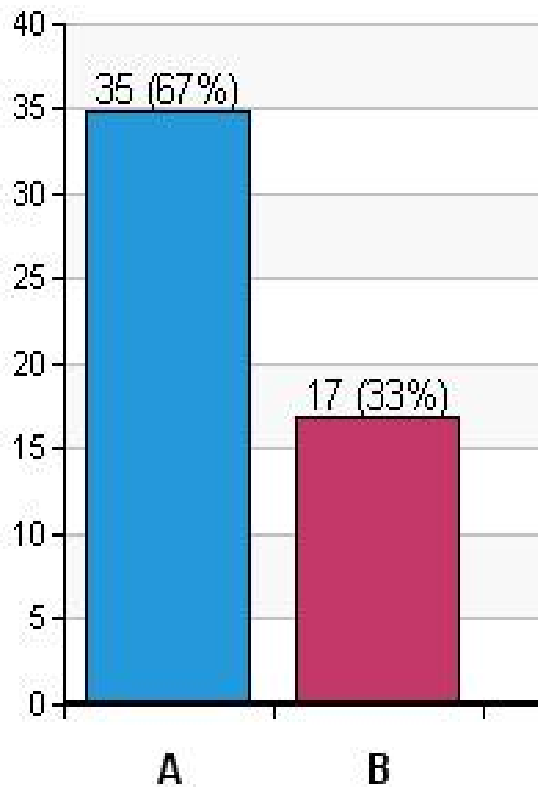
- We were looking at an array of intuitive responses in familiar and unfamiliar moral dilemmas, as well as a series of related questions about moral understanding, e.g., in the form of reactive attitudes.
 - We compared the predictive and explanatory value of contemporary “dual-process” models of these phenomena
 - ... with an alternative, model-based and approach involving the simulation and evaluation of possible acts, motivational structures, and feelings.

For example

What if you learned a friend had thrown the switch in Switch?

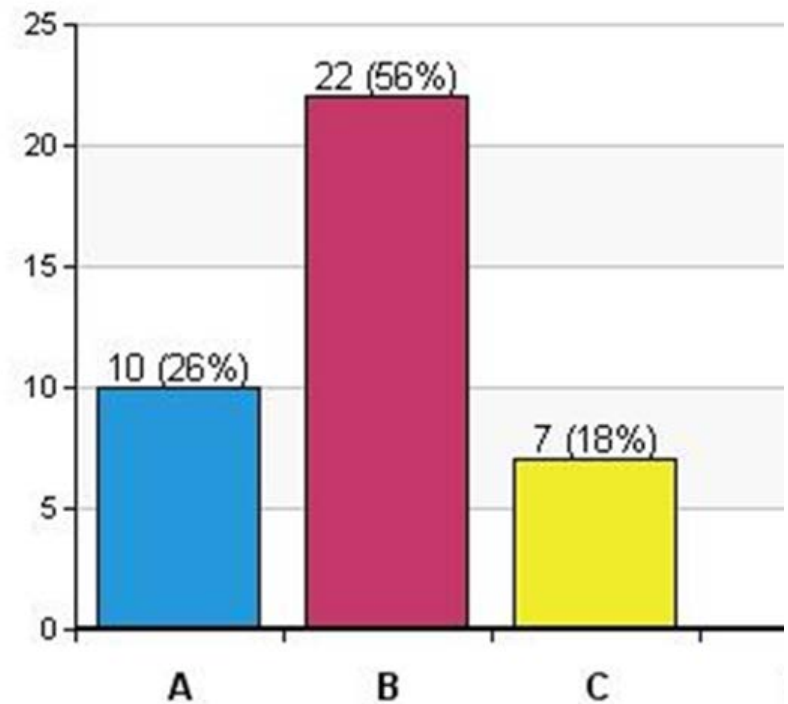
Switch

A = pull B = do not pull



Switch aftermath

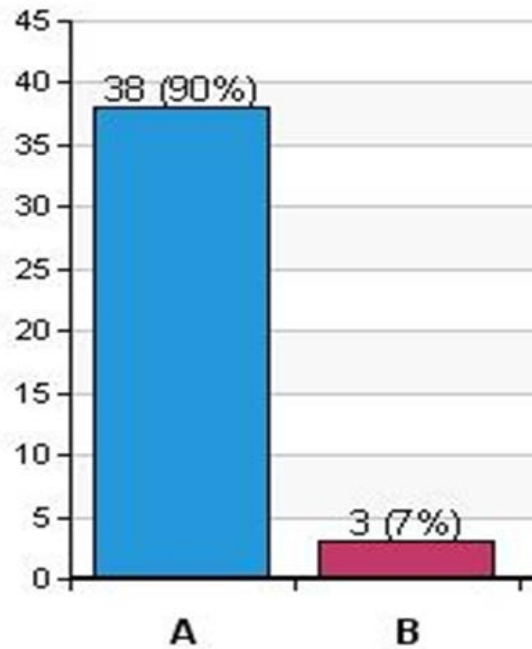
More, same, less trusting



What if you learned a friend had pulled the switch in Loop?

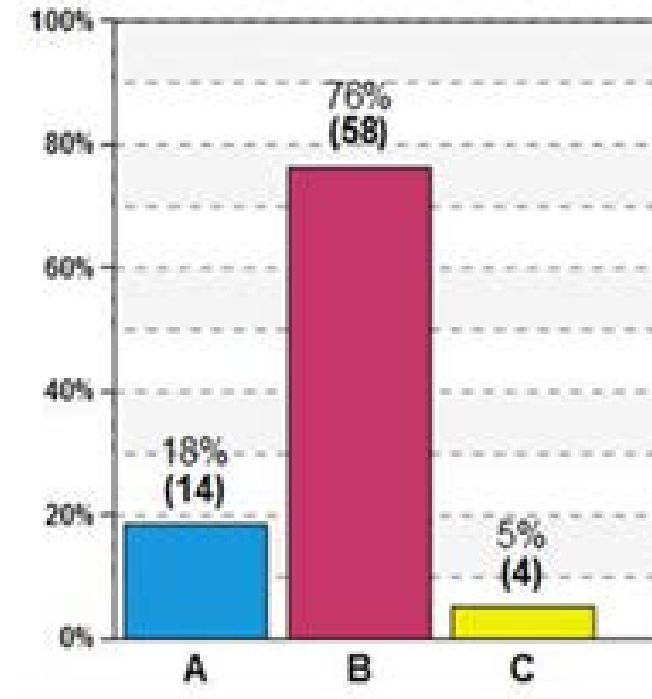
Loop

A = pull switch B = do not pull



Loop aftermath

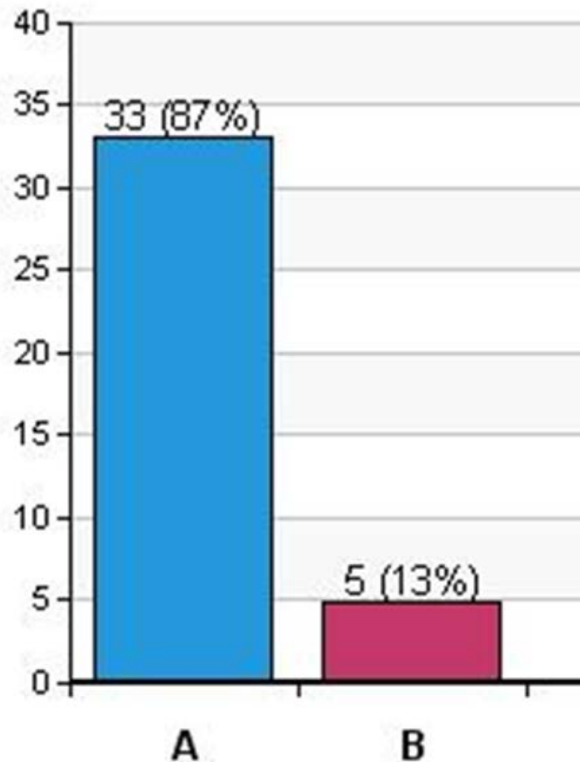
More, same, less trusting



What if you learned a friend had waved to the workers to move in Wave?

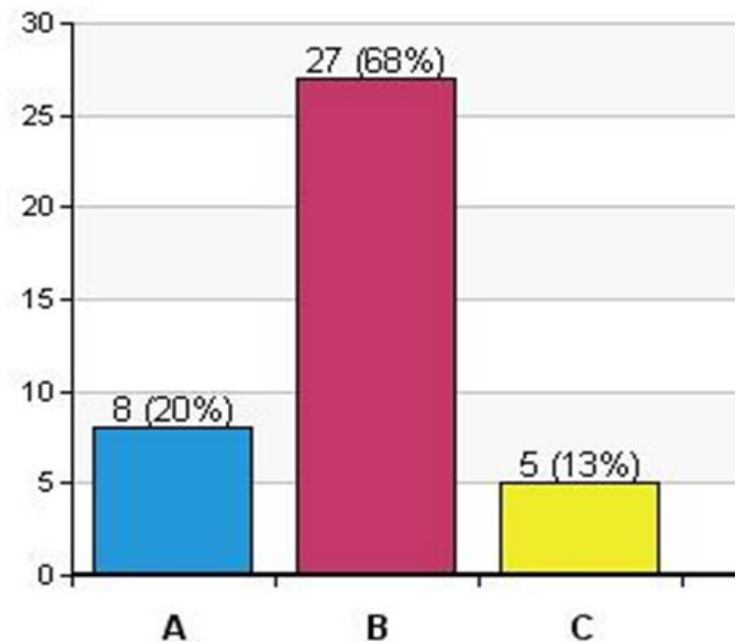
Wave

A = wave B = do not wave



Wave aftermath

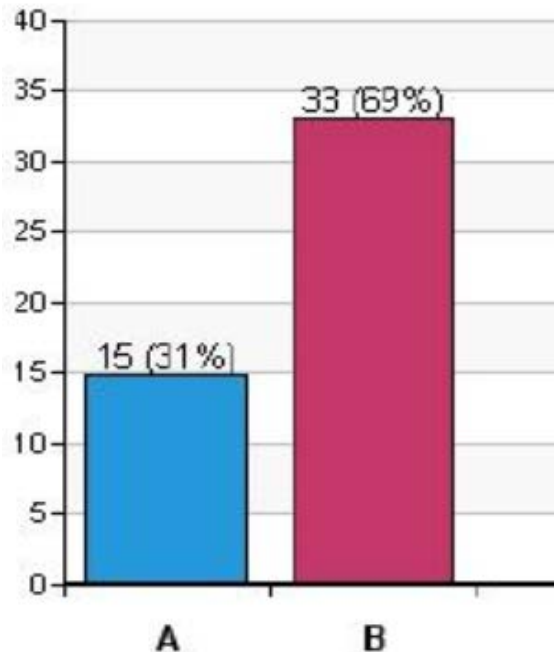
More, same, less trusting



What if you learned a friend had pushed the large gentleman in Footbridge?

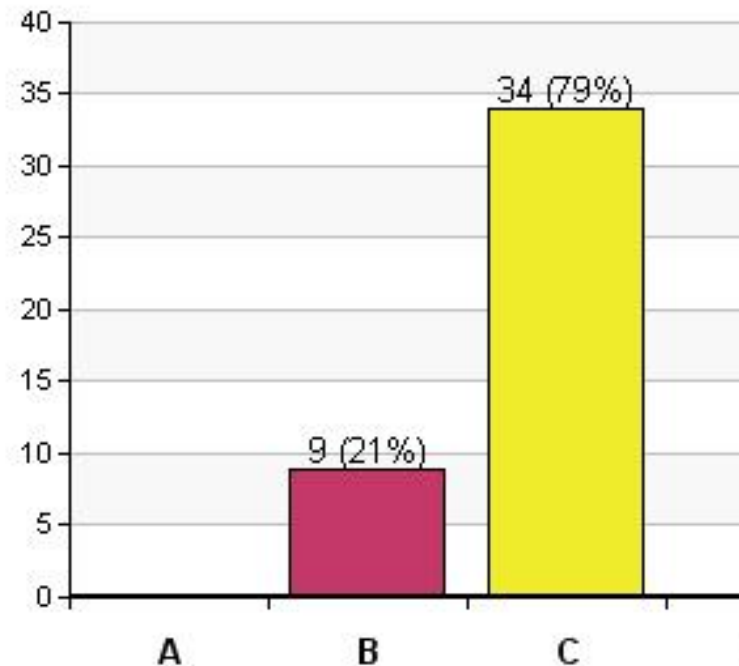
Footbridge

A = push B = do not push



Footbridge aftermath

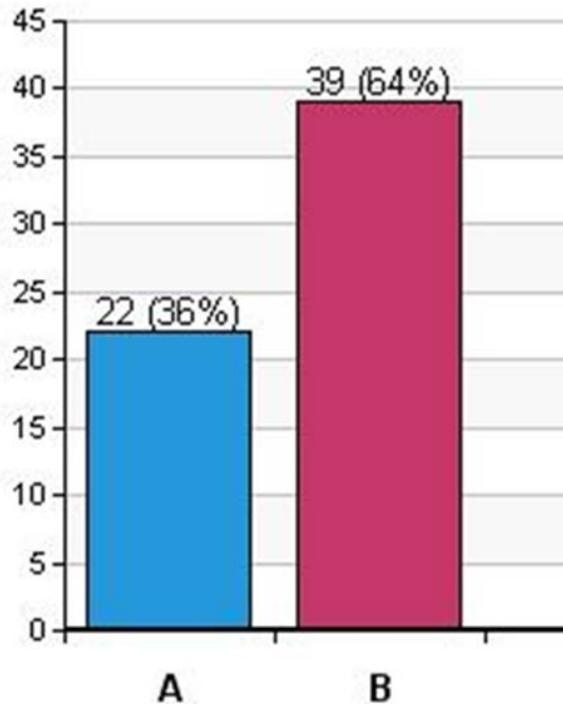
More, same, less trusting



What if you learned a friend had beckoned the large gentleman in Beckon?

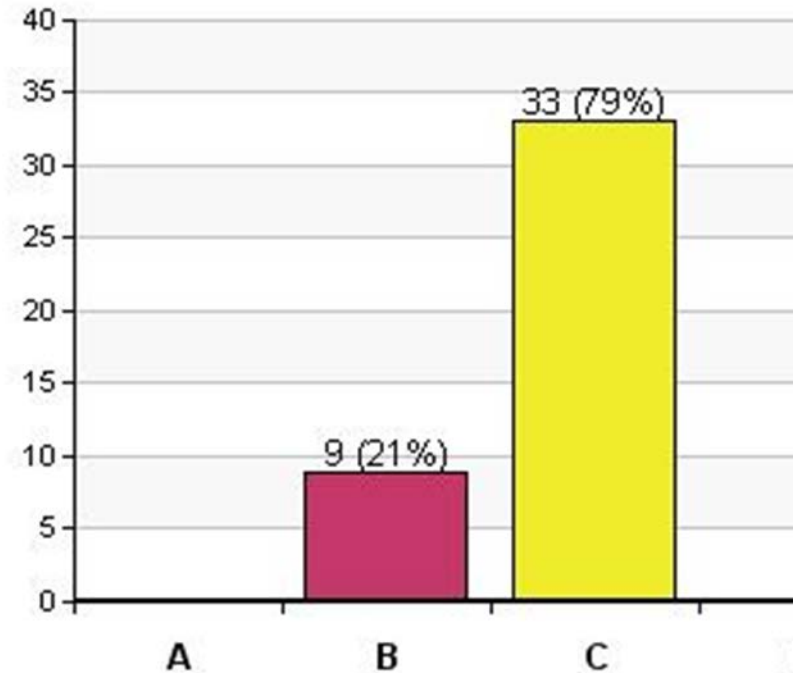
Beckon

A = beckon B = do not beckon



Beckon aftermath

More, same, less trusting

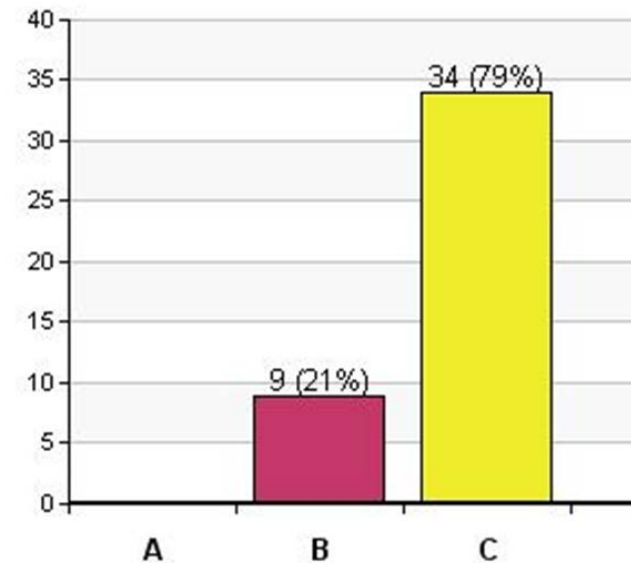
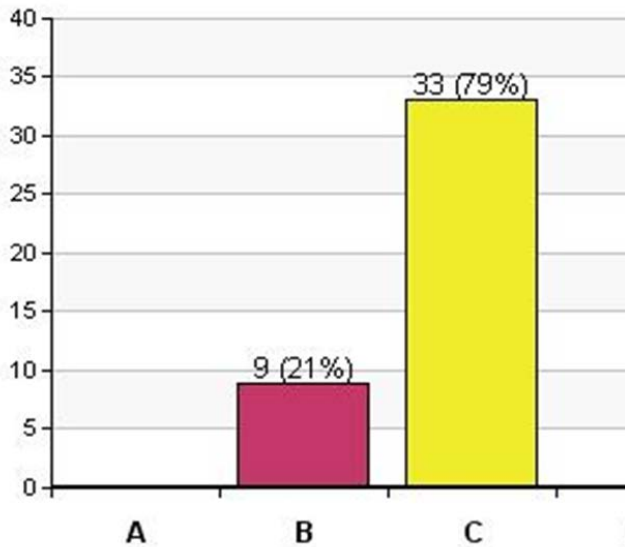
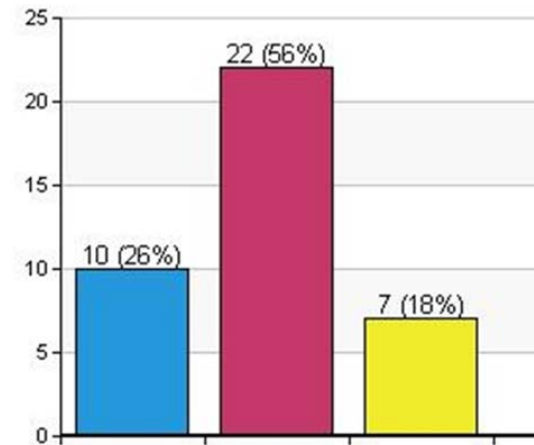
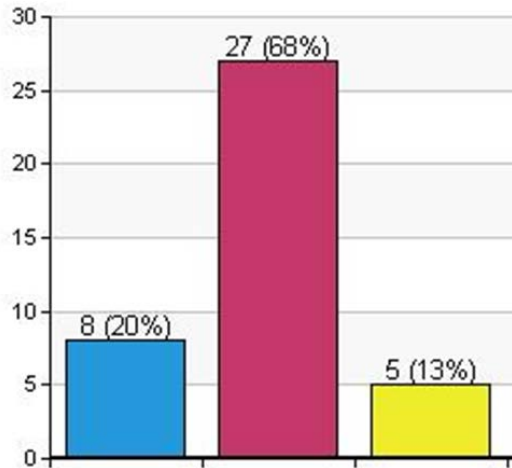


Trust

Wave and Switch

Beckon and Footbridge

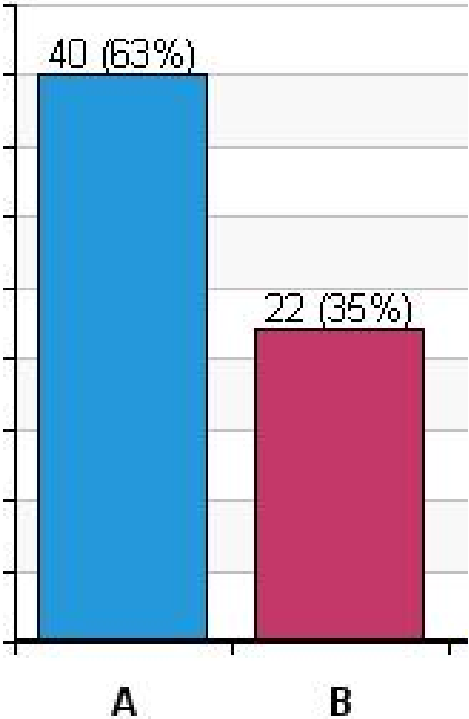
More, same, less trusting



What if you learned a friend had pushed the large gentleman in Bus?

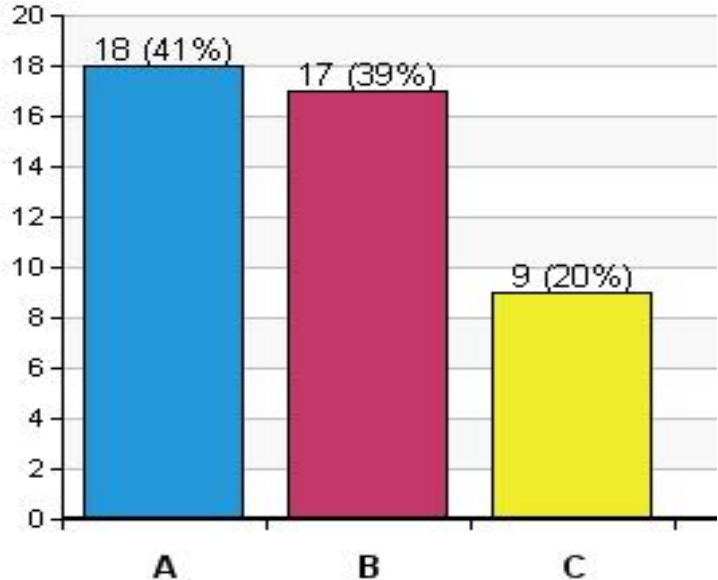
Bus

A = push B = do not push



Bus Aftermath

more, same, less trusting



Collateral evidence

- ... that my students were picking up on something real:
- Bartels & Pizarro (2011), Gao & Tang (2013), and Kahane *et al.* (2014), found that likelihood of giving a “push” verdict in Footbridge-like scenarios was not correlated with general altruism, but with rating on psychopathy scale, egoism, and disregard for moral violations generally.
- Conway & Gawronski (2013), Gleichgerrcht & Young (2013), Weich *et al.*, 2013) found decreased levels of empathy, harm-aversion, and perspective-taking in those giving push-like responses in Footbridge-like scenarios.
- Duke and Begue (2014) found that higher alcohol level predicted greater tendency to give “push”-type verdicts.

Models of the agent mediate moral intuitions




- Uhlmann *et al.* (2013) found that a projected model of the agent as lacking in empathy and character mediated judgments in trolley cases.
- Everett *et al.* (2016) found that “inverse inferences” were made of trustworthiness of agents in trolley scenarios.

“Reactive attitudes”: Switch vs. Footbridge

A = regretful and sympathetic, reasonable hope

B = regretful, guilty, and sympathetic, some hope

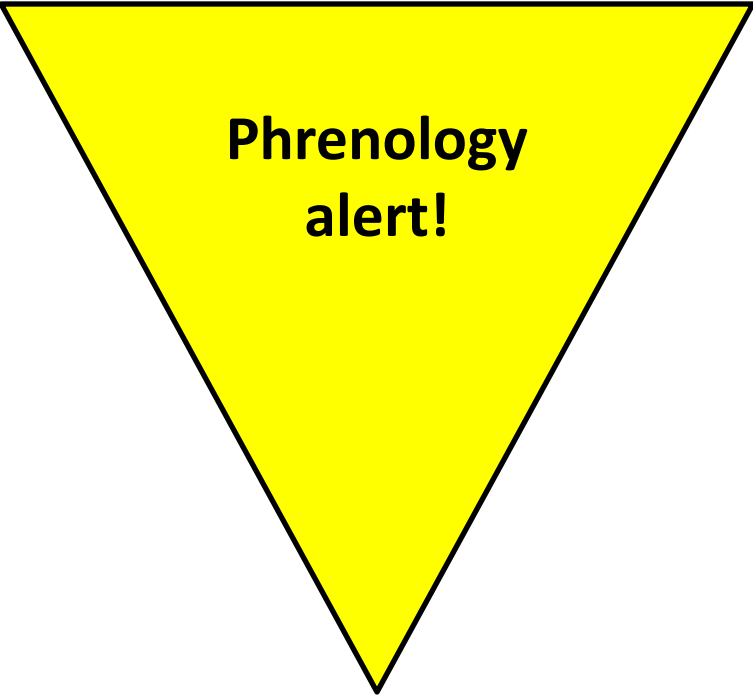
C = regretful, ashamed, and sympathetic, little hope

Response		Vote %	Votes
A		32%	16
B		46%	23
C		20%	10

Response		Vote %	Votes
A		12%	6
B		33%	17
C		54%	28

Descriptively and explanatorily

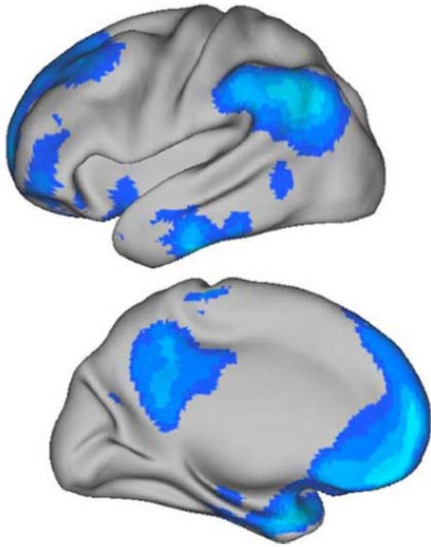
- ... a model-based account does well with these judgments, and with a range of others we discussed.
- Is there independent reason to think that general model-based capacities are at work in moral judgments?



**Phrenology
alert!**

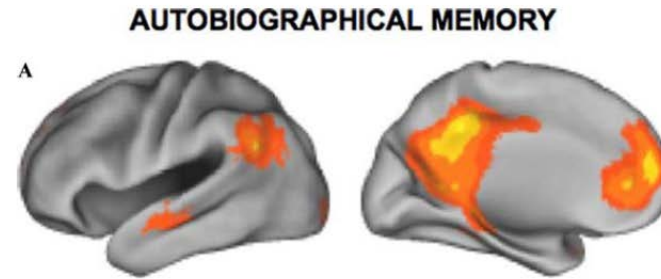
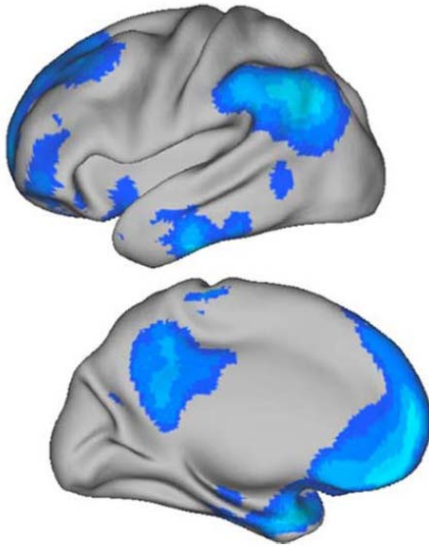
Default network

(Buckner *et al.*, 2008)



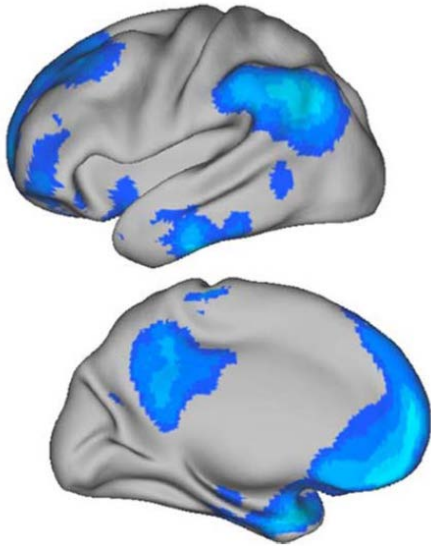
Default network

(Buckner *et al.*, 2008)

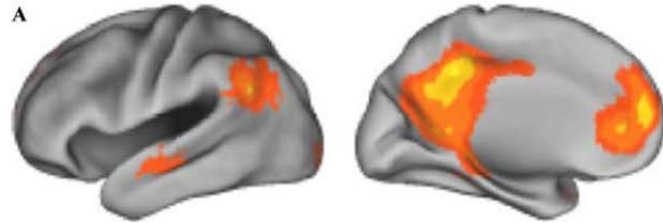


Default network

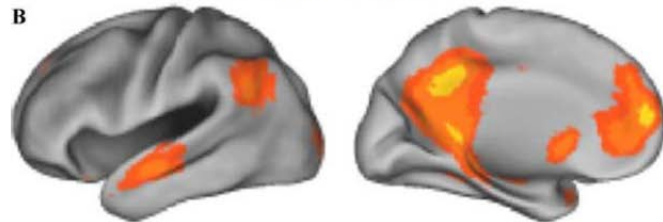
(Buckner *et al.*, 2008)



AUTOBIOGRAPHICAL MEMORY

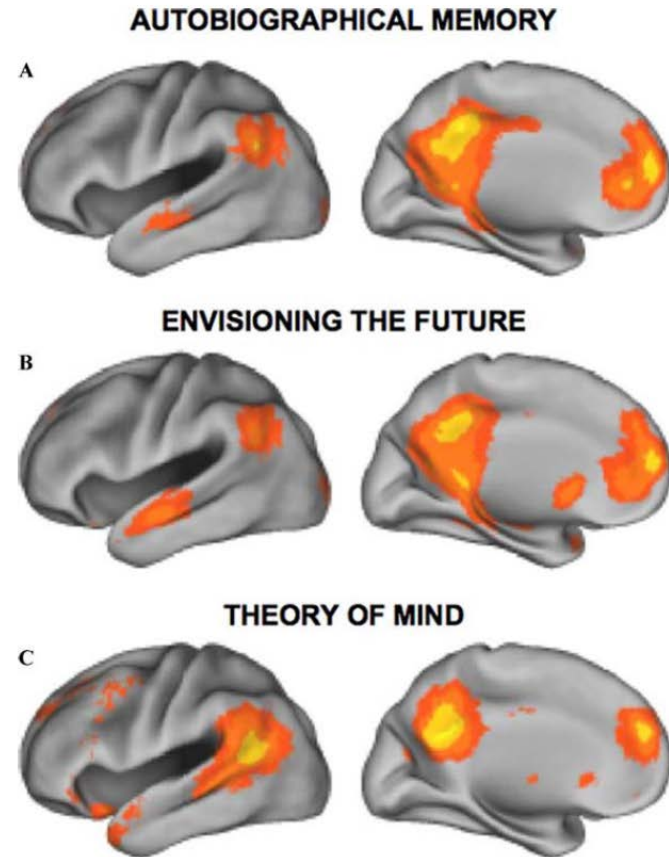
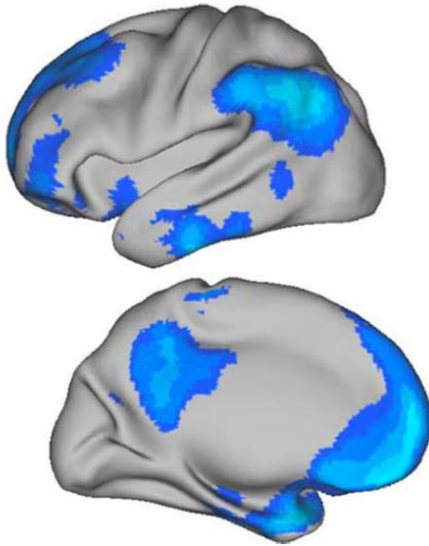


ENVISIONING THE FUTURE



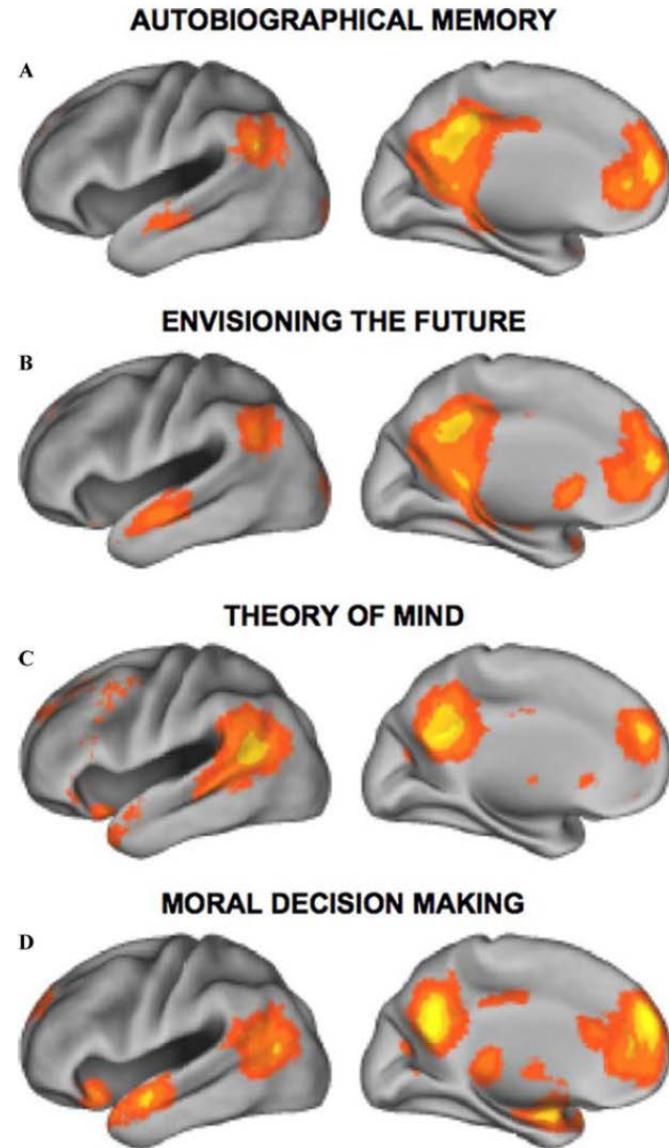
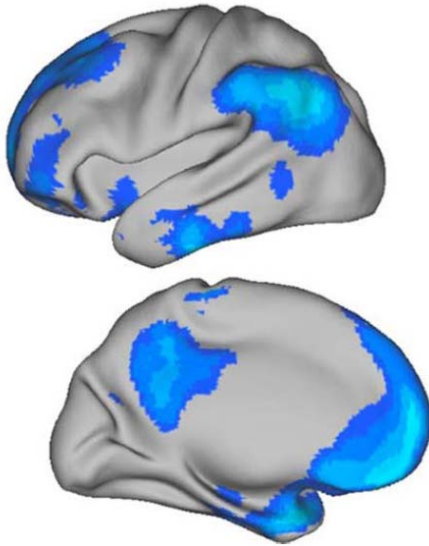
Default network

(Buckner *et al.*, 2008)



Default network

(Buckner *et al.*, 2008)



A more unified picture of evaluation and action

- The default mode, one of two fundamental modes of operation of the brain, involves integrated networks that recruit information widely.
- Evidence suggests that a chief function of this mode is the prospective simulation and evaluation of actions and outcomes recruit information widely (Buckner & Carroll, 2006; Hassabis & Maguire, 2009; Moll, *et al.*, 2005; Shenhav & Greene, 2010).
 - Such simulation, we have seen, can promote a fuller representation of the physical or social environment and its possibilities (Buckner *et al.*, 2008; Daw *et al.*, 2016; Seligman *et al.*, 2016).

Imaginative rehearsal and self-expression

- Such on-going imaginative rehearsal enables the mind to explore possibilities mentally and use experience to *prepare* action in ways that can help explain fluency.
 - Aristotle was right:
 - Extensive experience is normally required for acquiring the rich, accurate models that underlie genuine skills.
 - This is not “habit” in the modern sense, however—it representationally-rich, flexible “practical intelligence”.
 - And we see most clearly the depth of someone’s skill in their ability to act *spontaneously*. Such action is not “automatic”, but *self-expressive*.

(6) Evolutionary concerns and the development of moral skill

Some evolutionary psychologists and evolution-inspired philosophers ...

- ... have suggested that we should not expect humans to be equipped with a capacity to track morally-relevant considerations in their own right.
-

A word from anthropology

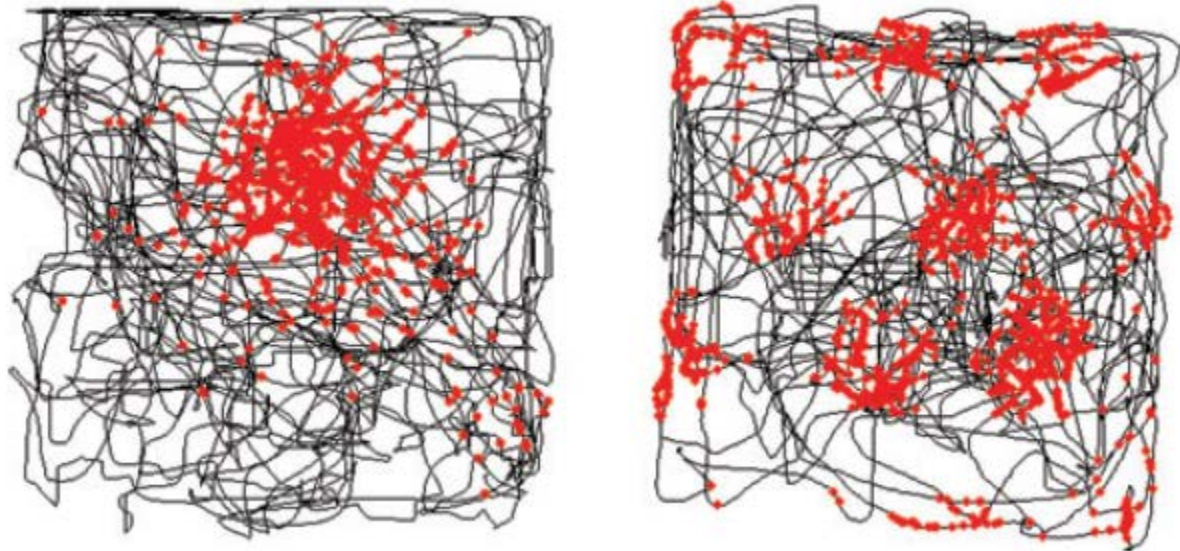
- Based on the most systematic anthropological and archaeological studies we have of hunter-gatherer societies, current and historical, it would appear that:
 - During the longest period of human evolution, the Late Pleistocene, *Homo sapiens* seem to have lived in hunter-gatherer groups. Studying such groups today, and drawing upon archaeological evidence, such Late Pleistocene hunter-gather groups appear not to have had dominance hierarchies, and practiced high levels of sharing within the group even among those not closely genetically related (Boehm, 2014), and often with movement between groups (Marlowe, 2004).

Still ...

- ... claims about evolution are highly speculative.
- Perhaps we could look more directly at whether humans display the kinds of capacities needed for apt responsiveness to moral reasons as such?
 - Keep in mind: it is not a prerequisite for apt responsiveness to moral reasons that one *conceive* them in moral terms—what matters is how one responds to morally-relevant considerations in thought and action.
 - So: we can look at the question developmentally.

Recall: Perspectival and non-perspectival spatial mapping—place and grid cells in the rat

(Moser *et al.*, 2008)



Non-perspectival *social* mapping

- Highly intelligent social animals, like chimpanzees, spend a considerable amount of time observing third-party behavior.
 - They use these observations to form accurate expectations about
 - which tasks require cooperation
 - what sort of cooperation is needed
 - which other individuals would make the best cooperation partners for them (Melis *et al.*, 2006).
- Moreover, the mammalian reward structure is highly flexible, and can take as its objects social relations and abstract values.

To push further, ...

- ... we will need to look at the literature on a particular branch of the primate bush, *Homo sapiens*,
- ... focusing on early infant learning, starting at a time when explicit instruction would be rare, if possible at all.

Non-perspectival *epistemic* mapping

- Infants aged 2-8 months use passive observations of the environment to form accurate expectations of visual statistics and phonetic sequences, suggesting domain-general processes (Sobel & Kirkham, 2007; Kidd *et al.*, 2012).
- By 12 months, infants distinguish reliably between “unable” vs. “unwilling” behavior among adults (Woodward *et al.*, 2009).
- By 16 months, they show heightened attention to mistaken labeling and labelers in learning words (Koenig & Echols, 2003)

Non-perspectival *epistemic* mapping

- By 36-48 months, infants use third-person behavior by adults to discriminate adult accuracy, knowledgeability, competence, reliability, deceptiveness, and quality of will, and use these discriminations to guide their own behavior (Doebel & Koenig, 2013; Lane et al., 2014; Sobel & Corriveau, 2010.)
- By the fourth year, infants pay increased attention to the domain-relevance of imputed adult traits in deciding what to learn from whom (Sobel & Corriveau, 2010).

Default, defeasible trust

- Such epistemic mapping could be seen as made possible by an infant's capacity for default, defeasible trust.
 - As we argued earlier, incapacity to extend default trust to one's senses and faculties, or to other people, would render the infant incapable of acquiring the information needed to gain evidence of the reliability of their senses or of others.
- We can think of such default trust as an epistemic *prior* that enables learning and participation in an epistemic community. Infants who experience unstable, unreliable environments have difficulty developing trust, and subsequent difficulties in learning and social interaction (ref.).

Epistemic autonomy and objectivity

- At the same time, a capacity for default, defeasible trust equips a child for *autonomous* learning—by relying upon their own experience as well as others, they can achieve some independence from relations of personal affiliation or authority.
 - For example, with growing experience, infants become increasingly willing to rely upon information from unfamiliar individuals who display greater epistemic competence or reliability than a familiar caregiver (Harris & Corriveau, 2011).
- With time, in effect, their epistemic mapping gains objectivity

Non-perspectival *moral* mapping

- Infants' modeling of the intentional or narrative structure of third-party adult interactions conform to the predictions of Bayesian causal inference (Hamlin *et al.*, 2013),
- More controversially, infants in the first year also show marked *preference* for third parties whose behavior exhibits morally-favored patterns (Hamlin, 2013).
 - Infants as young as 4-6 months follow with interest “morality plays” involving puppets who help or hinder, and show a reliable preference for helpers. By 8 months, infant sophistication with intentional structure is such that they prefer puppets who *hinder* a hinderer (Hamlin *et al.*, 2011).
- We can't rest great weight on these very early results—but they are suggestive about possible *priors* in infant cognition.

What mediates such mapping and preferences?

In part, empathy

- Another way to approach the question: Hume argued that human behavior generally gave evidence of a capacity for “sympathy” in which one responded directly to the interests of others without mediation by self-interest.
 - Such sympathy is perfectly general, and can be elicited by vivid representation of the condition of others, even in very remote circumstances.
 - It permits non-perspectival representation of the values at stake in social situations and interactions, and can generate from this representation a positive personal response to acts, types of acts, or states of character that contribute to well-being generally.
- Hume insists that this capacity is not a distinctively “moral sense”—it is equally involved in our capacity for language, for thinking about our own futures, for understanding others, and for effective participation in social life, and. It is form of *general purpose cognition that also affords a basis for moral learning.*

Intrinsic empathic motivation

- Do we see evidence of this? Even in the first year of life infant response to the distress of others has begun to shift from “empathic distress” to “empathic concern”,
 - ... so that by 9-10 months, infants show signs of spontaneously attempting to help those in distress (Geangu *et al.*, 2011).
 - And as their physical abilities grow, so do infant attempts to assist others showing evident need for help or to comfort those in distress (Roth-Hanania *et al.*, 2011)
- By 12-16 months, infants engage in attempts to assist those in distress or need even in the absence of external encouragement or reward (Warneken & Tomasello, 2006).

Not empathic concern alone

- Empathic concern appears to make infants spontaneously sensitive to, and motivated by, morally-relevant features of third-party interactions that involve helping or harming.
- Moreover, there is recent evidence that, by 15 months, infants are also spontaneously sensitive to *unequal* or *unfair* divisions of rewards to third parties in circumstances where the recipients of the rewards are equally-situated.
 - Moreover, this sensitivity appears to be more than a sensitivity to violations of convention, since it is closely related to actual behavior in sharing with others even at some expense to the self (Schmidt & Sommerville, 2011).

Default, defeasible trust and cooperation

- Just as we can see how acquiring a selective ability to trust others epistemically depends upon a measure of initial, default reliance upon one's senses and faculties, and other people's testimony,
 - ... so does acquiring a selective ability to trust and cooperate with others depend upon an initial, default willingness to trust others and attempt to cooperate with them. We can think of this as a pro-social prior that helps equip an infant for the social cognition and motivation involved in effective participation in a community.
- As in the prisoner's dilemma, an initial extension of unsupported cooperation can elicit the cooperation of others, and help avoid becoming trapped in non-cooperative cycles.

Autonomy and objectivity

- At the same time, having some measure of default trust in one's own experience with, and responses to, others promotes feedback and enables one to acquire a greater degree of autonomy.
 - With time, an infant's map of the morally-relevant features of its social environment can become less dependent upon relations of personal affiliation or authority.
- The same kinds of pressures to generalize and abstract found elsewhere in model-based learning apply here, helping account for a growing ability to use considerations of analogy, perspective-taking, and consistency in moral thought. These are forms of enhancing objectivity and reducing dependence upon purely subjective or arbitrary considerations.

One of the most striking examples ...

- ... of infants' spontaneous capacity for this kind of autonomy in their understanding of morally-relevant features of their world and trust in their own responses is the large body of cross-cultural evidence that, by age 3-4,
 - ... infants reliably distinguish moral violations from mere violations of authority or inconveniences,
 - ... and moreover attribute this difference to the presence of harm or benefit in moral cases,
 - ... and further show intrinsic motivation to follow, and later to enforce, moral norms even in the face of contrary authority (Smetana, 1989; Turiel, 2002).

Why call this *moral* modeling or learning?

- Consider first the epistemic case. Again, we believe that individuals can be aptly responsive to epistemic reasons without deploying normative epistemic concepts. What does matter?
 - Do they represent evidentially- or epistemically-relevant information in its own right?
 - Does the individual's representation of these epistemically-relevant features encode an understanding of their nature, and appreciation of how or why they are relevant?
 - Do these representations orient thought and action—including evaluation and motivation—in ways appropriate to the epistemic relevance of these features?

Why call this *moral* modeling or learning?

- Consider now the moral case. Again, we believe that individuals can be aptly responsive to moral reasons without deploying normative moral concepts. What does matter?
 - Do they represent morally-relevant information in its own right?
 - Does the individual's representation of these morally-relevant features encode an understanding of their nature, and appreciation of how or why they are relevant?
 - Do these representations orient thought and action—including evaluation and motivation—in ways appropriate to the moral relevance of these features?

It appears, then, ...

- ... that, just as infants construct non-perspectival, general representations of spatial, causal, and epistemic relations in the world around them,
- ... so do they begin the project of constructing representations of *morally-relevant* features of actions or situations that are:
 - (a) non-perspectival,
 - (b) general,
 - (c) consistent,
 - (d) thought- and action-guiding,
 - (e) independent of authority or sanction,
 - (f) concerned with harms and benefits, or fair sharing.

This would be an example of “wittes skile”

- Learning how to respond aptly to *kinds* of reasons—epistemic or moral—by
 - ... being alive to relevant factors
 - ... representing them non-perspectivally
 - ... being capable of regulating thought and action accordingly
 - ... even if one cannot articulate the underlying understanding upon which they are based.

(7) Normative relevance?

Normative relevance?

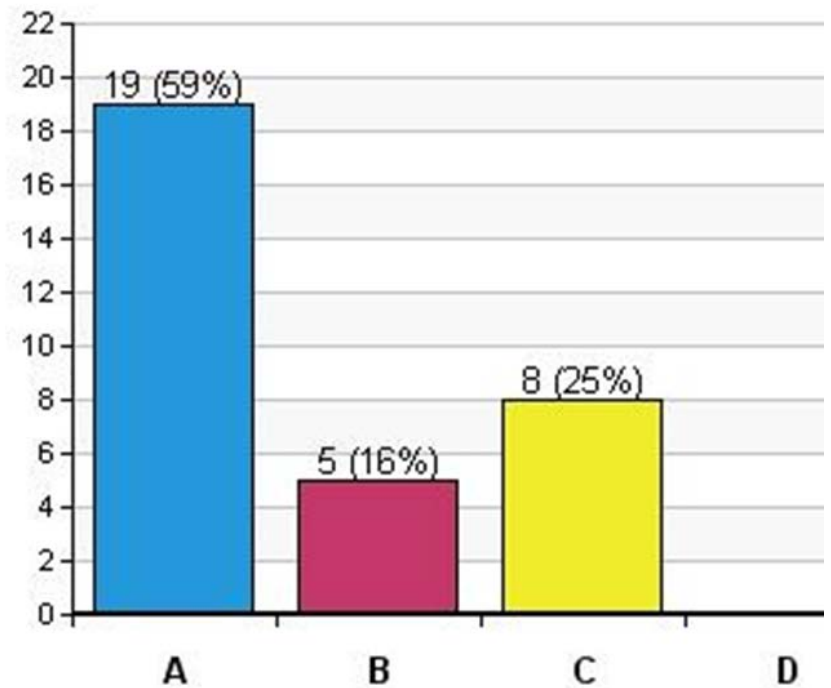
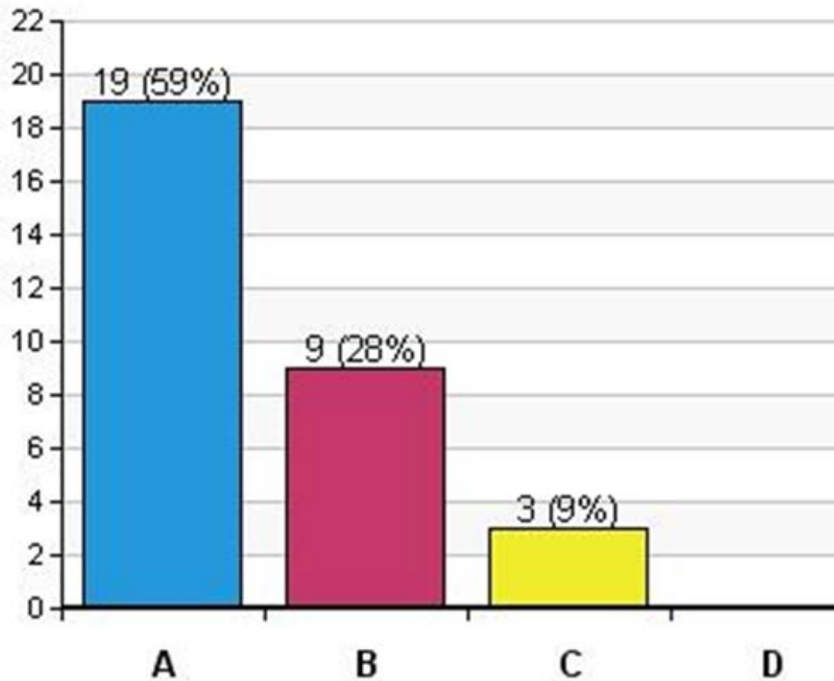
- If right, these arguments, and this evidence, can defeat the debunking arguments of “dual process” and evolutionary accounts.
 - We may well be intuitively equipped for tracking morally-relevant features and responding aptly.
 - Of course, such skills are always imperfectly developed, and liable to the influence of one’s particular situation, interests, inherited prejudices, and so on.
- But this might be enough to lend some credence to the picture of moral judgment found in Aristotle and Hume--the idea that our access to questions about the appropriateness of behavior goes via questions about the kind of agent who would perform the action.

Imaginative “proximity” of potential victims

A = all six B = single man C = the five workers

- **Switch**

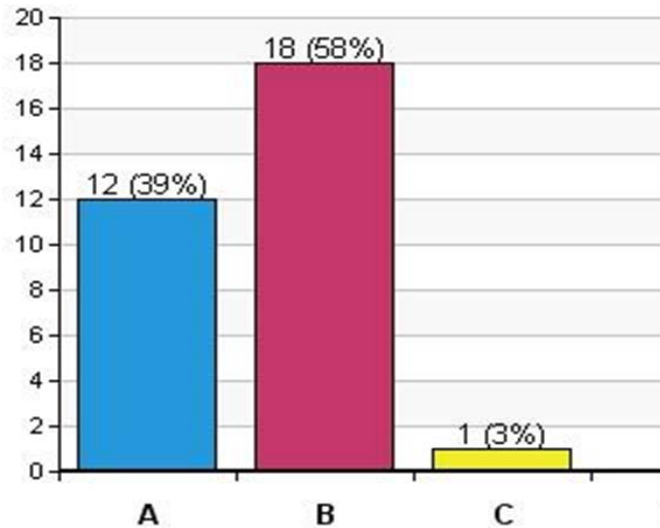
- **Wave**



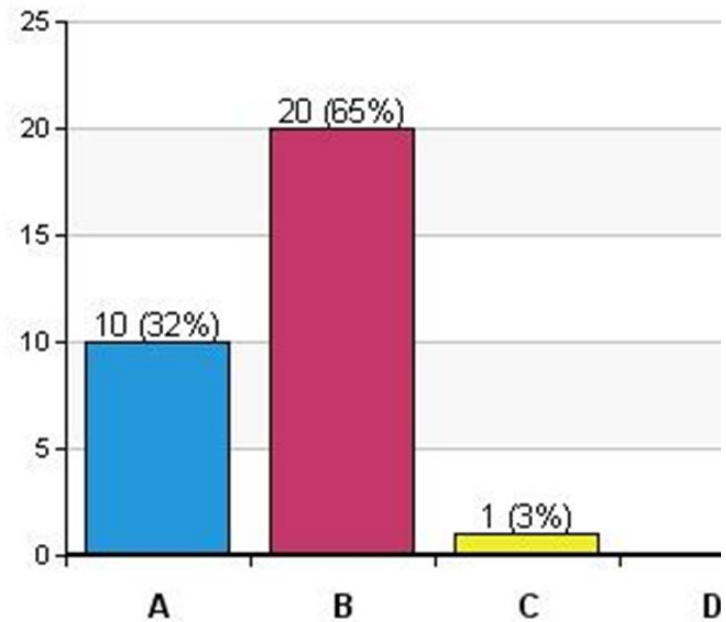
Imaginative “proximity” of potential victims

A = all six B = single man C = the five workers

- **Footbridge**

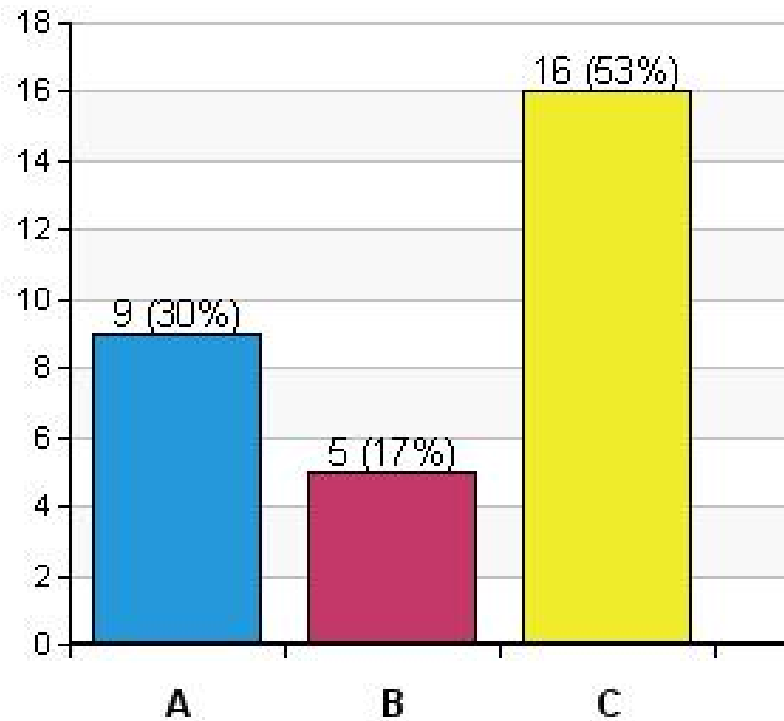


- **Beckon**

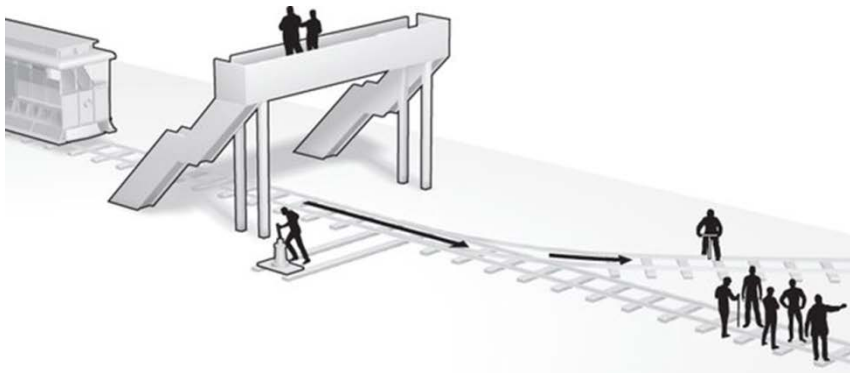


When imaging the Bus scenario, which potential victims seemed to you the most “proximate”

A = all six B = man exiting bus C = people on the bus



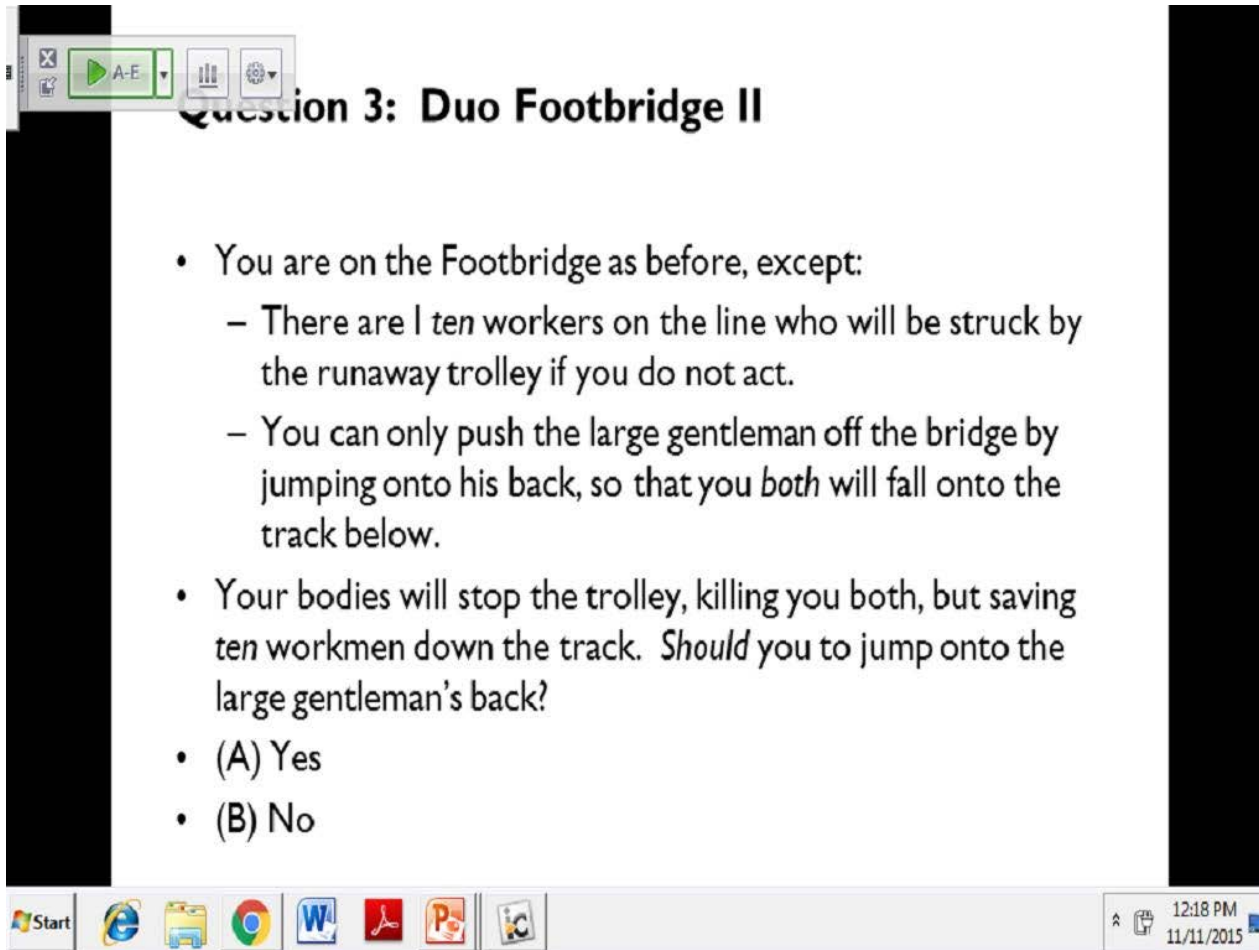
Footbridge vs. bus



We saw ...

- ... in the case of self-driving cars that removing the agent can remove asymmetries and yield stable assessments.
- Can we improve the motivation in Footbridge?

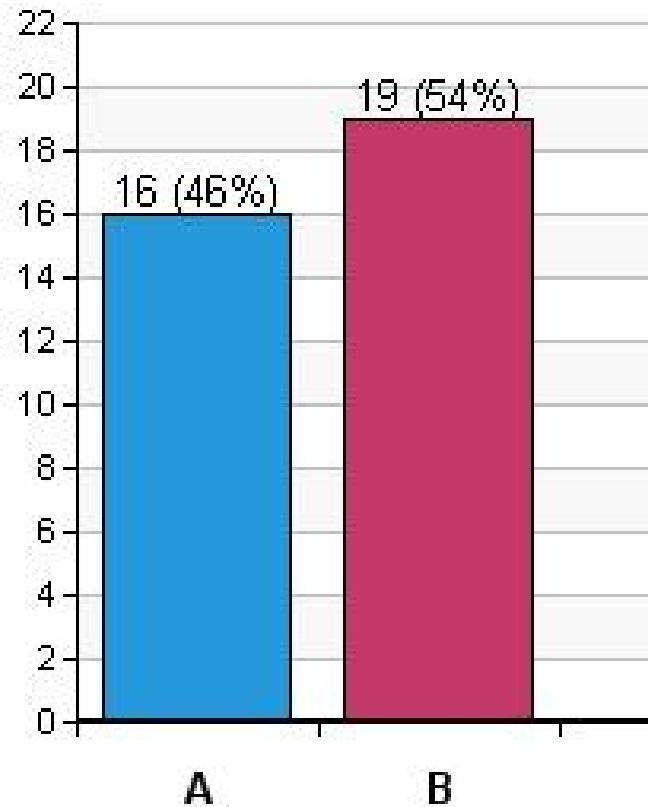
Duo Footbridge



Question 3: Duo Footbridge II

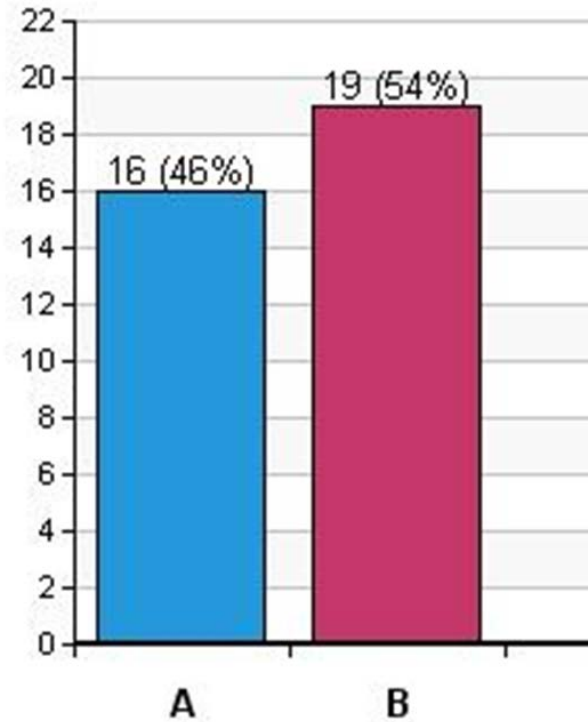
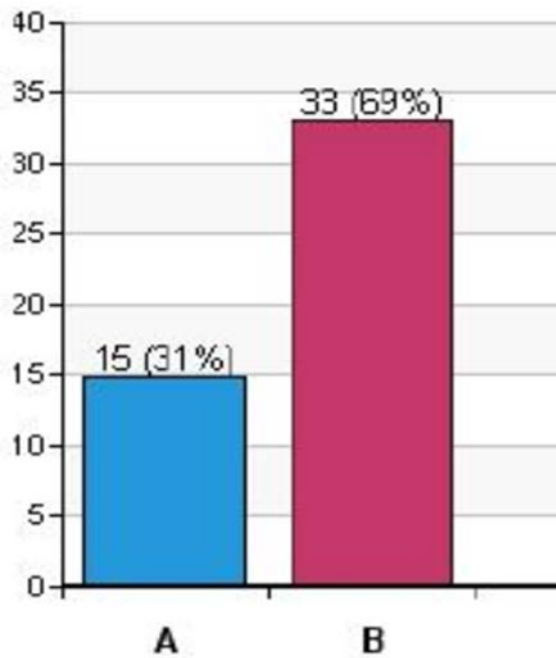
- You are on the Footbridge as before, except:
 - There are *ten* workers on the line who will be struck by the runaway trolley if you do not act.
 - You can only push the large gentleman off the bridge by jumping onto his back, so that you *both* will fall onto the track below.
- Your bodies will stop the trolley, killing you both, but saving *ten* workmen down the track. *Should* you to jump onto the large gentleman's back?
- (A) Yes
- (B) No

Should you jump on the back of the large gentleman, so that you both block the trolley?



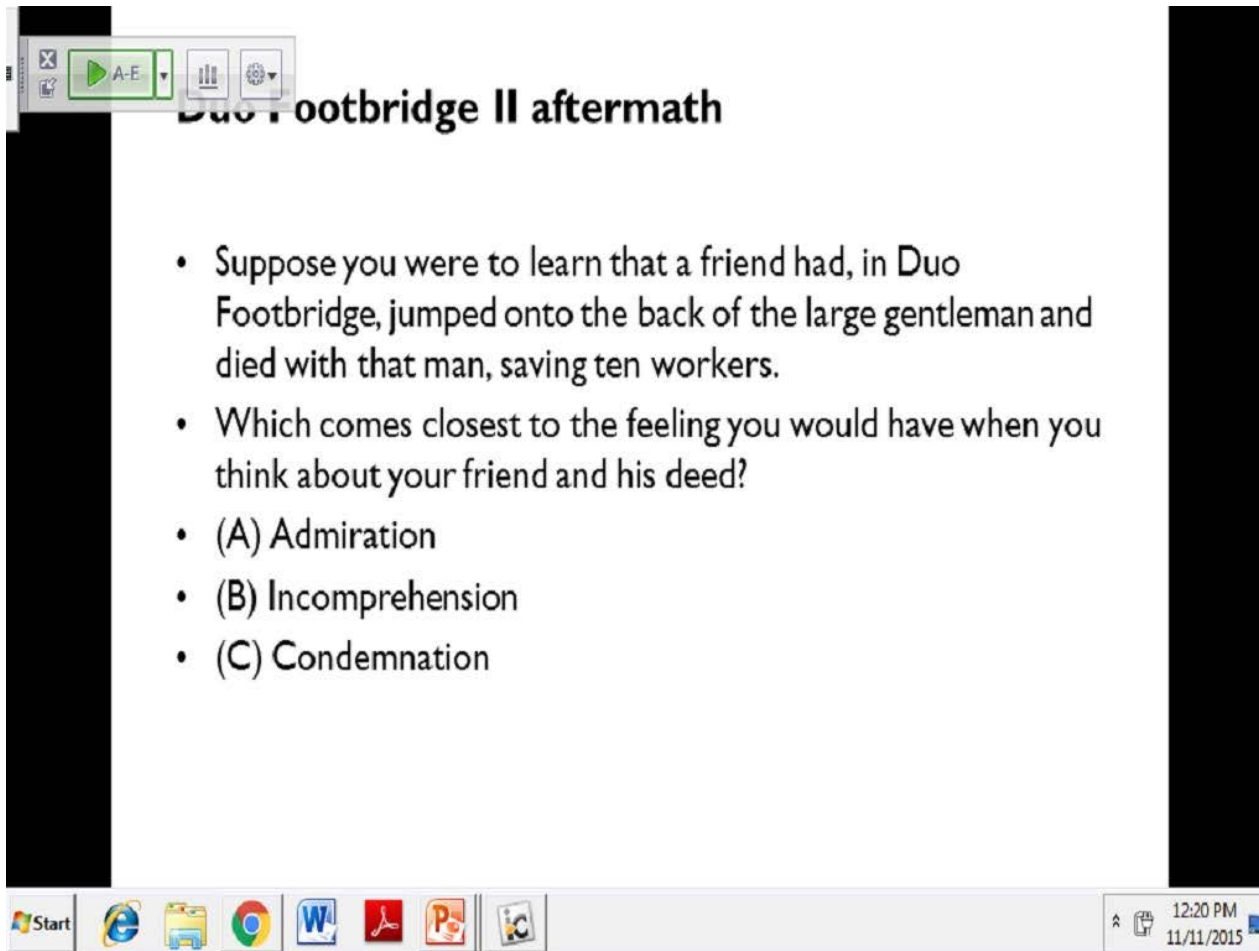
Solo vs. duo Footbridge

A = push B = don't push



Is this a matter of *optimal* underlying character?

In the aftermath of Duo Footbridge



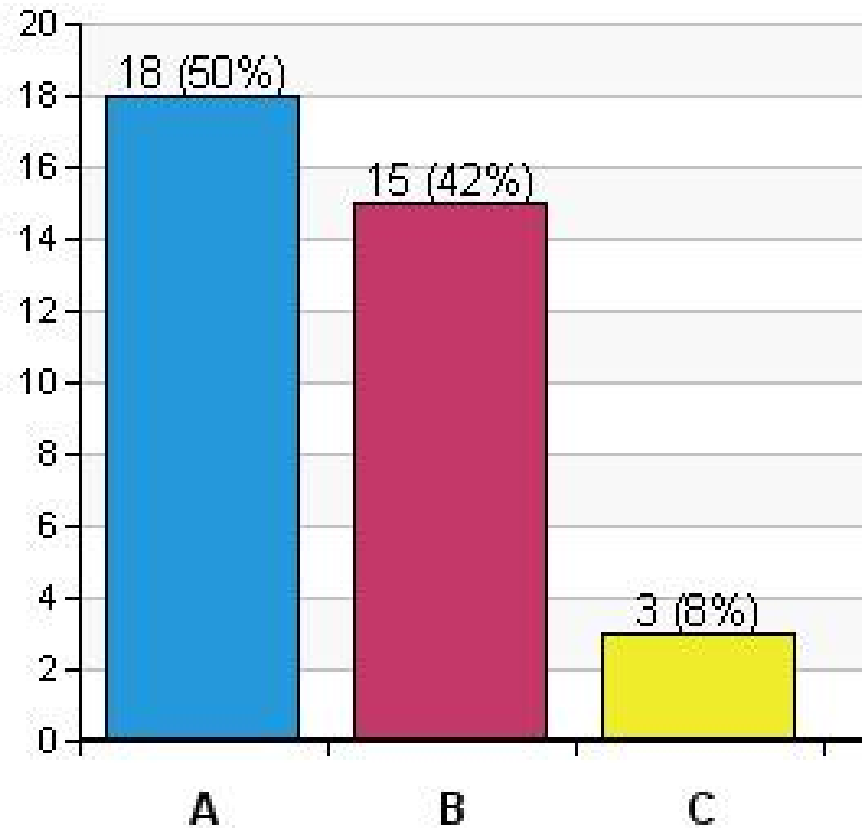
Duo Footbridge II aftermath

- Suppose you were to learn that a friend had, in Duo Footbridge, jumped onto the back of the large gentleman and died with that man, saving ten workers.
- Which comes closest to the feeling you would have when you think about your friend and his deed?
- (A) Admiration
- (B) Incomprehension
- (C) Condemnation

The image shows a presentation slide within a Windows 7 desktop environment. The desktop background is black. At the top, a window titled 'Duo Footbridge II aftermath' is open, displaying a bulleted list of questions and options. The taskbar at the bottom contains icons for Start, Internet Explorer, File Explorer, Google Chrome, Microsoft Word, Adobe Reader, and PowerPoint. The system tray on the right shows the time as 12:20 PM on 11/11/2015.

In the aftermath of Duo Footbridge

Admiration Incomprehension Condemnation



The effect of character ...

- ... on judgments of moral appropriateness or inappropriateness of actions may be further mediated by a certain sense of what one can *reasonably* expect of others, or what one would *condemn* others for omitting to do.
 - I won't pursue further the kind of normative theory this might be, except to note that there could be *consequentialist* or *virtue-theoretic* versions.
- Where does this leave *deontology*? If that means: a theory of *duty* or *duties*, then there can be a theory of duty or duties *within* virtue theory or consequentialism
- If that means: a theory where *rules* are at the bottom, then even Kant won't qualify—at the bottom is not rule-following “legalism”, but the unconditional value of a good will and our capacity to treat others as ends.

(8) The importance of explicit deliberation

This mention of duty, however, brings us to another critical dimension of moral competence:

- ... our capacity for explicit deliberation and normative guidance.

Of mice and men

- If we are able to respond to reasons in ways animals cannot, then it is *not* because:
 - ... we represent options and ends abstractly
 - ... we consider choices prospectively, looking at alternatives and weighing their advantages and disadvantages,
 - ... we follow to a good approximation norms of rationality in revising our expectations and choices.
- Animals can do all of this, and more. They can explore the world and its prospects mentally as well as physically, and choose in light of a weighing of competing goals.

But such intuitive understanding has inherent limitations

- Distinctive of humans is the introduction of new concepts and practices, the inheritance of such innovations through explicit instruction, culture, and the systematic development of more accurate and reliable forms of assessment that permit us to cooperate on a scale unprecedented in the animal world.
 - As we saw in the case of logic and mathematics: these can be understood, following some hints in Wittgenstein, as artificial standards to which we can conform our thought in order to achieve a much more accurate and general representation of the world.
 - Similarly for the development of standardized measures, laws, scientific methods, ... and principles of justice.

Of course, we must have “wittes skile”

- We would have no hope of handling complex questions of justice or policy without the introduction of measures and rules.
- Recall the three families of normative concepts.
- Logic, rules, metrics, laws, and so on belong to the family of *regulatives*.
 - If we are to use them to achieve *evaluatively* worthwhile ends, and improve our *deliberative* capacity.
- As we saw in the case of explicit rule-following, we are not forced into a dilemma:
 - Either posit “blind” dispositions to follow a rule
 - Or launch a regress in which a rule is needed to apply the rule.

A third way

- As the children who resisted authority tell us, we can be *intelligent* in our obedience to rules, without regress,
 - ... since what guides us is not a rule, but a skill that embodies very substantial knowledge and understanding.
 - Even at age 4.
- As any traveler knows, humans have remarkable skills in all three of the normative families, and use them fluently in deciding how or when or in what manner to act.

Thus we conclude for now ...

- ... our project of building from more basic elements the complex human capacity for responding aptly to reasons for action.
- But ...

(9) ... where does this leave us meta-ethically?

Internalism and motivation

- *Motivational judgment internalism* has been the most fundamental source of cleavage among meta-ethical views for nearly a century.
- *Motivational judgment internalists* about moral judgment hold that there is a necessary, conceptual connection between judging (say) that *x is morally right* or *x is morally good* and being in some degree motivated favorably toward *x*.
 - This is thought to capture the idea that the agent's moral judgments must have *practical force* for the agent.
- Non-cognitivists and expressivists consider it a decisive advantage for their view that they can capture this “internal” connection between moral judgment and motivation, since they hold that the state of mind expressed in moral judgments is a motivating one, and not a mere belief (cf. Gibbard 2003).

Beliefs and desires

- In *Ruling Passions*, for example, Simon Blackburn argues that “eighteenth century [philosophy of mind] got it right” in dividing mental states into cognitive, Apollonian states and passionate, Dionysian states.
 - Dionysian: *Emotions (passions, arousals, etc.) and desires (impulses, whims, lusts, urges).*
 - Apollonian: *Attitudes (stances, etc.), and representations (knowledge, truth, reasons).*

Moral knowledge?

- Regarding the possibility of moral knowledge, Blackburn writes:
 - “There is an insuperable obstacle to keeping ethics under the rule of Apollo. Suppose we think our ethics is entirely exhausted by our beliefs. What then? Even the most magnetic star does not attract everyone. Beliefs do not normally explain actions: it takes in addition a desire or concern, a caring for whatever the belief describes.” (90)
 - “The practical role ... is what ethics is for. If there is such a thing as ethical knowledge, it is a matter of knowing how to act, when to withdraw, whom to admire, more than knowing that anything is the case.” (1)

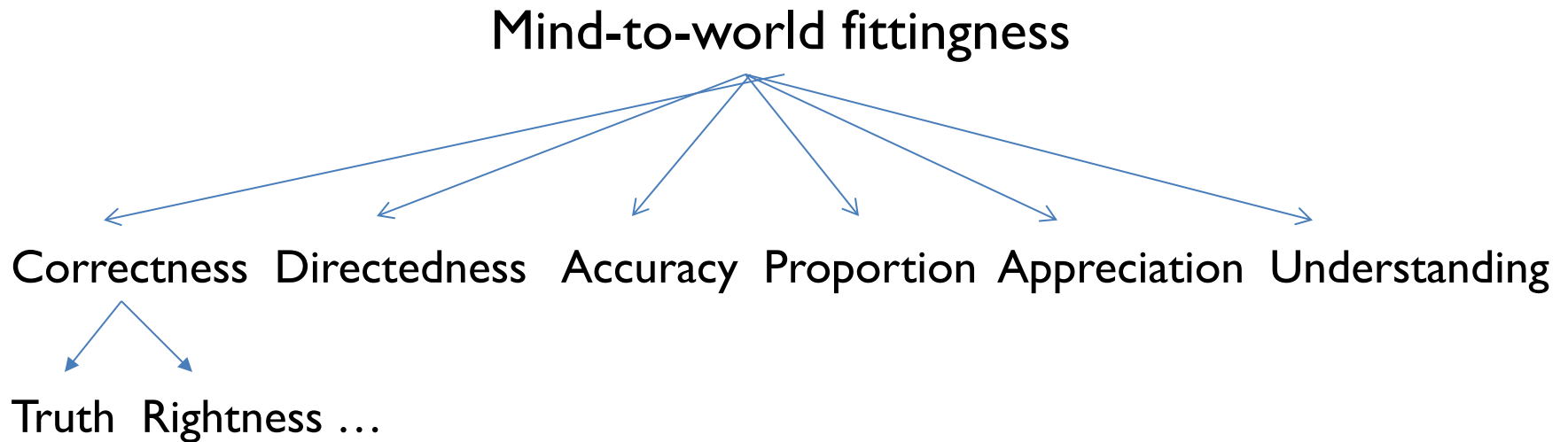
However, ...

- ... we have seen that *affective states*, which are by their nature motivating or action-guiding, can have mind-to-world direction of fit.
 - Affective states *represent* the world in certain ways, and thus constitute forms of cognizing it.
 - *Fear* is a paradigm. It presents the world as possessing certain risks or threats, can be more or less accurate, well-directed, reasons-responsive, and so on.
- Fear does indeed help us to “know how to act, when to withdraw”, and so on, and so can be action-guiding, but it can also be knowing in the ways cognitions can: more or less accurate, misdirected, disproportionate, etc. For Aristotle’s courageous man, experience-calibrated fear helps him *know* danger: *identify* its sources, *appreciate* its magnitude, and *understand* what it is to be at risk.

Moral knowledge

- Doesn't knowledge require *truth*?
- It requires some notion of *getting things right* or *appreciating something for what it is* or *understanding its nature*.
 - And affective states can qualify in all these ways as more or less knowing.
- And on the account offered here, affective states such as belief involve *expectations* that admit talk of *correctness* or *mistakenness* as well. Thus we can be *mistaken* in our fear, our anger, or our degree of confidence or belief.

There are multiple dimensions of mind-to-world fit



Moral knowledge and practical force

- It thus is unwarranted to conclude that moral knowledge is not possible from the fact that moral judgment has practical force: moral knowledge is a matter accurate, well-directed, proportional understanding and appreciation of moral reasons.
 - Motivation alone is in any event a poor proxy for the practical, *normative* force of moral judgment. In itself, as critics have argued, motivation is not normative.
- However, we have seen how compound states that combine affective attitudes with the regulation of action-tendencies, such as desire and belief, can provide a *fitting* recognition of value and reality, and shape our practice accordingly.

Two kinds of judgment

- The motivational internalists are partly right: the kinds of attitudes that constitute our moral perspective, and that are expressed in our moral judgments, are typically motivating.
- But this is a matter of the nature of a moral perspective, or of moral knowledge (if that term be allowed), not a conceptual truth about the meaning of moral concepts.
 - Thus, it is not a *linguistic* mistake to make a moral judgment in the absence of corresponding motivation—e.g., affective disorders that affect motivation need not, and in themselves should not, change our moral views.
- Thus, an unadorned judgment that “x is wrong” can be made, and made sincerely, in the absence of motivation.

Two kinds of judgment

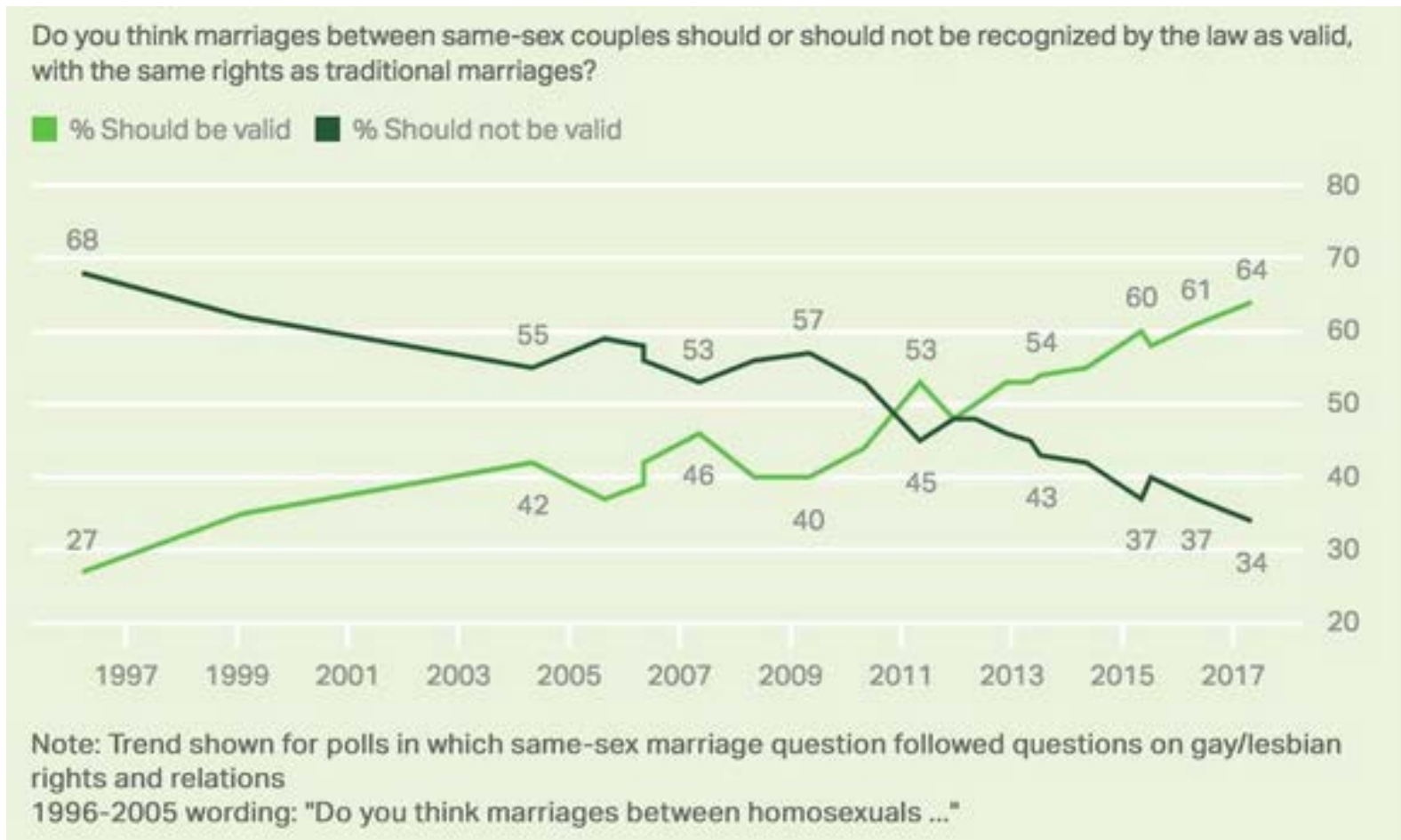
- This is also what Kant meant in saying that one could make a merely *theoretical* judgment of permissibility or impermissibility by applying the categorical imperative test,
- ... but we also require a *practical* moral judgment, expressing an *affective* attitude (the “moral feeling”) that can motivate through *an appreciation* of the value of a good will, or *respect for* others as ends. The categorical imperative is morally compelling to us, and action in accord with it has moral worth, insofar it is a way of expressing this appreciation or respect in action—not simply because it is a form of rational consistency.
- This distinction is vital for moral *change*. (Or aesthetic—for Kant, the closest analogy. E.g., for Thoreau at Walden Pond—appreciation could proceed, and lead, judgment.)

Two kinds of judgments

- Or consider the massive “natural experiment” that took place when millions of gay individuals made their sexual orientation publicly known.
 - At the beginning of this period, a great majority of people in the US thought that allowing gay couples the right to marry was *undesirable* and *wrong*.
 - But many in this majority group discovered that people they know and admire are gay. This enabled them to *appreciate* that being gay was not a moral flaw, even before they changed their “official” judgment. (A kind of inverse inference involving their models of character as a source of behavior.)

Legal recognition of same-sex marriage, US, 1996-2017

(Gallup, 2017)

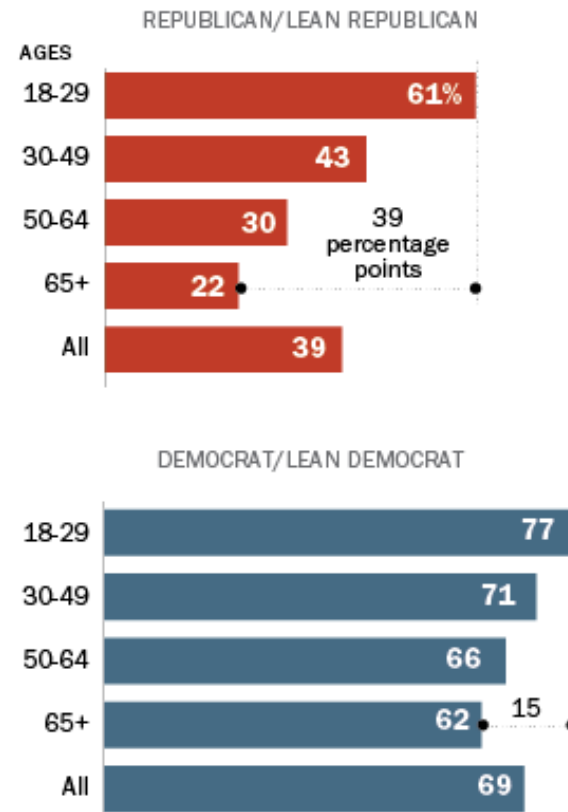


Political identity vs. personal experience

(Pew Trust, 2014)

Most Young Republicans Favor Same-Sex Marriage

Percent who favor allowing gays and lesbians to marry legally

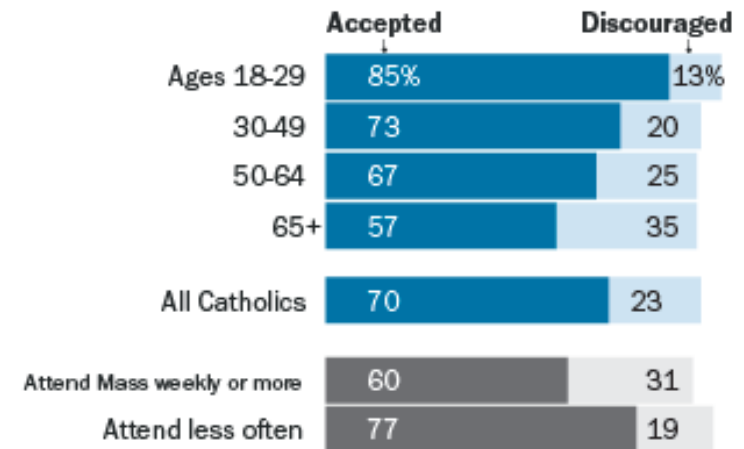


Inherited religion vs. personal experience

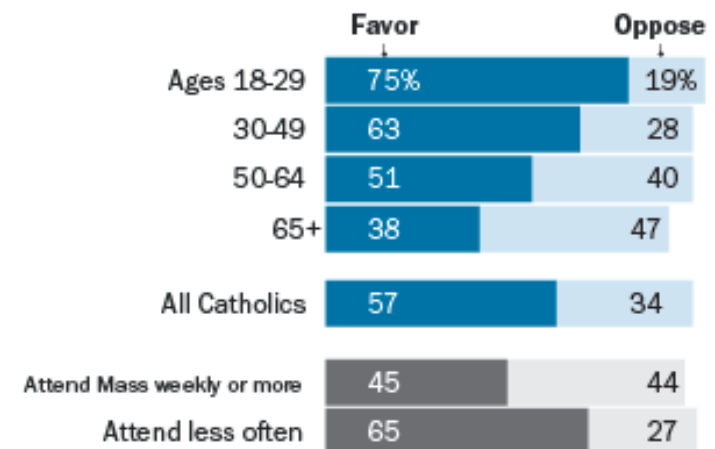
(Pew Trust, 2014)

U.S. Catholics' Views on Homosexuality, Same-Sex Marriage

Homosexuality should be ...



Support for same-sex marriage



Moral learning

- May be a powerful source of personal development
- ... and also social change.

- Prejudice against gays was millennia old, and backed by the full authority of most organized religions and even legal codes.
- Some evolutionary psychologists argued that such prejudice is “in our genes”.
 - Yet more basic learning processes—perhaps involving empathic simulation of others—could challenge and change this prejudice, even within the span of a lifetime.

Moral learning and moral realism?

- These cases make clear why the view under consideration is not a form of *subjectivism*.
 - The reasons for change in moral views are grounded, not in attitudes, or even ideal attitudes, but in the facts about human life that would ground and explain these changes in attitude.
- The ground is, like the ground of all value, *subject-involving* or *subjectual*. But it is a set of objective facts about such subjectual questions.
- In *The View from Nowhere*, Nagel argued our distinctive normative situation reflects the intersection of subjectivity and objectivity—this is but one example.

Subjectual but objective

- Consider for example the view in “Moral Realism” and “Facts and Values” according to which our best evidence (typically) of what is good for a person is what that person would, if fully informed and widely experienced, desire to desire for the circumstance of being in one’s actual shoes.
 - The ground here are those relational facts that would explain the particular second-order desires, not the desires themselves.
- The view is therefore *subjectual* and *relational*—like facts about nutrition—and not opinion-independent and standpoint-independent.

Naturalism and non-naturalism

- Nothing here that violates naturalist strictures,
 - ... but also there is nothing that non-naturalists who recognize the need for a plausible psychology would need to object to.
- Normative concepts can be part of job descriptions satisfied by natural properties, as I have tried to argue here for the case of the concept, <apt responsiveness to reasons>.
 - Far from trying to *replace* this concept, I have been trying to make the world safe for this normative concept—to show how something we do might satisfy it.
- Thus is a division of labor possible.



To Derek Parfit (1942-2017)