

# **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 5:

“Moral Intuitions, Moral Judgment, and Moral Agency”

Peter Railton

(University of Michigan)

Oxford, May 2018

# **Discussion seminar tomorrow morning**

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

# Today's shameless appeal to authority:

## David Hume

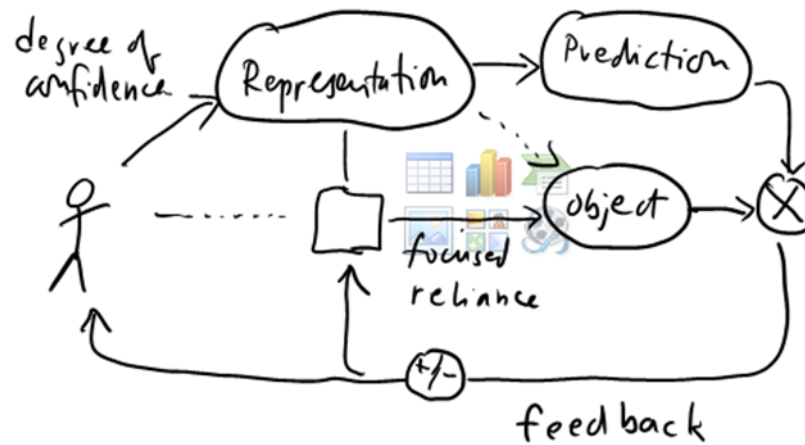
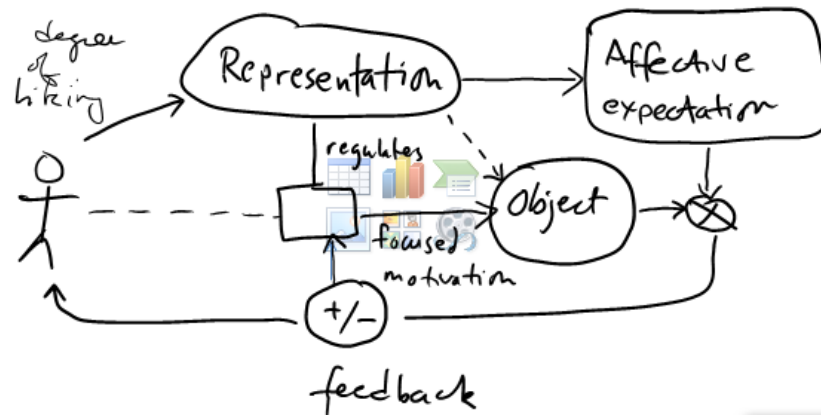
- “... every particular man has a peculiar position with regard to others; and 'tis impossible we cou'd ever converse together on any reasonable terms, were each of us to consider characters and persons, only as they appear from his peculiar point of view. In order, therefore, to prevent those continual *contradictions*, and arrive at a more *stable* judgment of things, we fix on some *steady* and *general* points of view; and always in our thoughts, place ourselves in them, whatever may be our present situation.” (SBN 581-82)

# **(I) Skill and intuition**

## In attempting to build ...

- ... a unified account of rationality in desire, belief, and action,
- ... I have emphasized how much the basic states of desire and belief can contribute to our capacity to be rational in the wide sense of being aptly responsive to reasons.
  - Desire and belief are, in effect, “intelligent regulators” of our dispositions to attend, perceive, recall, infer, and act.
  - As we have seen, intelligent regulators of such processes construct *models* of the systems they regulate—in this case, the physical and social world and its prospects and perils, as well as our own resources and capacities to act.

# Desire and belief



# Models

- Such models can be understood as network-like representations that use experience to build up relations of informational and causal relevance.
  - Simplest form: *<target values; costs; if  $\rightarrow$  then projections; feedback; updating>*.
  - More complex: open to learning with respect to target values and costs; hierarchical and abstract in representational structure.
- Capacity and cost permitting, there is an inherent push toward hierarchical and abstract representation, since these support greater predictive and generative scope and power.

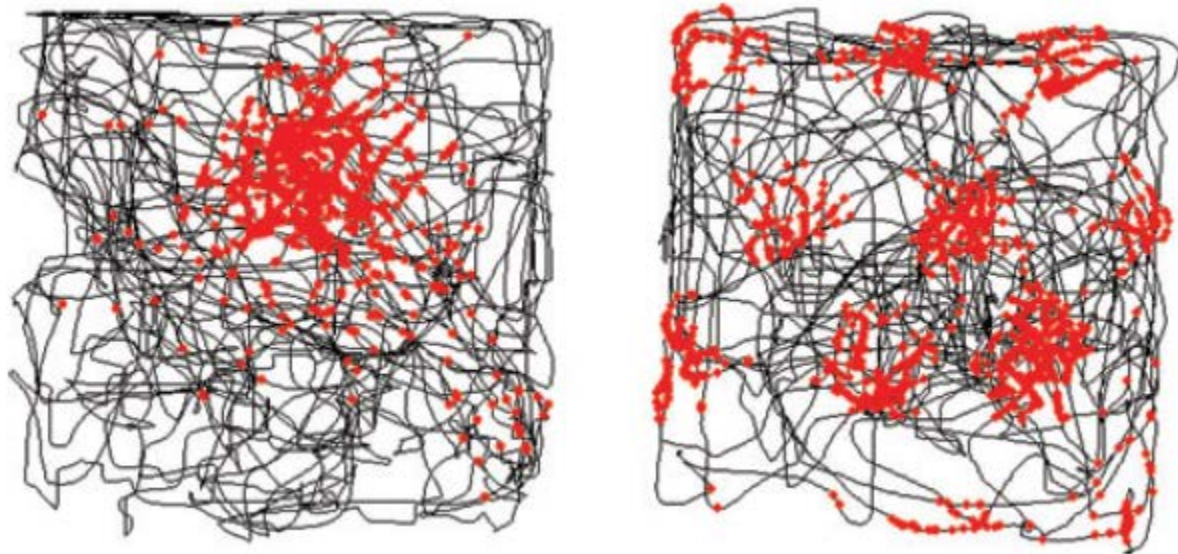
# Models

- These models can constitute “practical representations” in the sense that they are used in both a *forward* and an *inverse* manner to guide action.
  - *Forward*: Because they model the individual, her range of possible action, her aims, and the environment, they can generate not only predictions, but actual guidance of behavior.
  - *Inverse*: They permit focused adjustment and learning from feedback about outcomes.
- Some examples we’ve discussed:
  - Visual field and eye movements
  - Off-line and on-line simulation possible actions and choice

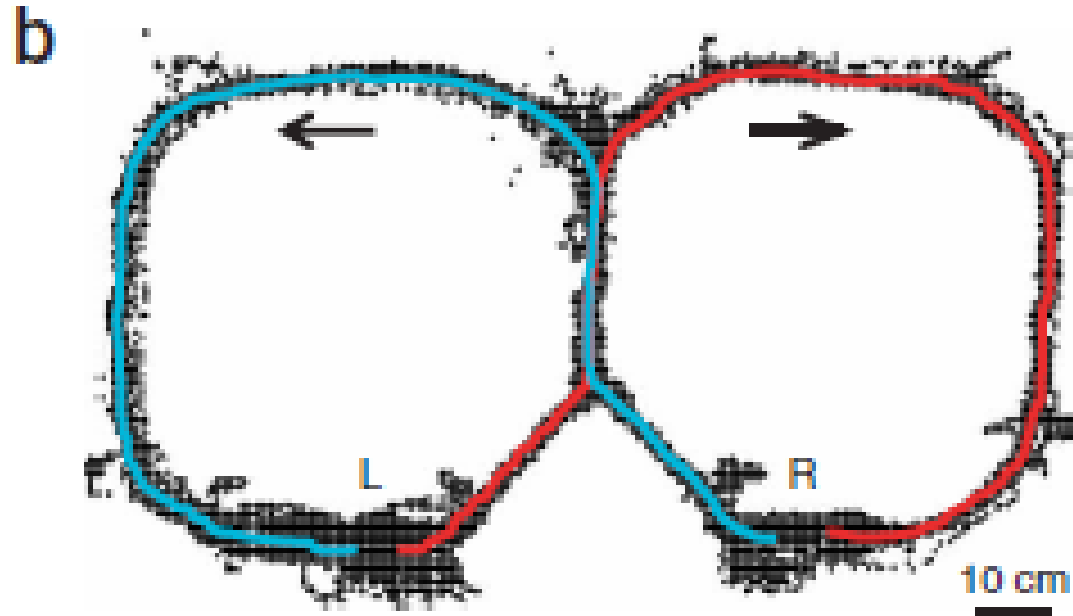


# Recall: “Allocentric” mapping of space

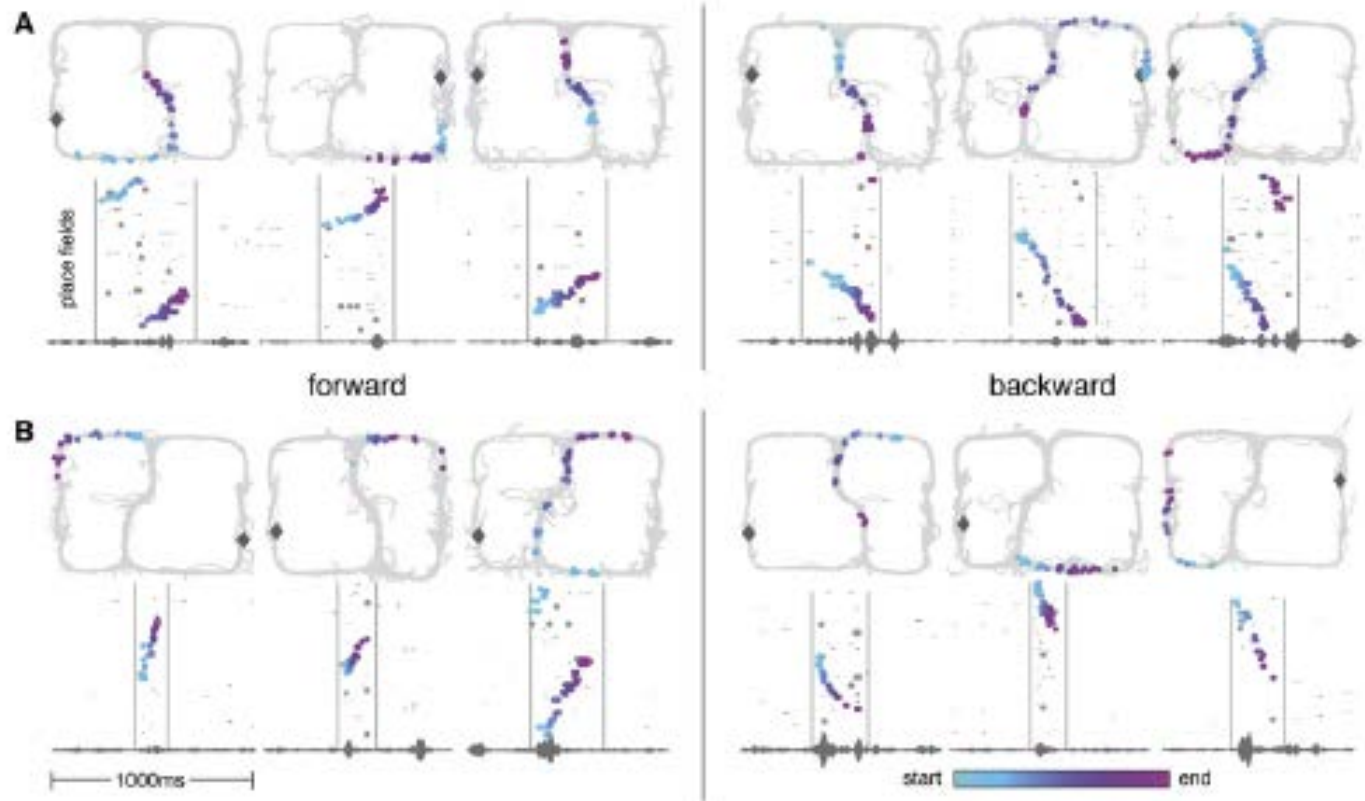
(Moser et al., 2008)



# Co-ordinated replay of map during sleep (Ji & Wilson, 2007)



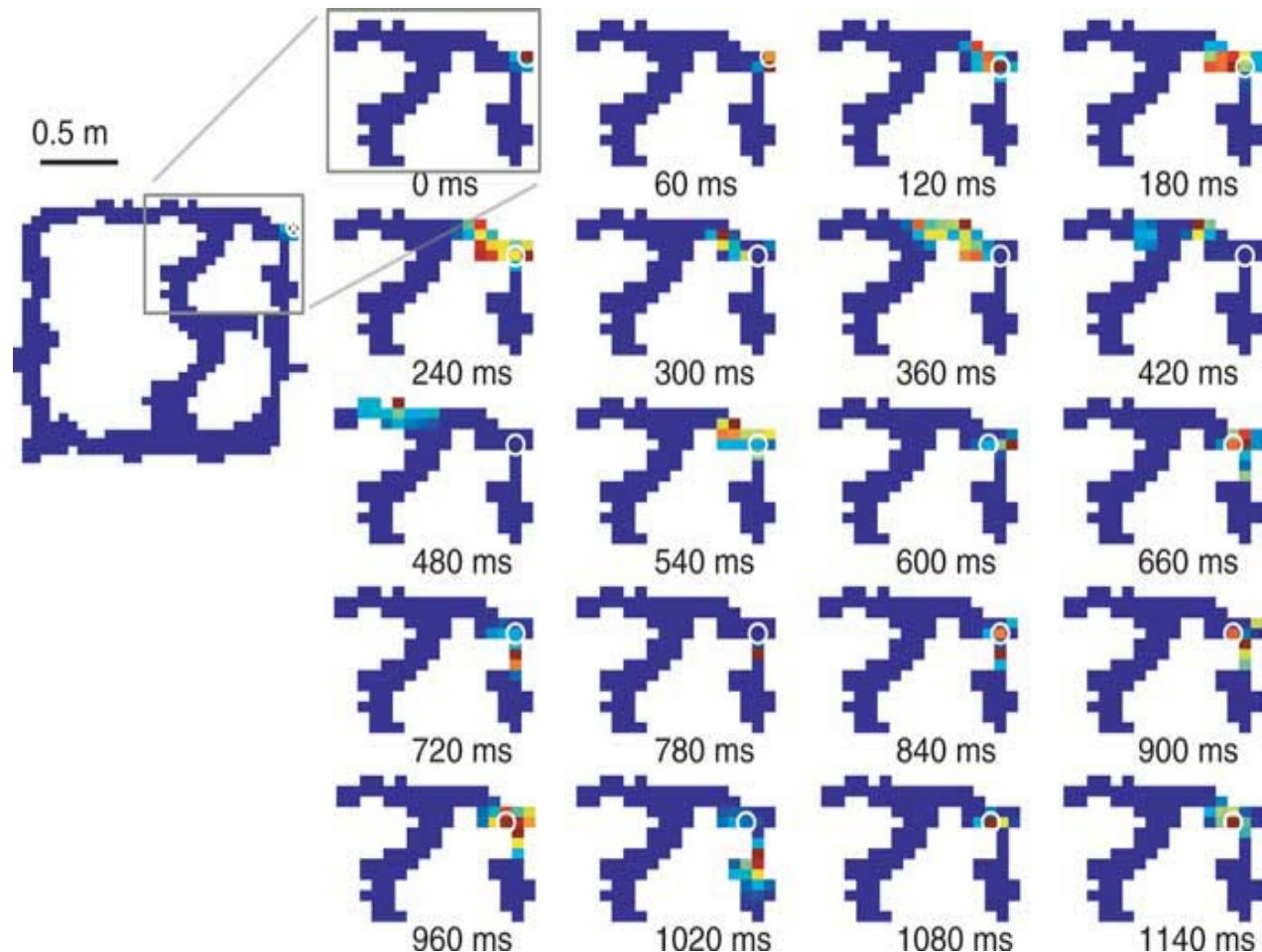
# Construction of novel paths in sleep (Gupta et al., 2010)



Examples of Forward and Backward Replay

# A rat following an evaluative representation

(Johnson & Redish, *J Neurosci* 2007)

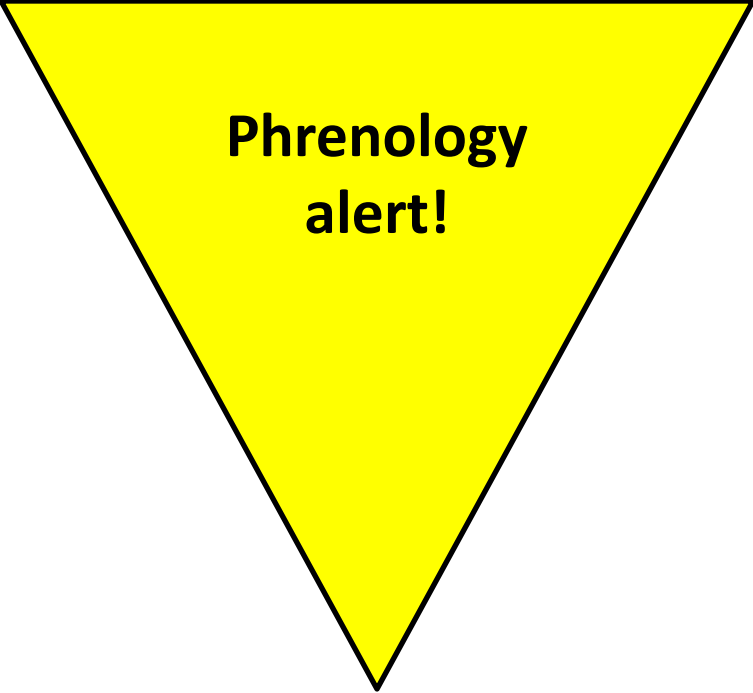


## And we've seen ...

- How the model-based, prospective character of desire and belief enables us to understand *skill with reasons and reasoning*.
  - In Lecture 3, to explain the possibility of having intelligent dispositions—shaping what one notices, what one calls to mind, what options one considers, and so on—in order to act intentionally in response to reasons for action, without needing to form a prior intention to do so.
  - Or, in Lecture 4, similarly intelligent dispositions to recognize situations as apt for applying and self-consciously following a rule without needing to follow a rule in order to do so.

# Prospective simulation

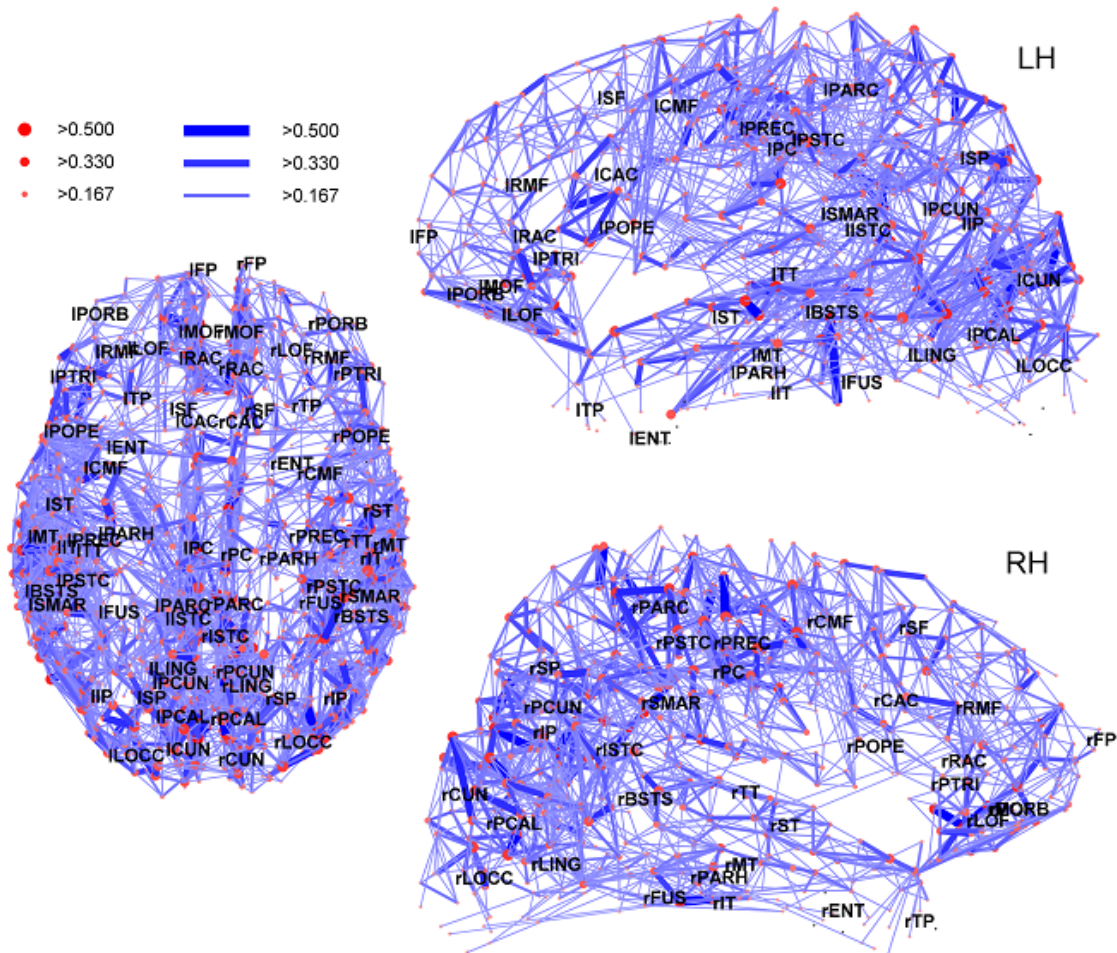
- The prospective simulation of the evolution of physical and social environment, and of one's potential actions and their likely outcomes, is sufficiently important in intelligent animals to be a recurrent, central activity of the mind, supported by a large-scale functional network, the default mode.

A large yellow equilateral triangle pointing downwards, centered on the page. It has a black outline.

**Phrenology  
alert!**

# Connectomic view of mind

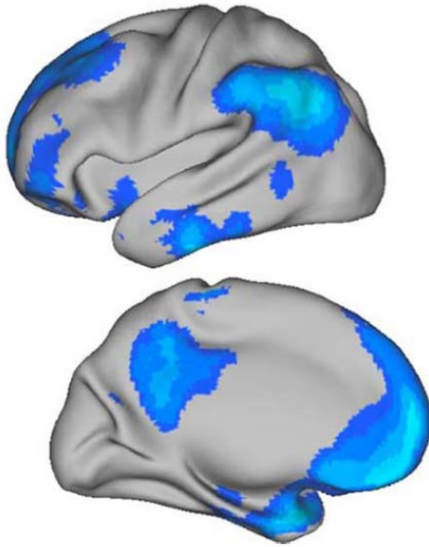
(Hagman *et al.*, 2008)





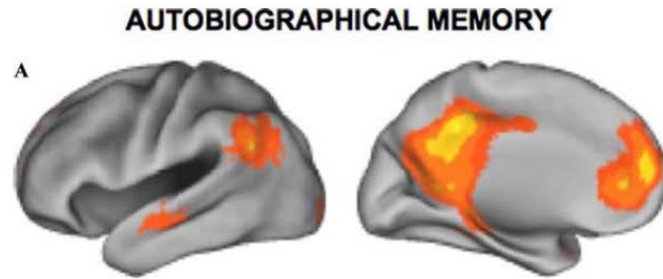
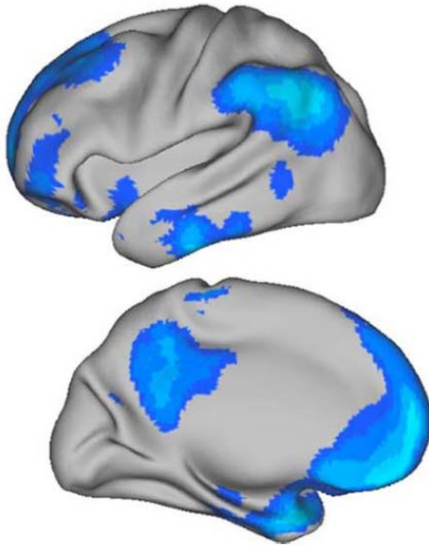
# Default network

(Buckner *et al.*, 2008)



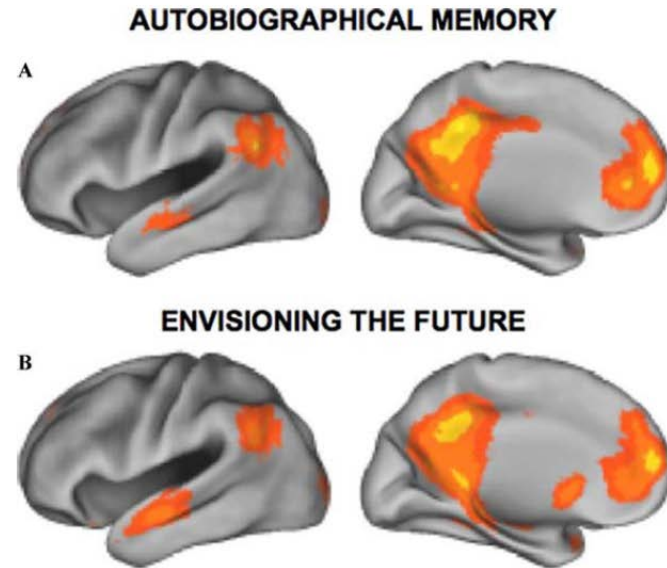
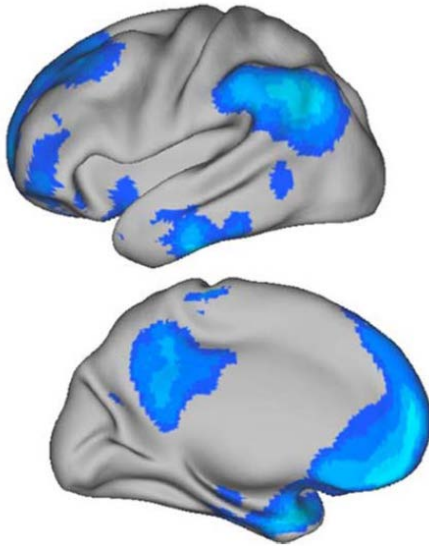
# Default network

(Buckner *et al.*, 2008)



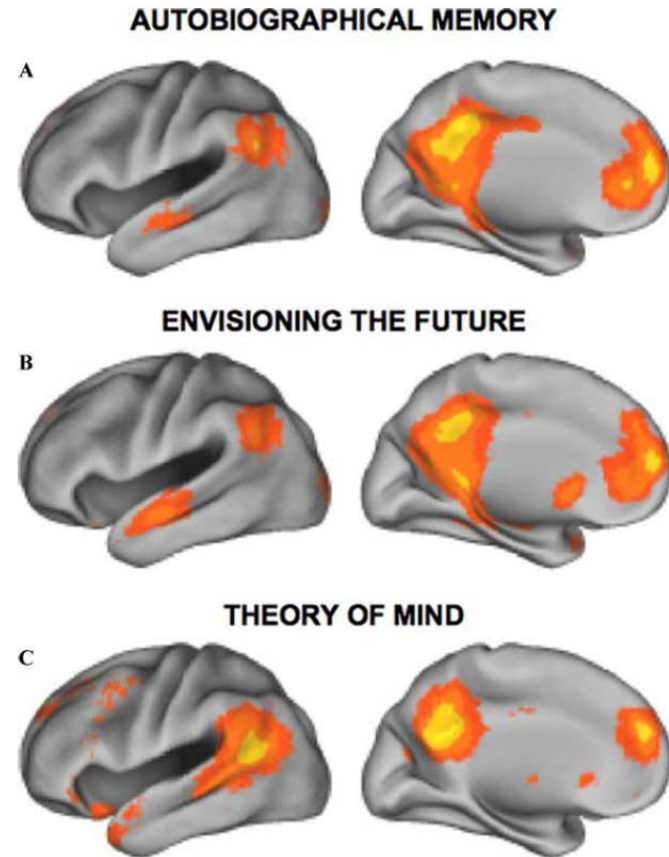
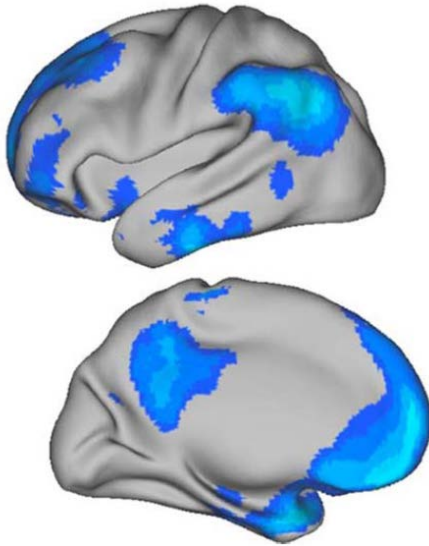
# Default network

(Buckner *et al.*, 2008)



# Default network

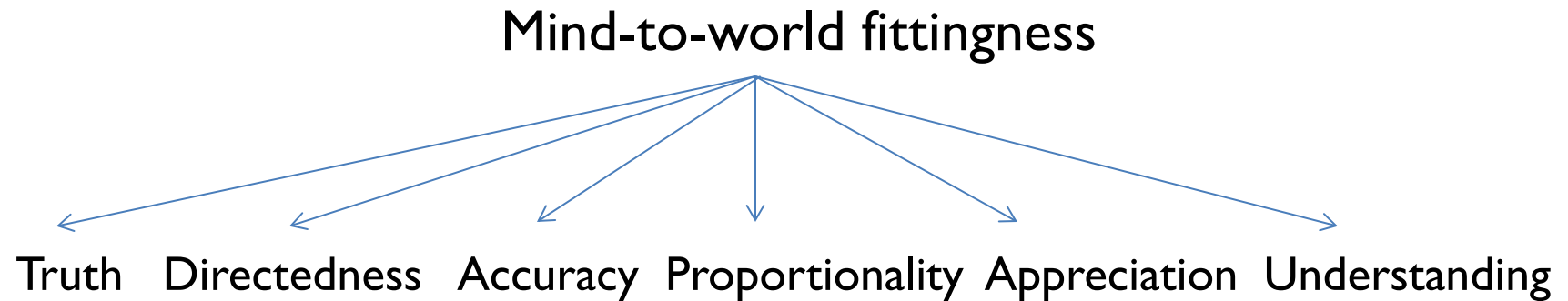
(Buckner *et al.*, 2008)



# Skill and intuition

- Skills thus involve extensive representational structures that can guide attention, perception, thought, and action in situation- and goal-appropriate ways,
  - ... *fluently*, without needing to call upon self-conscious deliberation or inducing the interference or regress this would involve. (Compare fluency in speech.)
- Skill is not “muscle memory” or “fixed action patterns” that have become habits—it permits “spontaneous” adaptation to changing and novel contexts and challenges via complex representational structures with mind-to-world direction of fit that can guide motor control.
- Such skilled capacities are sometimes called *intuitive*.

# Multiple dimensions of mind-to-world fit



## In this lecture ...

- ... we turn to the moral case, and start with the question of whether our moral judgments might be reflections of similar underlying acquired competencies—intelligent, fluent capacities for responses to situations, actual or hypothetical, of the kind called “moral intuitions”.

# Moral thought and practice ...

- ... rely extensively on intuitions at all levels.
  - Principles
  - Particular judgments
  - Reflective equilibrium
- The classical Intuitionists had a story about the nature of these intuitions, and why they have normative standing—they are self-evident, synthetic *a priori*, rational insights.
  - Yet few now accept this account, and no systematic account of moral intuition has emerged to replace it.
  - So we lack a widely-accepted account of the origin, nature, or authority of moral intuitions.



# Field notes: some of “obvious” features of intuitions in general

- (i) We often find ourselves with a spontaneous “sense” that some thought, action, state of affairs, etc. is plausible or implausible, right or wrong, trustworthy or dubious, dangerous, not working properly, etc. This sense:
- (ii) ... does not appear to require explicit, effortful, or controlled reasoning or judgment
- (iii) ... often emerges “immediately” in an actual situation, or in considering a hypothetical situation,
- (iv) ... typically arises non-voluntarily;
- (v) ... can be recalcitrant in the face of contrary *judgment*;

# Field notes: some of “obvious” features of intuitions in general

- (vi) ... yet is experienced as in some degree compelling or motivating;
- (vii) ... so that we are reluctant to give it up or ignore it;
- (viii) ... even when we cannot articulate a fully satisfactory explanation or justification for it.
- (ix) Moreover, intuition can spontaneously guide thought or action over time—think of a musician improvising—without need for deliberate planning, decision, or endorsement
- No doubt there are other features, but these I hope are relatively uncontroversial *descriptively*. (I don’t want to *presuppose* that our intuitions are *right* or *privileged*.)

# Field notes: What are *moral* intuitions? – Some truisms

- Moral intuitions have been subject to a range of psychological and philosophical critiques. A common form of these critiques is to say that, while intuitive moral judgment is a genuine phenomenon, it can be seen to be systematically responsive to *morally irrelevant considerations*, and unresponsive to *morally relevant considerations*.
- This requires a rough, consensual idea of what might count as morally relevant. Here are a few truisms to get started:
  - Moral considerations should be general, independent of parochial interests or perspectives, independent of sanction, and linked to considerations of cooperation, interpersonal trust, well-being, and respect for persons.

## Among the sources of challenges:

- *Evolutionary psychology*: Would natural selection have favored the evolution of cognitive and motivational systems of *Homo sapiens* capable of responsiveness to general consideration of the well-being of others, without parochialism?
- *Historical and social variability* in what is taken to be “intuitive”—dependence upon non-moral considerations like hierarchy.
- *Persisting disagreements* in intuition about moral questions, with no apparent method of resolving.
- “*Dual-process*” theories of the mind and moral judgment.

## **(2) “Dual-process” theories and moral psychology**

# Intuitions in contemporary psychology—“dual-process” accounts

- According to a considerable body of work in empirical research, people have both spontaneous, intuitive reactions and controlled, deliberative reactions when making judgments and acting.
  - The intuitive reactions rely heavily upon implicit “heuristics and biases” with “little understanding of logic and statistics” (Kahneman, 2012).
- The operation of intuitive processes is not introspectably available, so subjects often cannot say which features of a situation they are responding to, or explain sharply different judgments in similar-seeming but differently-framed scenarios.
  - Asked to justify their judgment or conduct, they may “confabulate”.

# Two systems?

- In the strong form that has been influential in discussing moral intuition, the dual-process picture often has taken the form of two *systems* that normally run in parallel:
- **System 1** – Intuitive, largely implicit, evolutionarily ancient
  - Fast, effortless
  - Affect-laden (Haidt, 2006), heuristic-based
  - “Push button” or “point-and-shoot” (Greene, 2006, 2014)
- **System 2** – Deliberative, largely explicit, evolutionarily recent
  - Slow, effortful
  - Draws upon scarce cognitive resources
  - Capable of logical and probabilistic reasoning

# Dual-process accounts and moral judgment

- But we need not insist upon two “systems” as such—it is enough if there are two persistent modes of processing, intuitive and deliberative, that can yield systematically different results.
  - This difference can be used to explain otherwise puzzling or inconsistent patterns in everyday moral judgment.



### **(3) Some well-known examples**

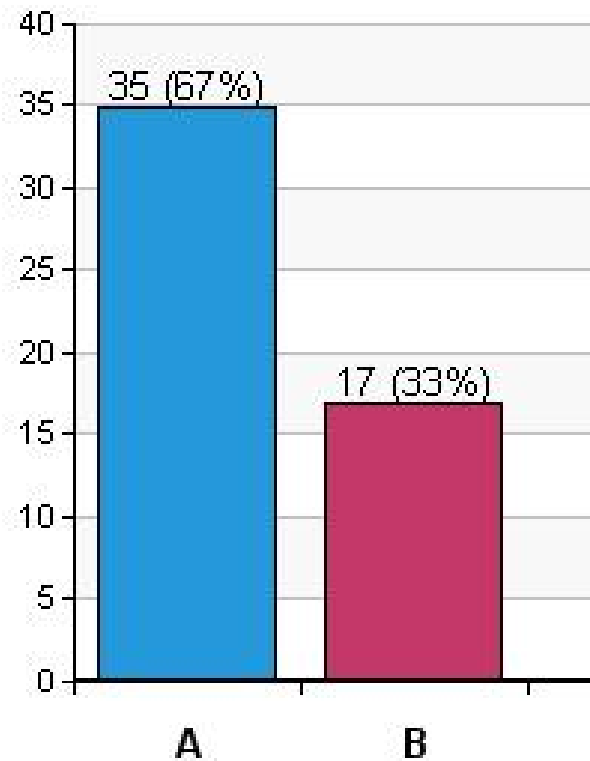
# Trolley Switch

(image from *New York Times*, 9 October 2010)



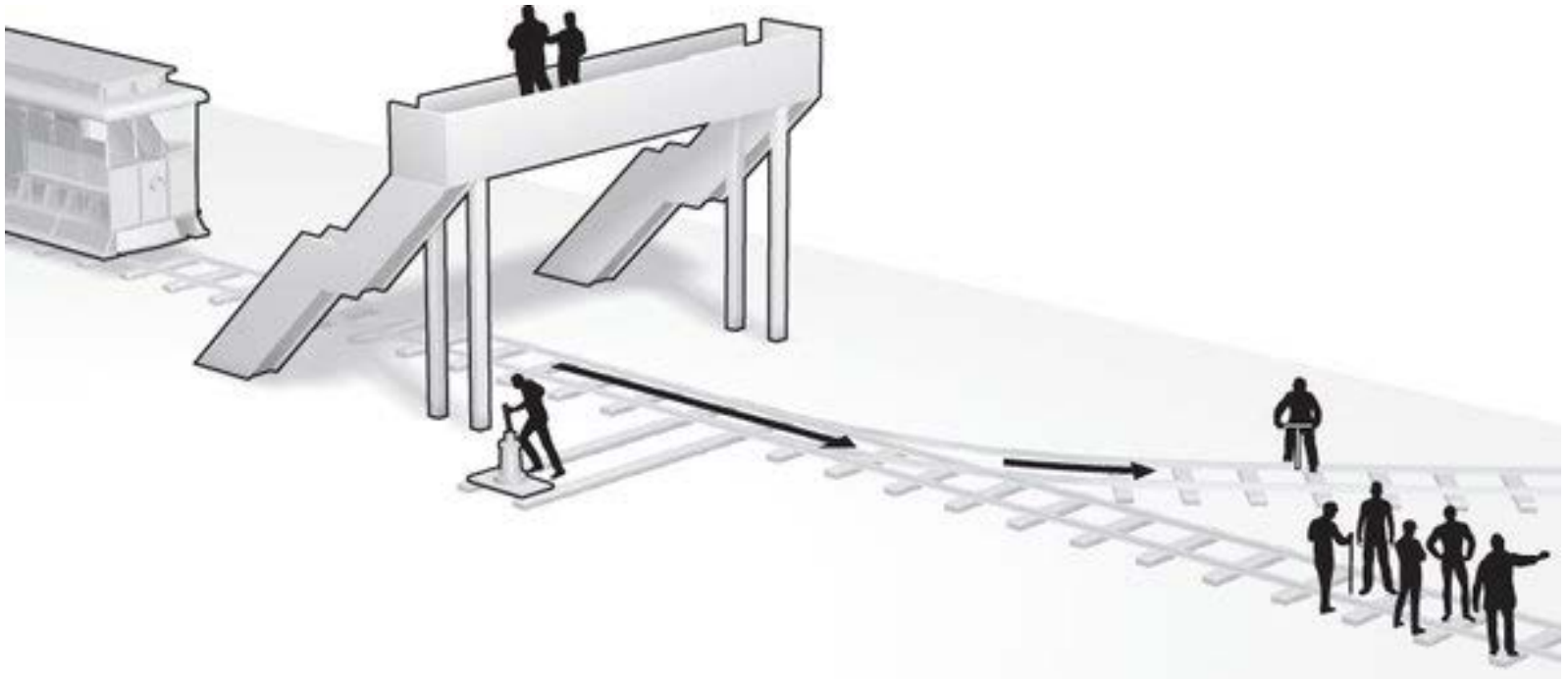
# Trolley Switch

A = Pull lever    B = Do not pull lever



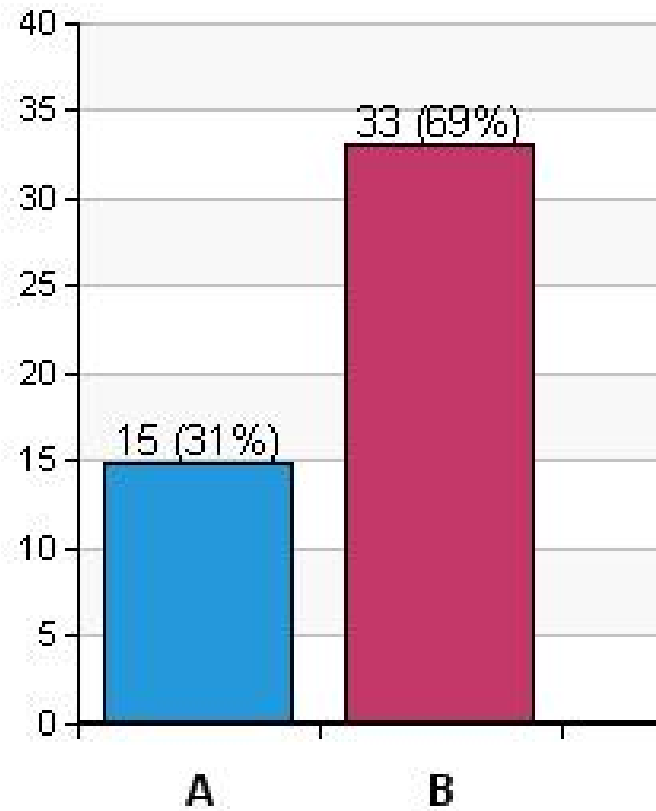
# Trolley Footbridge

(modified from *New York Times*, 9 October 2010)



# Trolley Footbridge

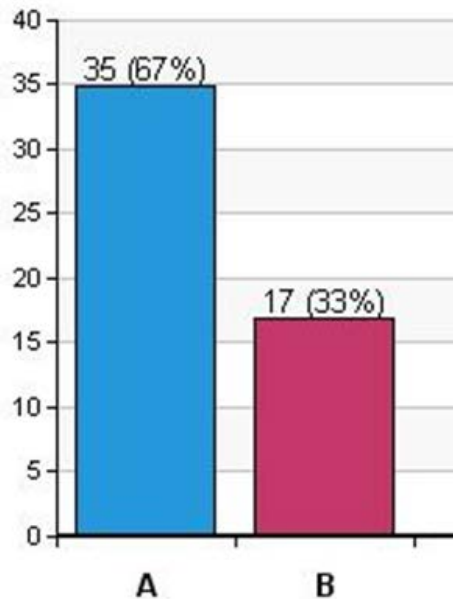
A = Push man    B = Do not push man



# Classic Trolley “asymmetry”

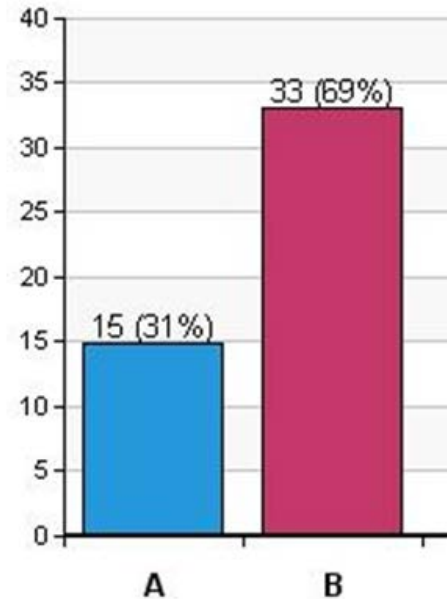
## Switch – Pull?

A = yes    B = no



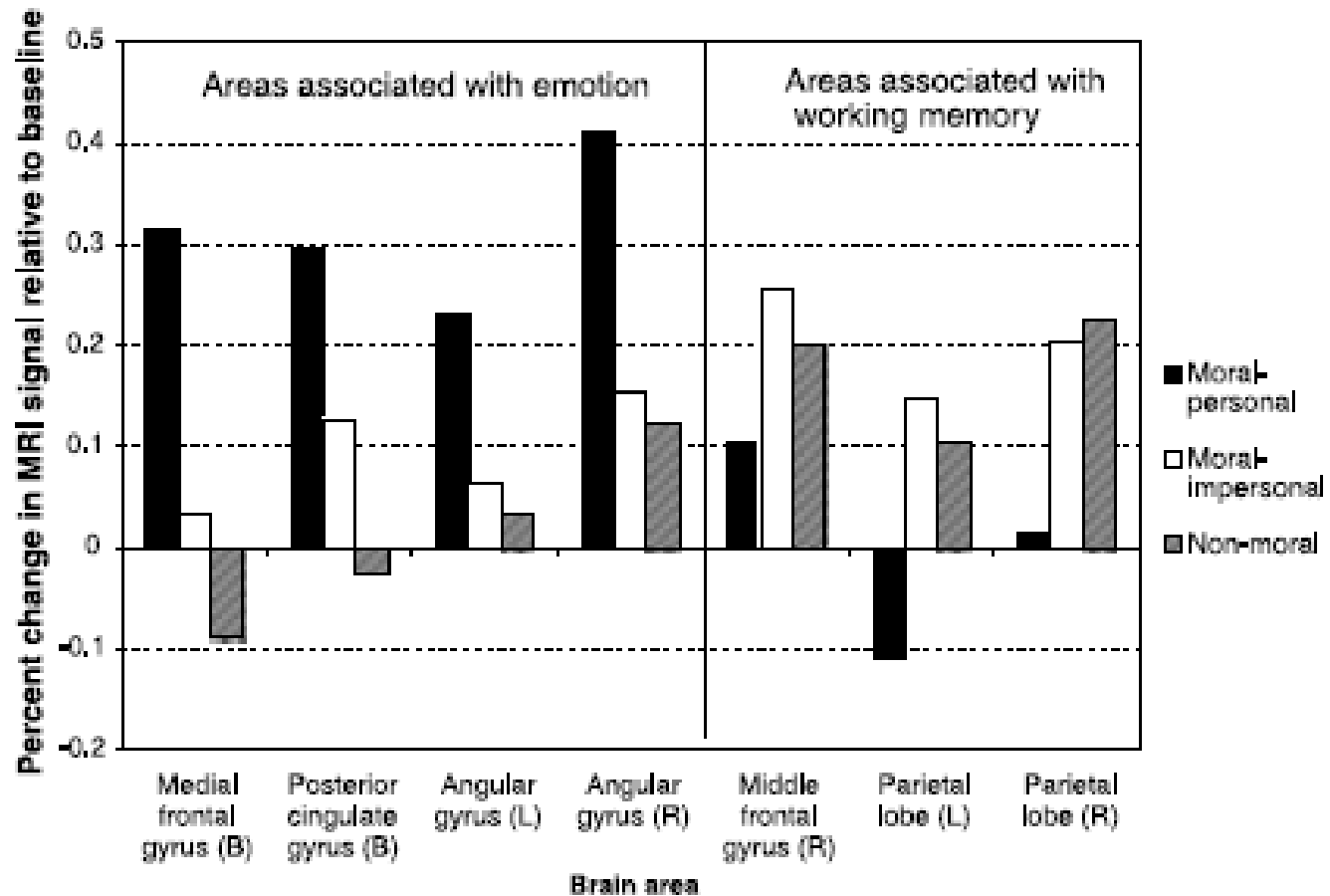
## Footbridge – Push?

A = yes    B = no



# fMRI investigation of moral judgment

(Greene *et al.*, 2001)



# Dual-process trolleyology

- Haidt and Greene (2002) and Greene (2013) propose that in **Footbridge** scenarios, subjects vividly imagine using their own muscular force to harm another, triggering in System 1 a “ME HURT YOU” heuristic, or an immediate aversion, preempting System 2 calculation of costs and benefits
- In **Lever**, imagining throwing a lever lacks this direct application of muscular force upon a victim, so System 1 sets off no affective “alarm bell” and the rule-based, cost-benefit calculation in System 2 predominates.
- In **Lever**, subjects *do* have access to their rationale, and uniformly give the loss-minimizing justification; in **Footbridge**, they have no insight into System 1, so have difficulty articulating any stable rationale, though their “intuition” typically remains firm and dominant.



# Normative relevance?

- Greene (2013) and others argue that the common verdict against pushing in **Footbridge** thus can be seen as reflecting morally irrelevant considerations (e.g., a direct aversion to the use of one's muscular force upon the victim, as opposed to indirect uses of muscular force), and thus should be given less weight than philosophers—especially deontologists--have typically assigned to it.
  - Note that an initially attractive principled explanation of the resistance to pushing, that we cannot use others as mere means, fails prey to examples such as:

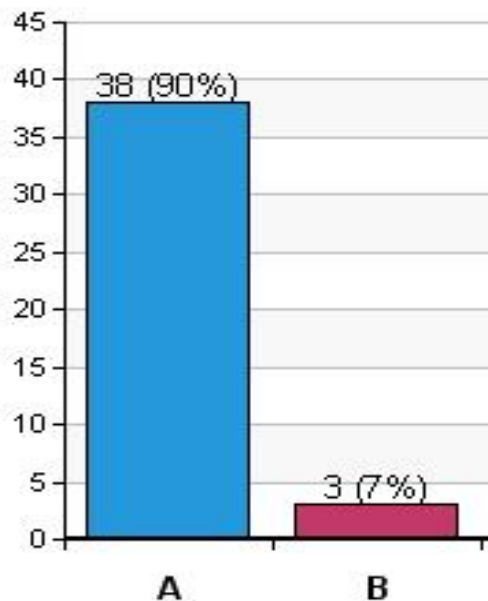
# Trolley Loop



# Loop Trolley “asymmetry”

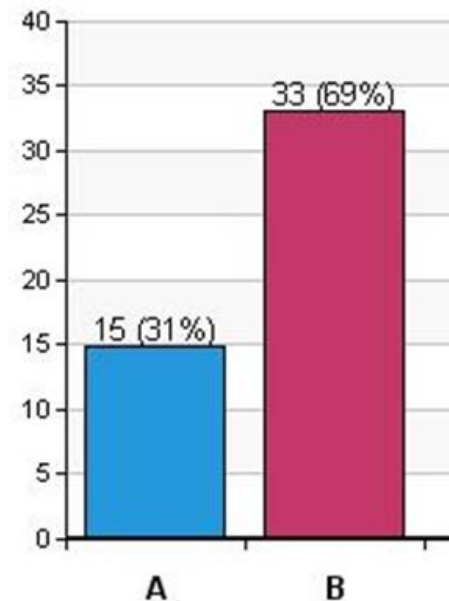
## Loop – Pull lever?

A = yes    B = no



## Footbridge – Push?

A = yes    B = no



## As Greene's account predicts, ...

- ... even though the man on the sidetrack *is* used as a mere means in **Loop**, throwing the switch in **Loop** does not involve direct use of muscular force upon the victim, so the dominant intuitive verdict is closer to **Switch** than **Footbridge**.
  - This suggests that a deontic rejection of using people as mere means is unlikely to lie behind Footbridge.
- The dual-process account also predicts, correctly, that people will have difficulty explaining why their willingness to sacrifice one to save five in **Switch** does not transfer to **Footbridge**. Since people lack introspective access to the sources of intuitive verdicts, it is unsurprising that they are often at a loss to explain the difference between the two cases.

# “Moral dumbfounding”

(Haidt, 2001)

- ***Julie and Mark** are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that, was it OK for them to make love?*

## **(4) My own kitchen chemistry**

# Informal classroom sampling

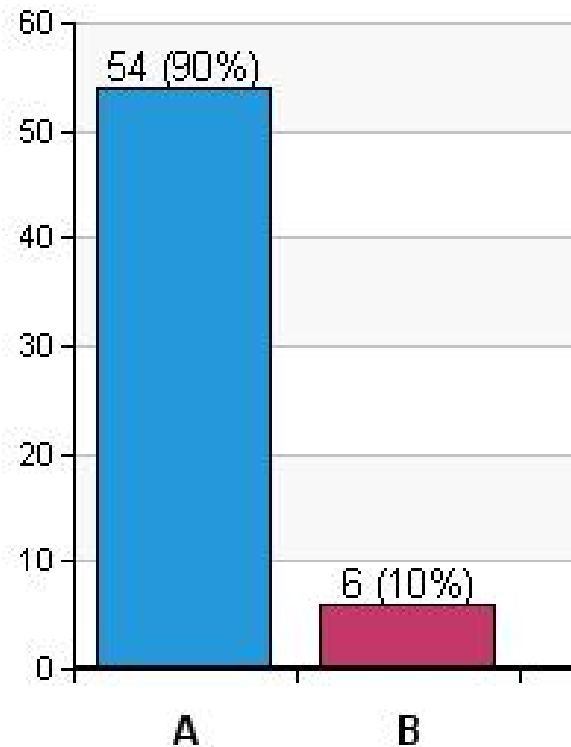
- The use of iClickers permits:
  - Rapid responses
  - Confidential responses
  - Probing beyond the initial scenarios
  - Display of summary responses in immediate aftermath of polling
  - Sampling across time, and in the wake of new information
- These are *not* controlled experiments ...
  - ... though there is some evidence that they accurately reflect what students think, and thus provide a representative sample (Stowell & Nelson, 2007).

**So let's try asking:**



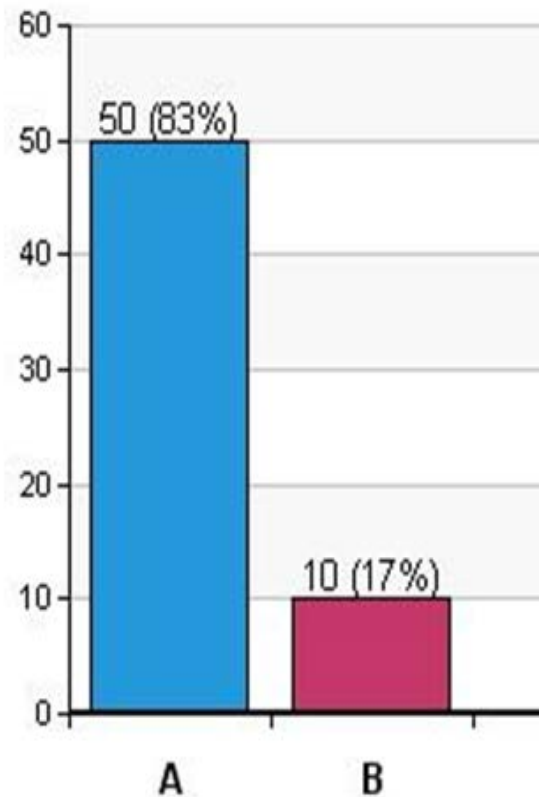
**Is it a moral reason *for* pushing the man from the footbridge that this would result in fewer deaths of innocent people?**

A = yes    B = no



**Is it moral reason *against* pushing the man that you would be directly causing his death, using your own arms?**

A = no    B = yes



## Let's introduce the missing cases

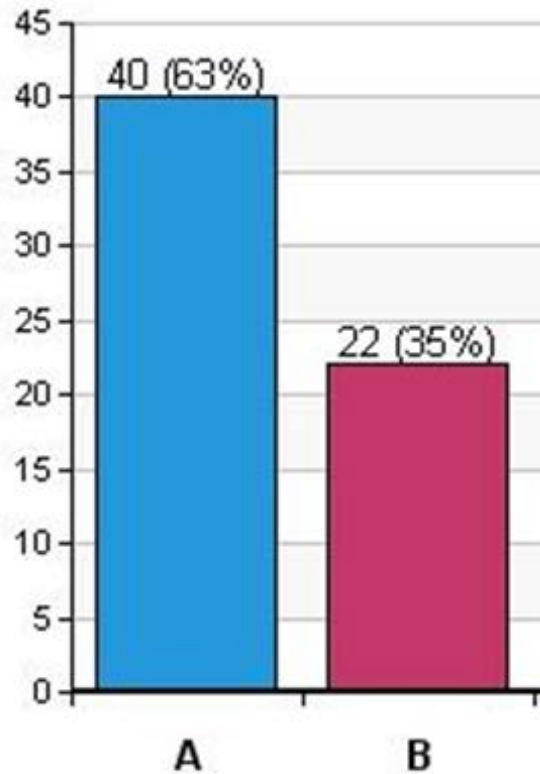
Trolley problem “intuitive” judgments	Intervention harming one to save five <u>should</u> <u>not</u> be done according to most subjects	Intervention harming one to save five <u>should</u> be done, according to most subjects
Use of direct muscular force to inflict harm	Footbridge	X
No use of direct muscular force to inflict harm	Y	Switch, Loop

**X = Bus**



## Bus

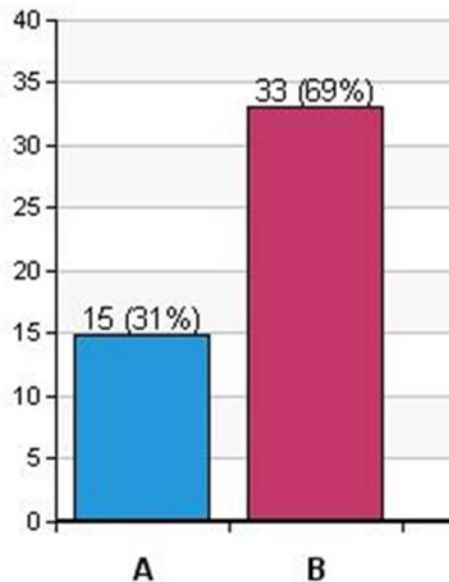
A = Push man    B = Don't push man



# Proximate, “direct muscular” cause of death

## Trolley Footbridge (personal force)

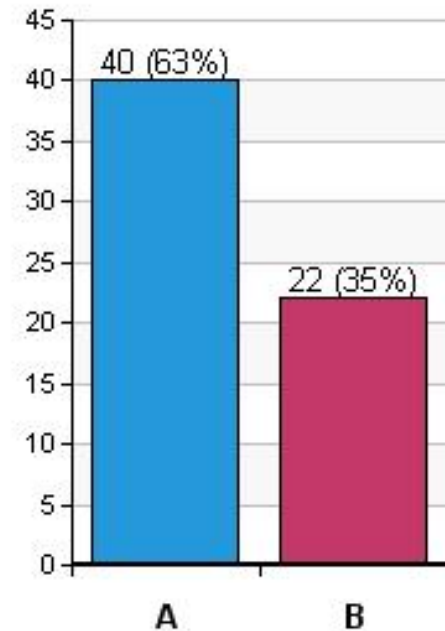
A = push    B = do not push



## Bus

### (personal force)

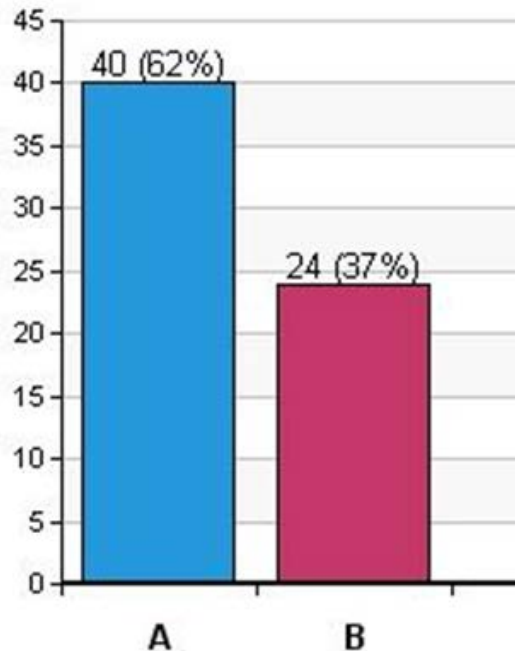
A = push    B = do not push



# Remote vs. proximate cause of death

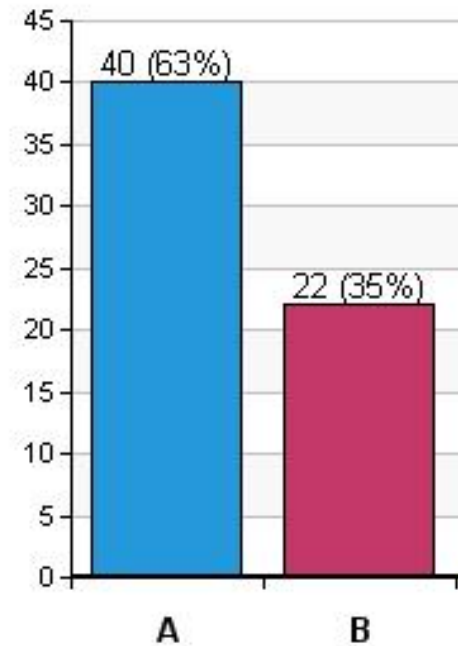
## Trolley Switch (remote)

A = pull B = do not pull



## Bus (direct force)

A = push B = do not push



**Y = Beckon**



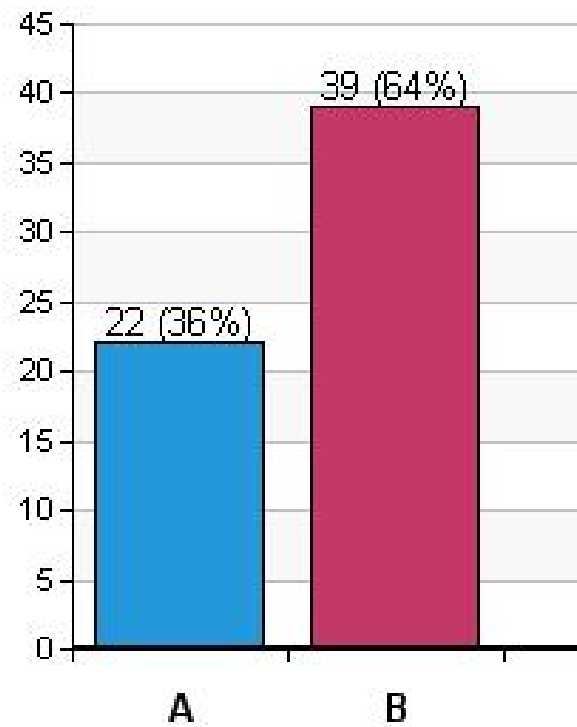


**Y = Beckon**



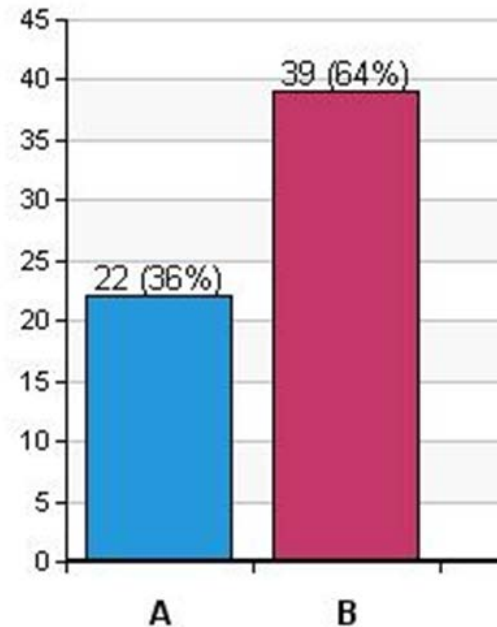
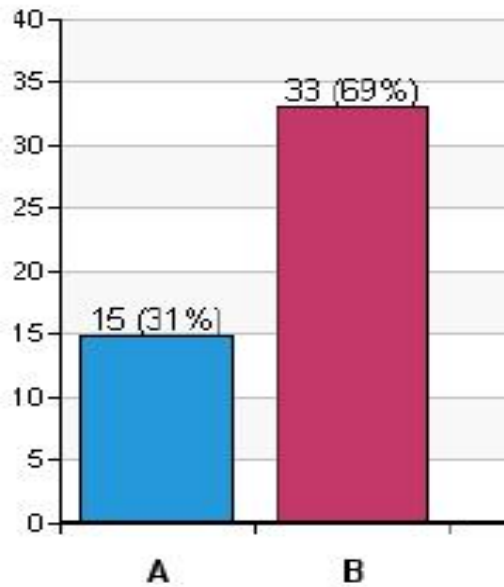
# Beckon

A = yes    B = no



## Footbridge and Beckon similarity

A = Perform act    B = Do not perform act

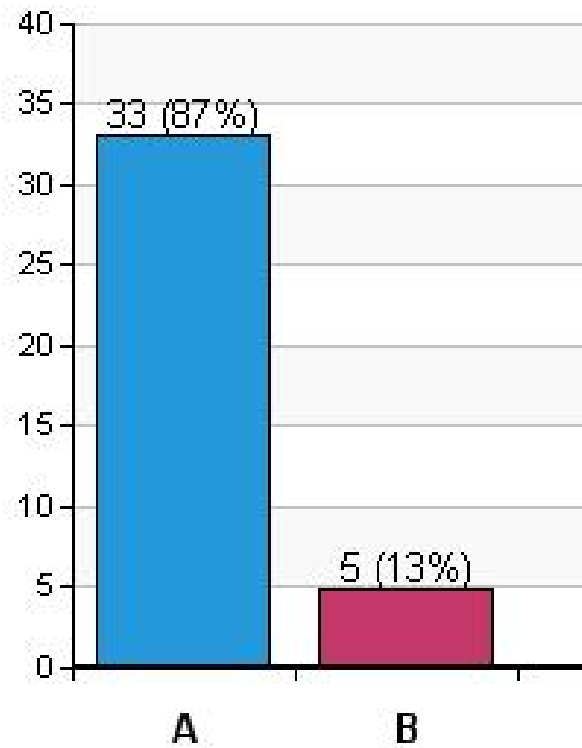


# Wave



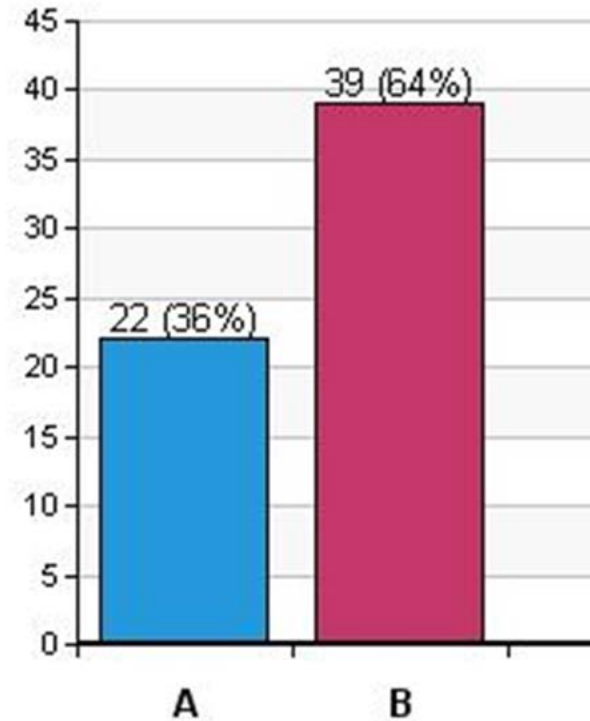
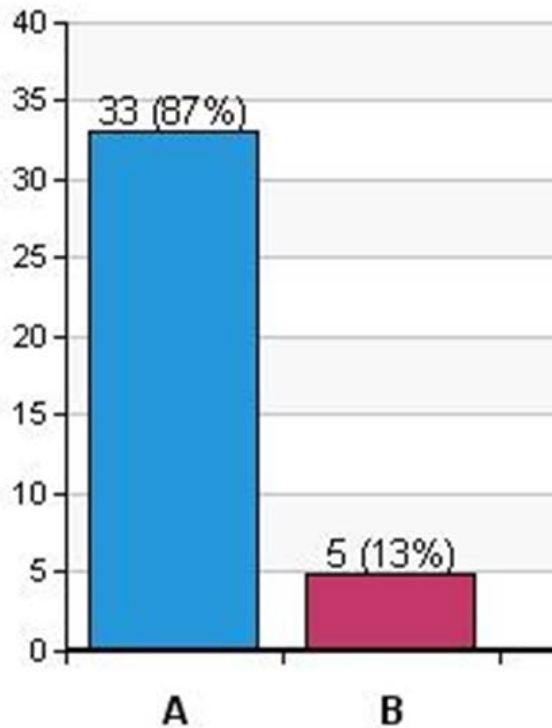
# Wave

A = Yes B = No



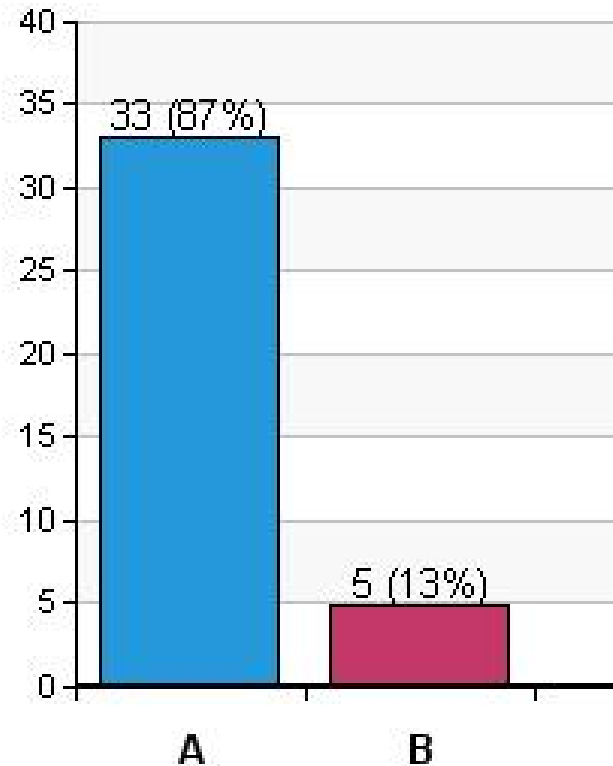
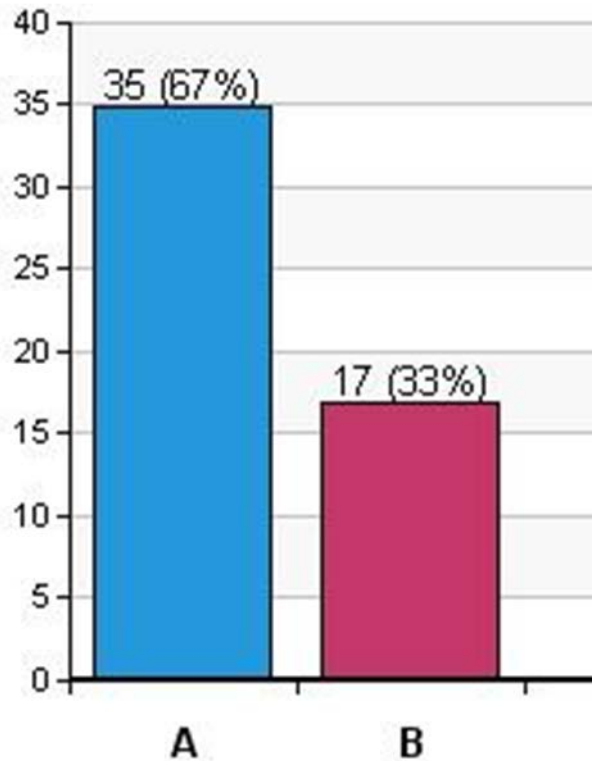
## Wave vs. Beckon “asymmetry”

A = Perform act    B = Do not perform act



## Switch and Wave similarity

A = Perform act    B = Do not perform act



# **Can an approach to intuitive moral judgment based upon prospective modeling ...**

- ... afford an explanation that unifies these seemingly diverse verdicts, which, taken together, do not fit either traditional deontological or utilitarian theories, and do not fit Green's "dual-process" theory.
  - And might this explanation tie the pattern in these cases to factors that are genuinely morally relevant?

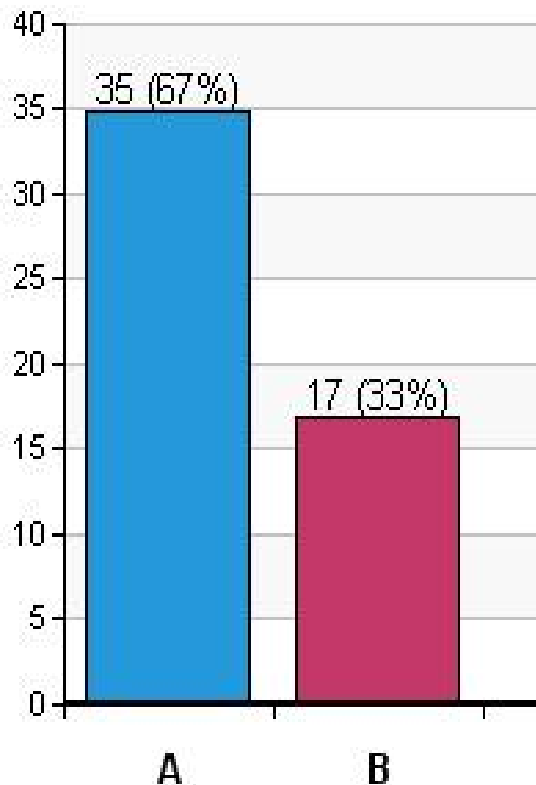


**(5) Is the *agent* being modeled?**

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

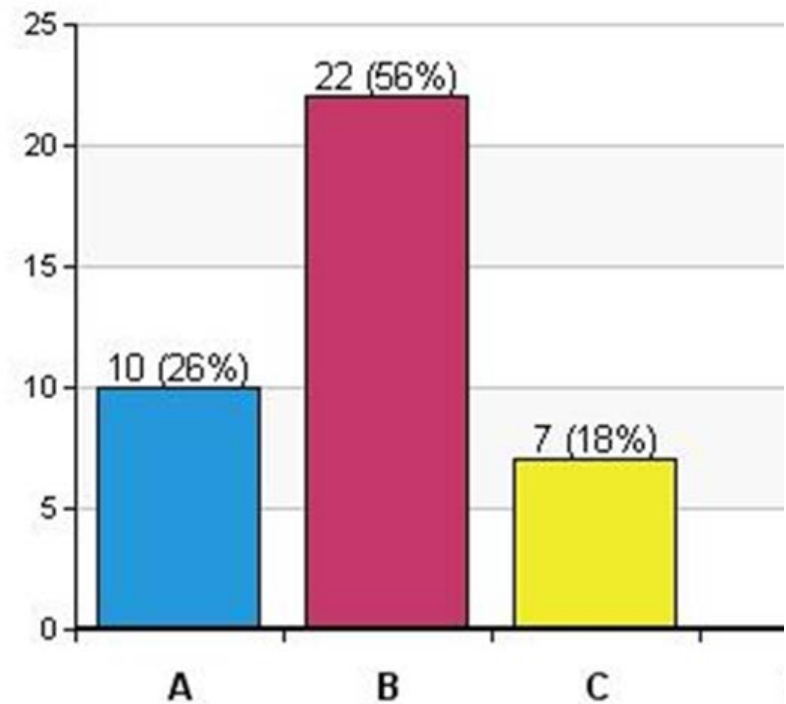
## Switch

A = pull B = do not pull



## Switch aftermath

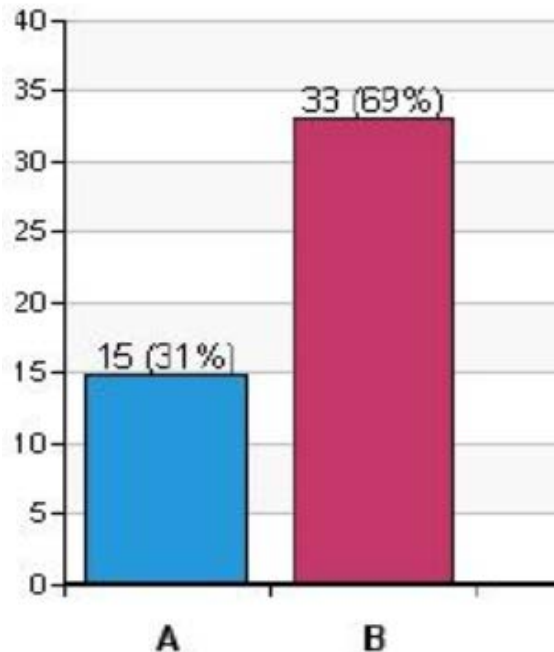
More, same, less trusting



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

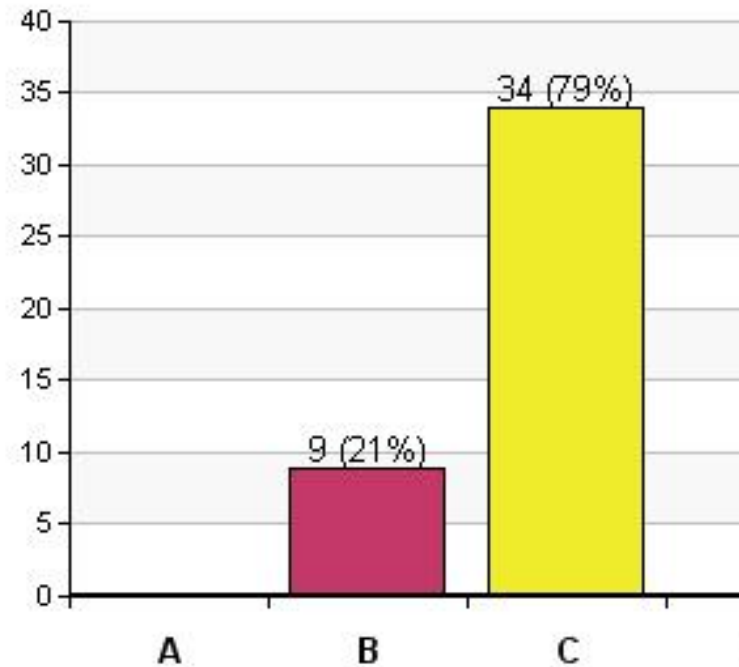
## Footbridge

A = push B = do not push



## Footbridge aftermath

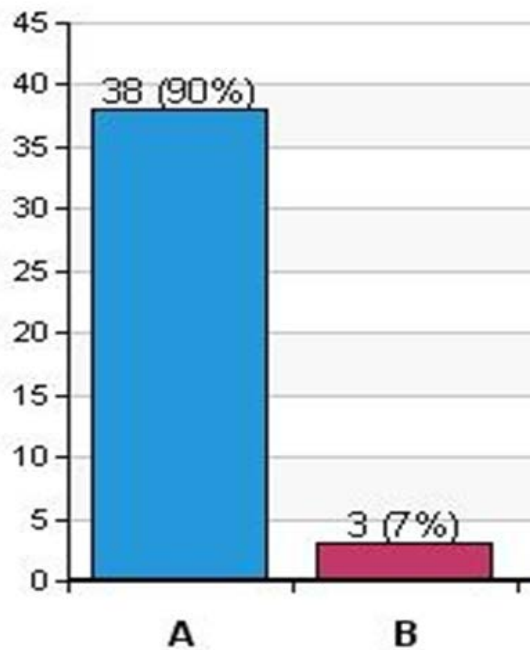
More, same, less trusting



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

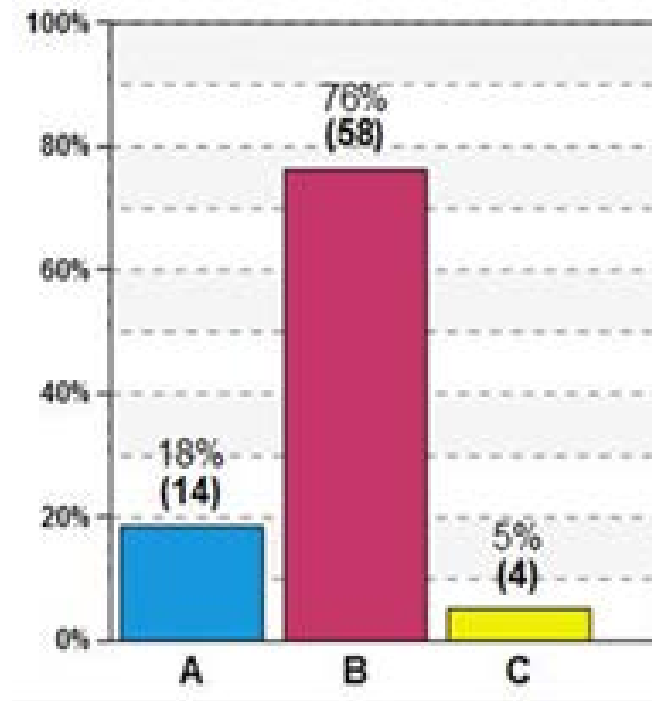
## Loop

A = pull switch B = do not pull



## Loop aftermath

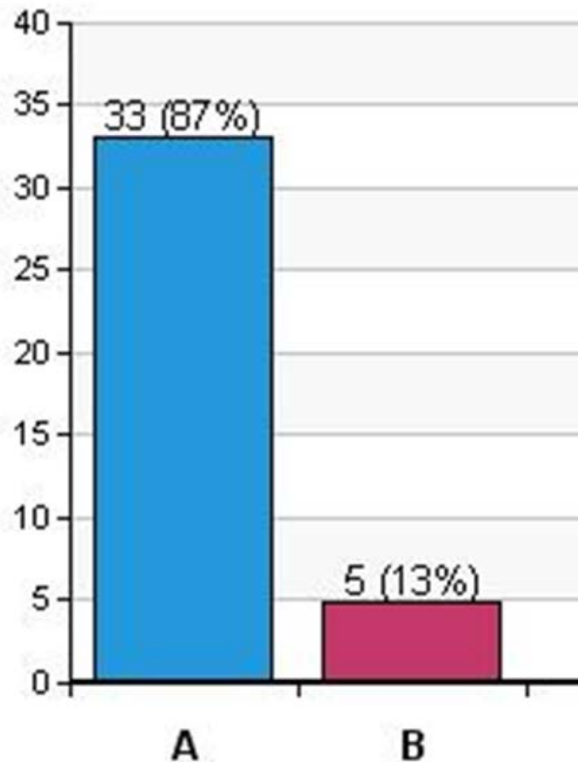
More, same, less trusting



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

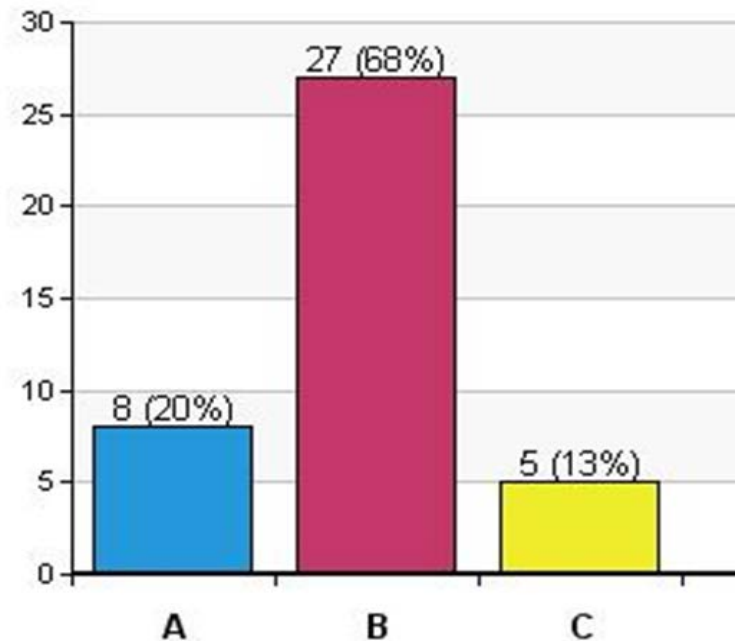
## Wave

A = wave B = do not wave



## Wave aftermath

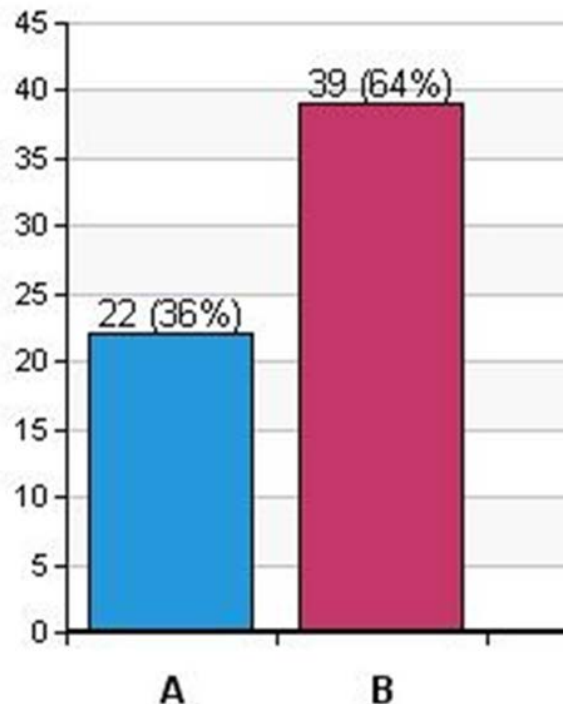
More, same, less trusting



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

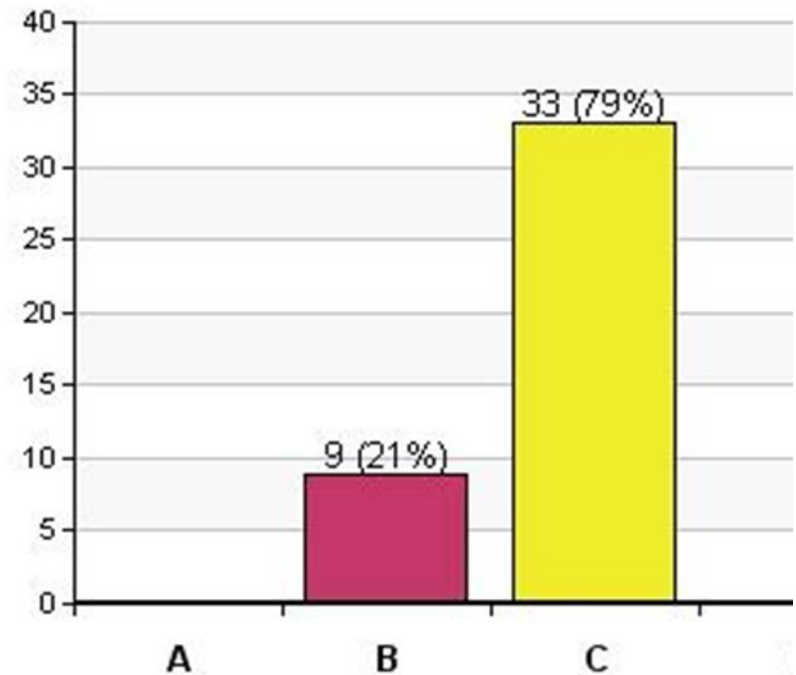
## Beckon

A = beckon B = do not beckon



## Beckon aftermath

More, same, less trusting

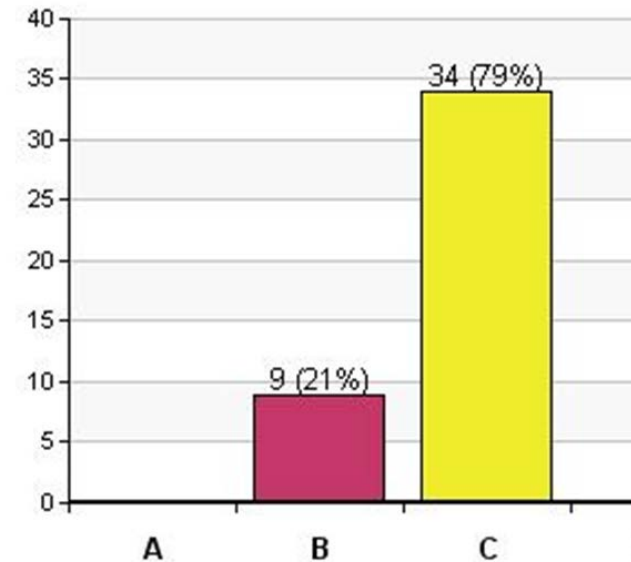
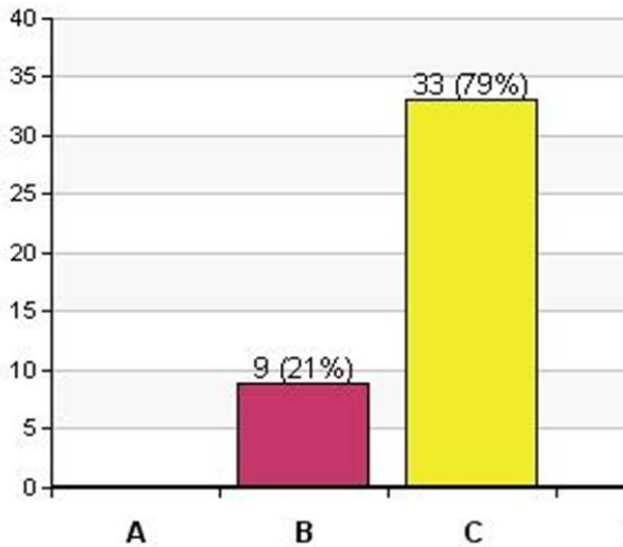
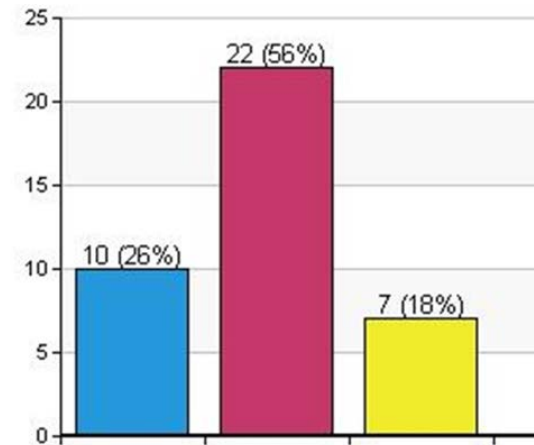
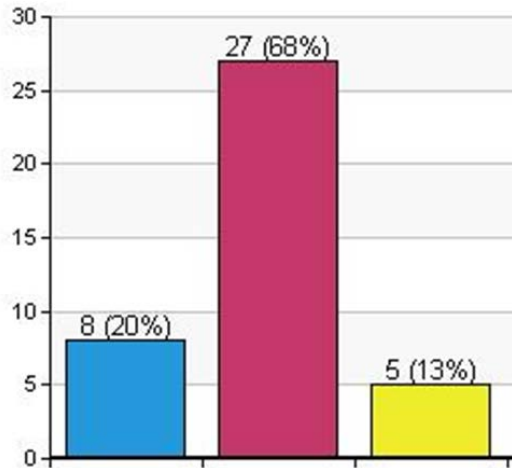


# Trust

## Wave and Switch

### Beckon and Footbridge

More, same, less trusting



# 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.



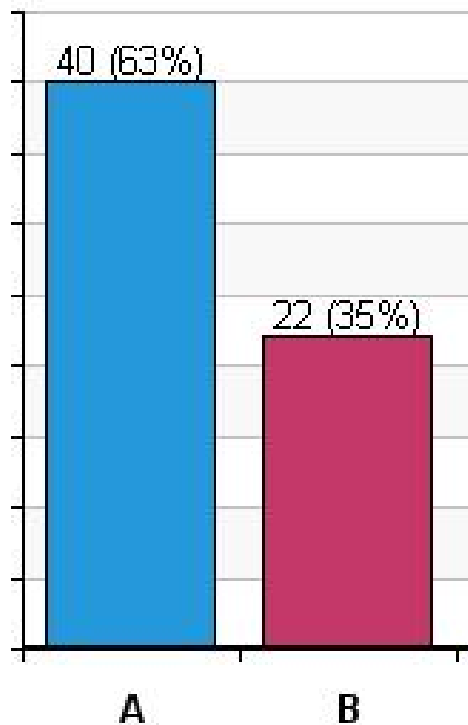
# Models of the agent mediate moral intuitions

- 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.

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

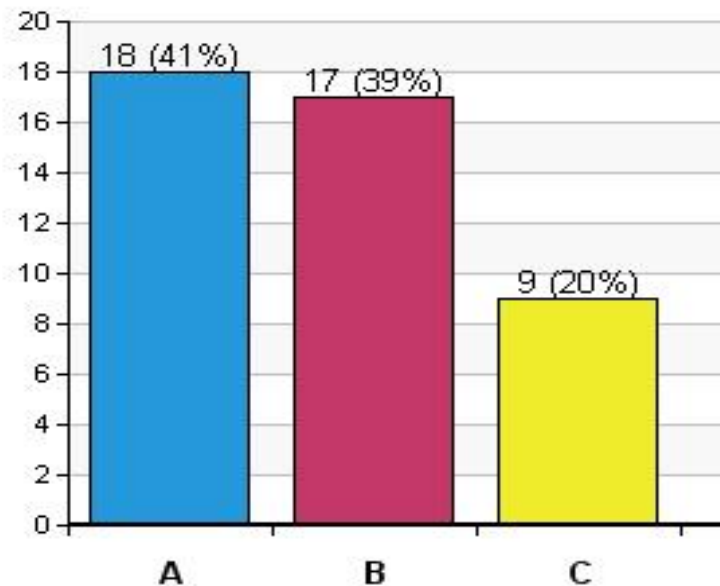
## Bus

A = push B = do not push



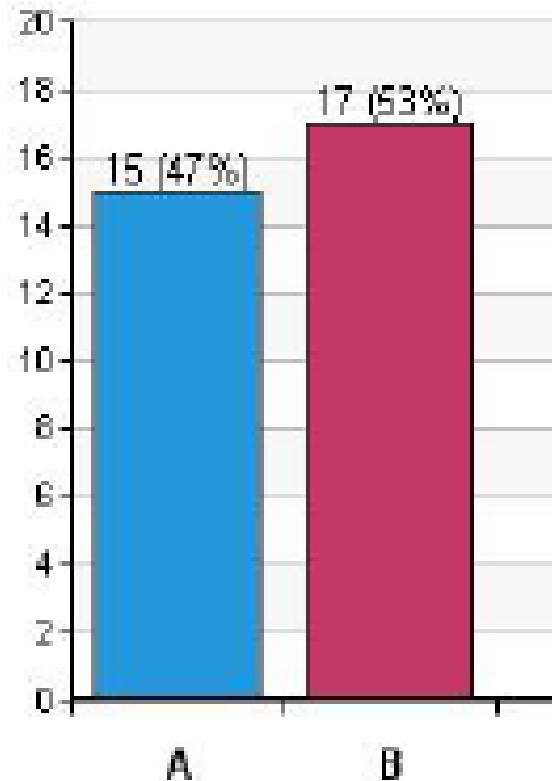
## Bus Aftermath

more, same, less trusting



**“Automatic aversion”? If you alone are leaving the bus, would you say you should throw *yourself* on top of the bomber?**

A = yes    B = no



**Are my students mentally simulating the situation and its possible outcomes?**

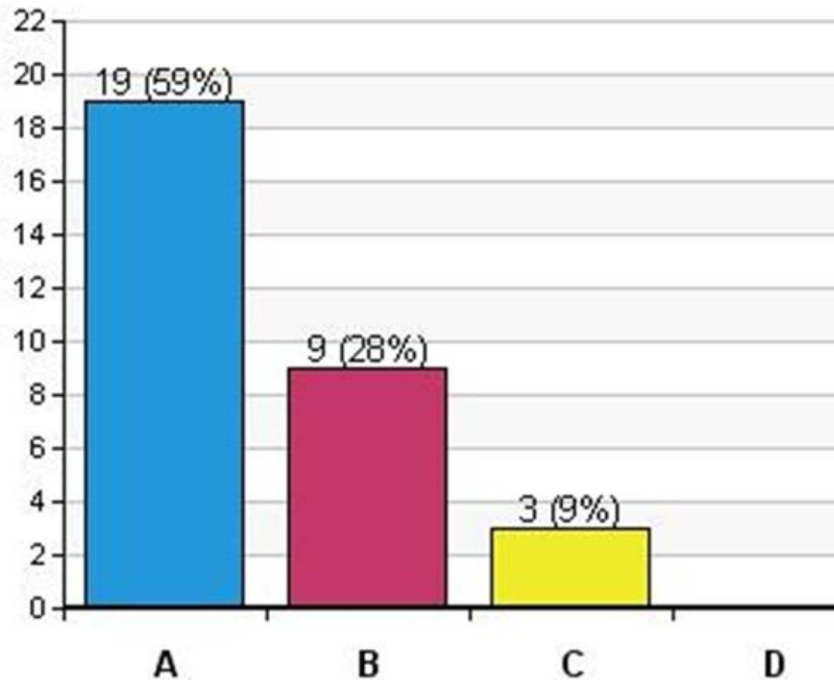
# Visualization and moral assessment

- Amit and Greene (2012) found that selectively interfering with visualization *increased* “cost-benefit” (e.g., “pushing”) responses in Footbridge-like dilemmas.
- Students’ self-reported “imaginative proximity” predicts their pattern of verdicts.

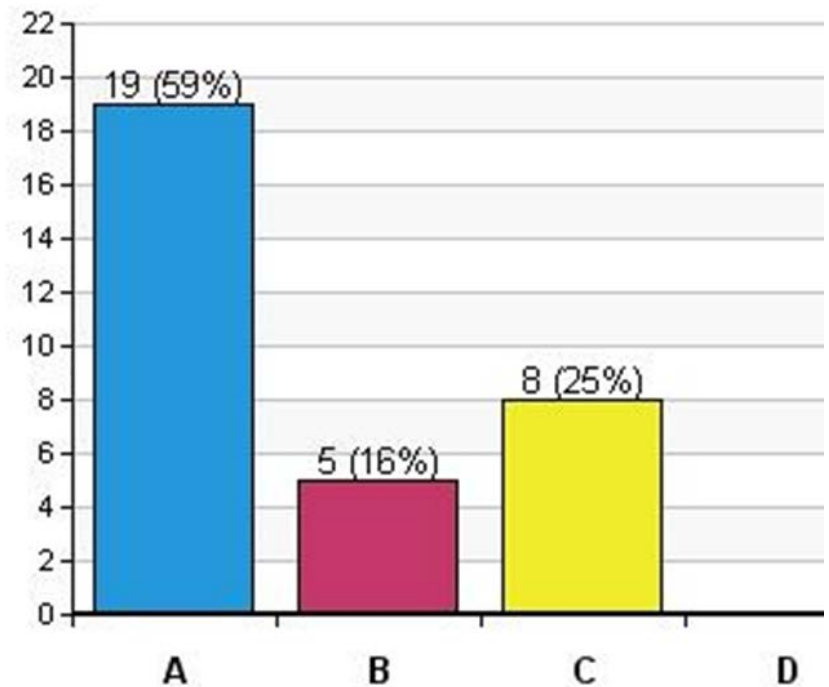
# Imaginative “proximity” of potential victims

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

- **Switch**



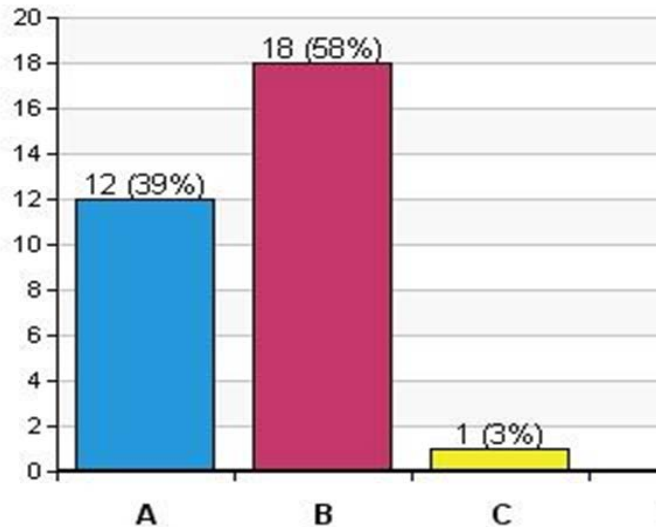
- **Wave**



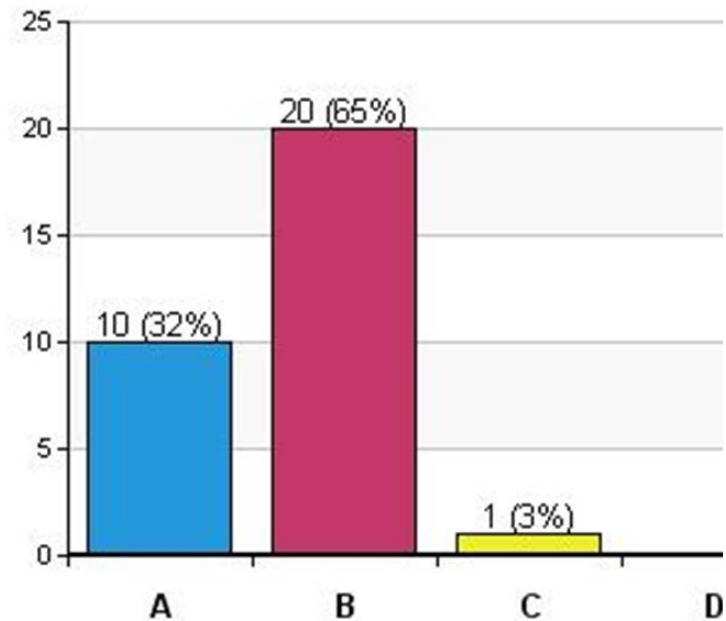
# Imaginative “proximity” of potential victims

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

- **Footbridge**

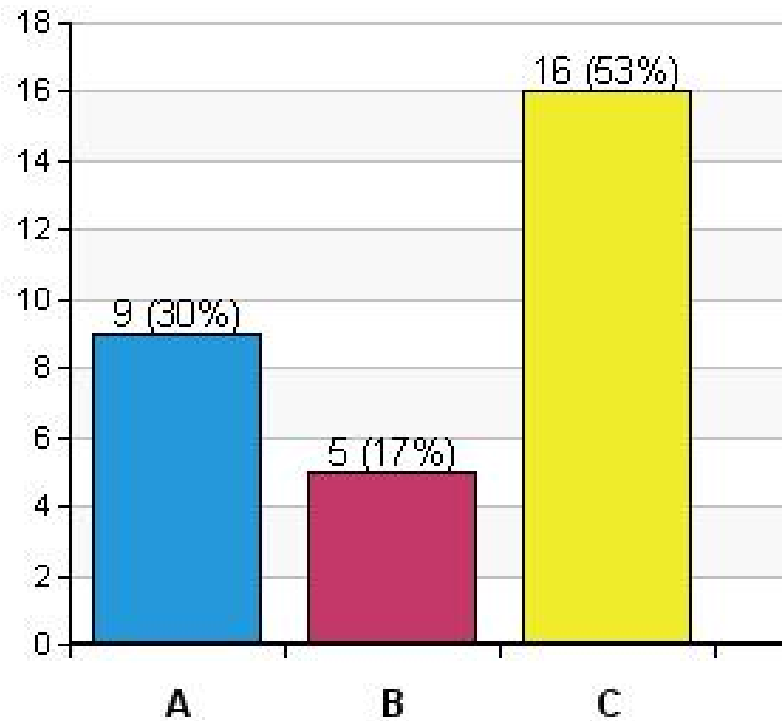


- **Beckon**



# 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





# Reactive attitudes – Switch




- Suppose that you had been in a Switch situation, and had pulled the switch, killing the worker on the sidetrack but saving the other five workers.
- Suppose you feel that you must now meet with the family of this man. Which comes closest to describing how you believe you would likely feel?
  - (A) Deeply regretful, sympathetic for their loss, and with a *reasonable hope* they might understand.
  - (B) Deeply regretful, guilty, and sympathetic for their loss, with *some hope* they might understand.
  - (C) Deeply regretful, ashamed, and sympathetic for their loss, with *little hope* they might understand.

## Switch aftermath:

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

# Reactive attitudes – Footbridge




- Suppose that you had been in a Footbridge situation, and had pushed the large man onto the tracks, killing him and saving the five workers.
- Suppose you feel that you must now meet with the family of this man. Which comes closest to describing how you believe you would likely feel?
  - (A) Deeply regretful, sympathetic for their loss, and with a *reasonable hope* they might understand.
  - (B) Deeply regretful, guilty, and sympathetic for their loss, with *some hope* they might understand.
  - (C) Deeply regretful, ashamed, and sympathetic for their loss, with *little hope* they might understand.

## Footbridge aftermath:

A = regretful and sympathetic, reasonable hope

B = regretful, guilty, and sympathetic, some hope

C = regretful, ashamed, and sympathetic, little hope

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




## Comparison:

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

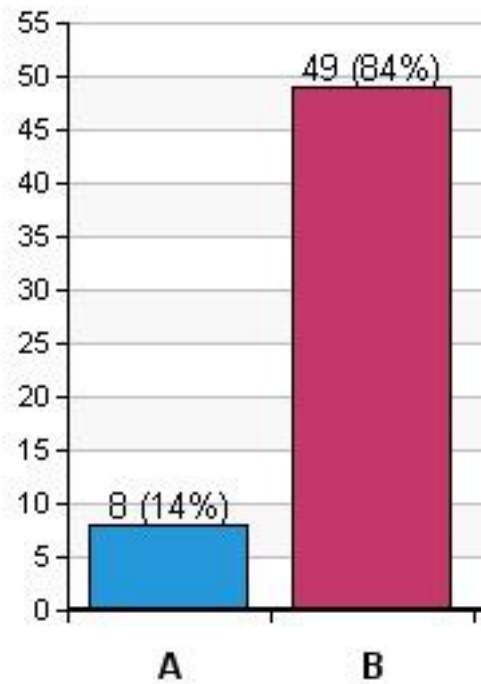
# The Knobe effect

# In the Boardroom, I

- *The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.' The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was helped. [Knobe, 2006]*
- Did the chairman intentionally help the environment?
- Yes?
- No?

# In the Boardroom, I

A = yes    B = no



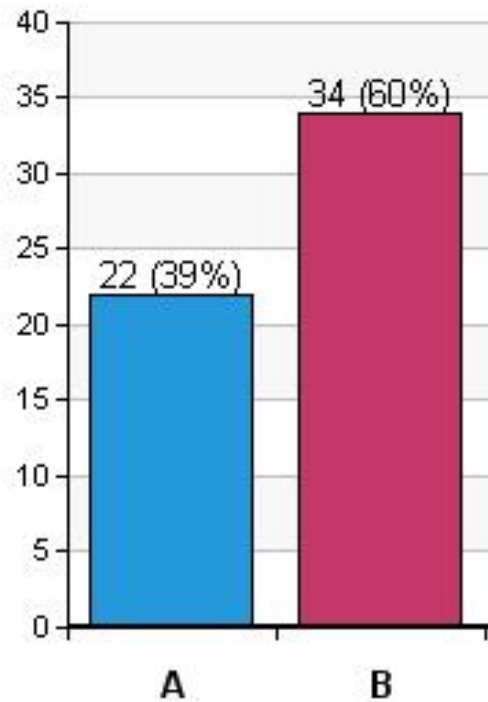


## In the Boardroom, II

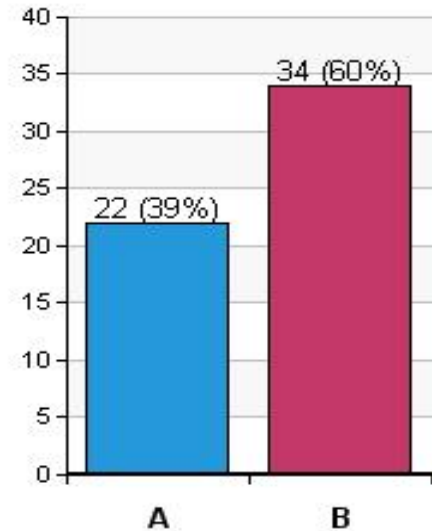
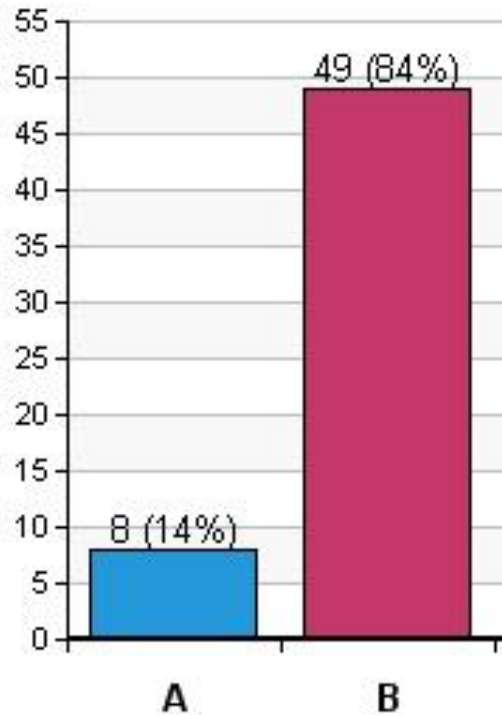
- *The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed. [cf. Knobe, 2006]*
- Did the chairman intentionally harm the environment?
- Yes?
- No?

# In the Boardroom, II

A = yes    B = no



# Boardroom I vs. Boardroom II asymmetry



# Goat Herder, I

- You are a herder living on a hillside. Above you is a neighbor with an orchard of olive trees. One day in town, you hear your neighbor being told, “Yes, you could use this spray on your trees, and it would kill all the bugs. And when the rain comes it will wash down onto your neighbor’s fields and kill the bugs that have been destroying her grass. It won’t hurt the goats, it will help them.”
- Your neighbor replies, not aware that you are present, “I don’t care at all about helping her goats, I just want to kill the bugs on my olive trees. Give me the spray.”
- Would you describe your neighbors attitude toward you as:
- ill will neutral good will

# Goat Herder, II

- You are a goat herder living in a remote village on a hillside. Above you is a neighbor with an orchard of olive trees. One day in town, you hear your neighbor being told, “Yes, you could use this spray on your trees, and it would kill all the bugs, but when the rain comes it will wash down onto your neighbor’s fields and poison her goats when they eat the grass.”
- Your neighbor replies, not aware that you are present, “I don’t care at all about poisoning her goats, I just want to kill those bugs. Give me the spray.”
- Would you describe your neighbors attitude toward you as:
- ill will neutral good will

# Modeling intentionality, not *moralizing* intentionality

- Sripada (2012) used structural equations analysis to find that an inferred evaluation of the motivational and evaluative attitudes of the CEO agent (a “deep self” model) mediates judgments in “Knobe effect” scenarios, leaving no residual effect of positive or negative moral assessment of the action itself.
  - This is a “lay scientist” effect, contrary to Knobe’s “lay moralist” diagnosis (2010)

## Still dumbfounded?

- ***Janet and Matt** are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried playing Russian Roulette. At very least it would be a new experience for each of them. Fortunately, when the spin the revolver's chambers, neither of them lands on the bullet. They both enjoy playing Russian Roulette, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that, was it OK for them to play Russian Roulette? [cf. Haidt, 2001]*

## **(6) What if we remove the human agent?**



# Models and framing

- Dual-process theorists emphasize that the *framing* of scenarios can have a large effect on intuitive judgments.
- If domain-general general modeling abilities, rather than dual-processing, lies underneath intuitive judgment, it should be possible, by exploring hypothetical situations in a number of ways, to overcome framing effects and reach more stable verdicts, as Hume claimed:
  - “In order, therefore, to prevent those continual *contradictions*, and arrive at a more *stable* judgment of things, we fix on some *steady* and *general* points of view; and always [well ... ] in our thoughts, place ourselves in them, whatever may be our present situation.” (SBN 581-82)

## Let us consider a case ...

- ... that is simpler than the trolley problems, because it removes *agency* from the actual hypothetical scenario.
- The model-based approach would suggest that we would *not* see the same kinds of asymmetries in these cases, since no assessment of the motivational character of the agent is involved.

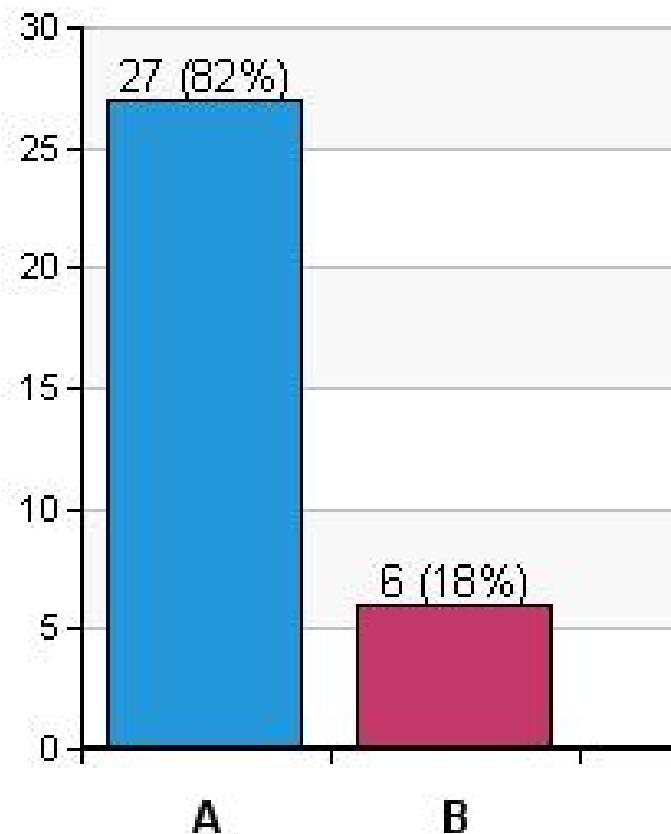
# A realistic trolley problem?

(Bonnefon et al., 2015)



# Should self-driving car swerve to the side, killing one pedestrian but saving five?

A = swerve    B = do not swerve



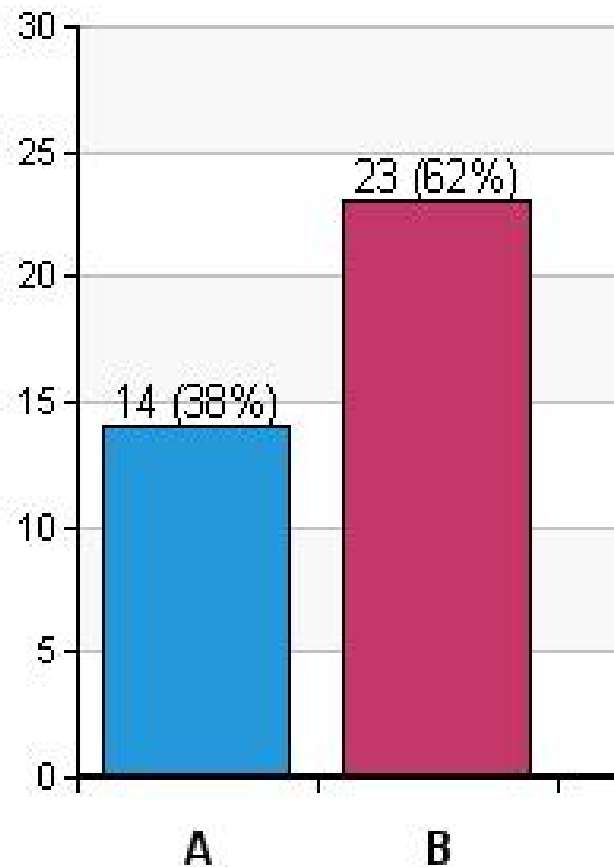
# A realistic trolley problem?

(Bonnefon *et al.*, 2015)



# Should self-driving car swerve into wall, killing car occupant but saving five pedestrians?

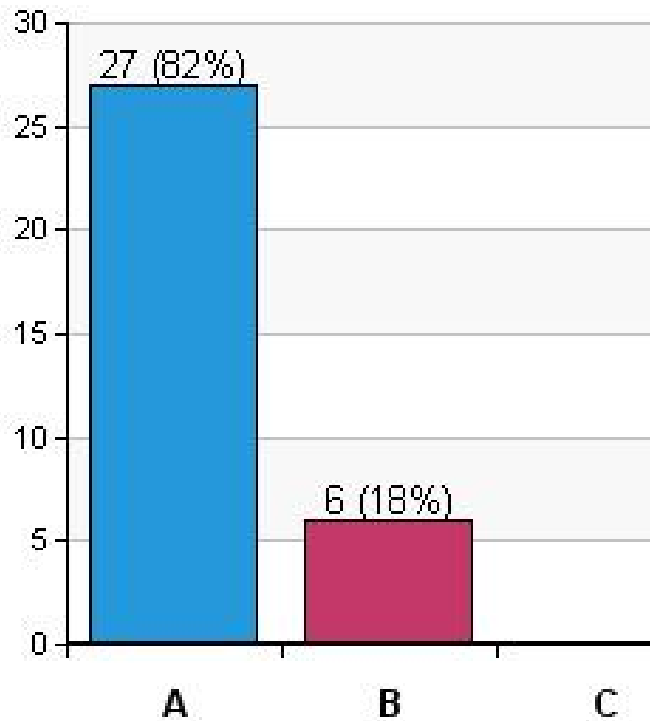
A = swerve    B = do not swerve



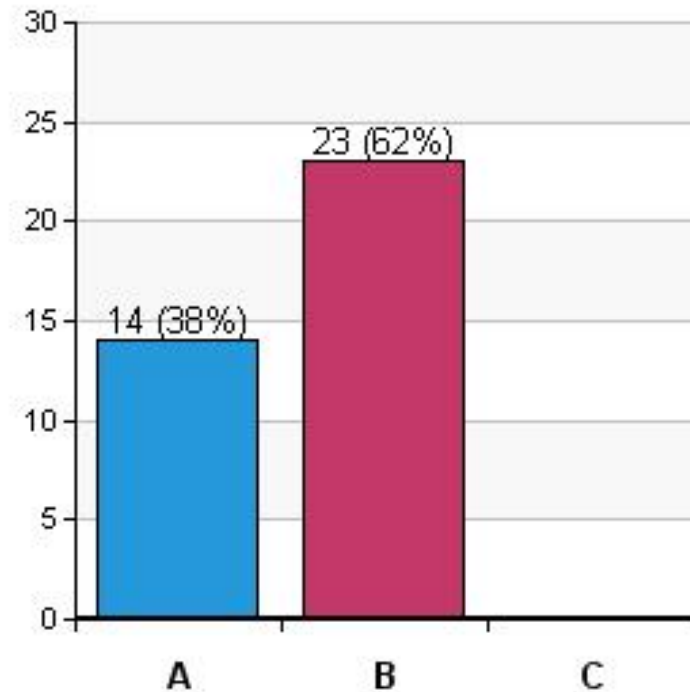
# Self-driving asymmetry?

A = veer B = don't veer

- **Pedestrian victim**



- **Rider victim**



## **We could stop there ...**

- ... and conclude that we still get asymmetries, even after removing the agent, owing to factors of the physical set-up that are unrelated to genuine moral considerations?
- Let's try the Humean experiment of altering imaginative perspective to see if the asymmetries survive.



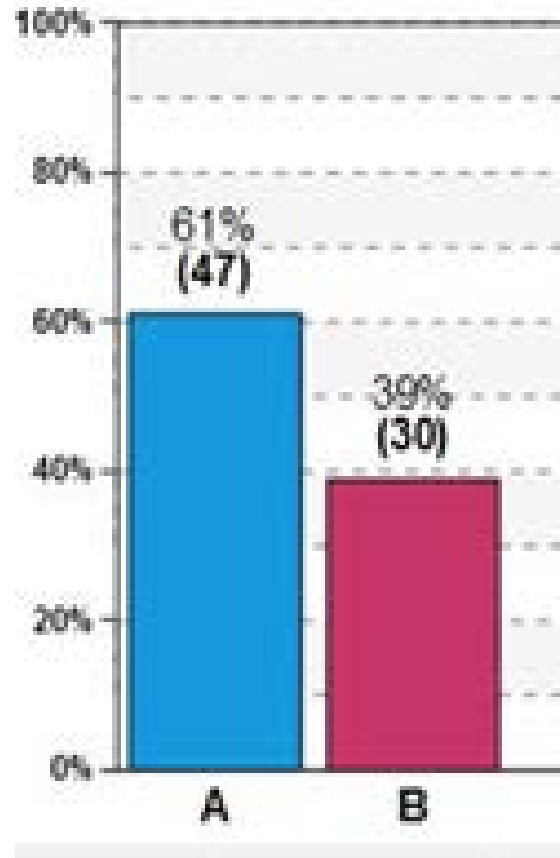
## Question 6

- You are vacationing in Sweden, where self-driving cars have become common on the roads. You don't have a car, and are traveling around the cities on foot. What do you think is best, should the self-driving cars in Sweden be programmed to swerve to avoid hitting five pedestrians in a crosswalk, even if this means hitting one pedestrian on the sidewalk (who is not now at risk)?
- (A) Yes
- (B) No



**In Sweden, should self-driving car veer into wall, killing a pedestrian  
not now at risk but saving five other pedestrians?**

A = yes B = no



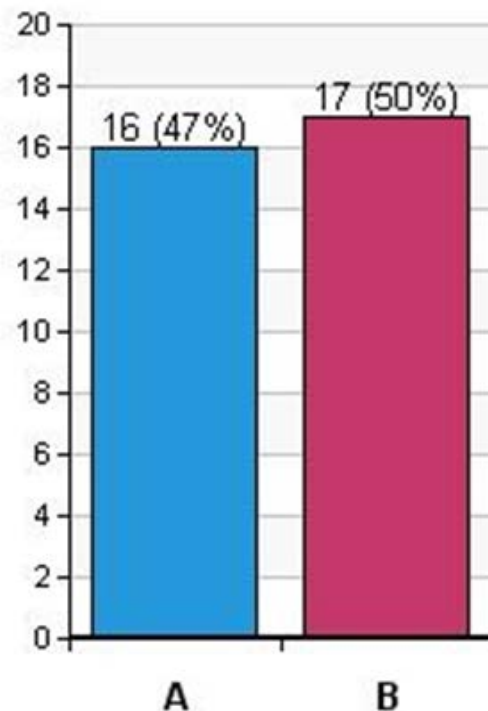
## Question 7

- You are vacationing in Sweden, where self-driving cars have become common on the roads. You don't have a car, and are traveling around the cities on foot. What do you think is best, should the self-driving cars in Sweden be programmed to swerve into a wall to avoid hitting five or more pedestrians in a crosswalk, even if this means hitting the wall with sufficient force to kill the passenger in the car (who is not now at risk)?
- (A) Yes
- (B) No



**In Sweden, should self-driving car veer into wall, killing the car occupant but saving five pedestrians?**

A = yes B = no



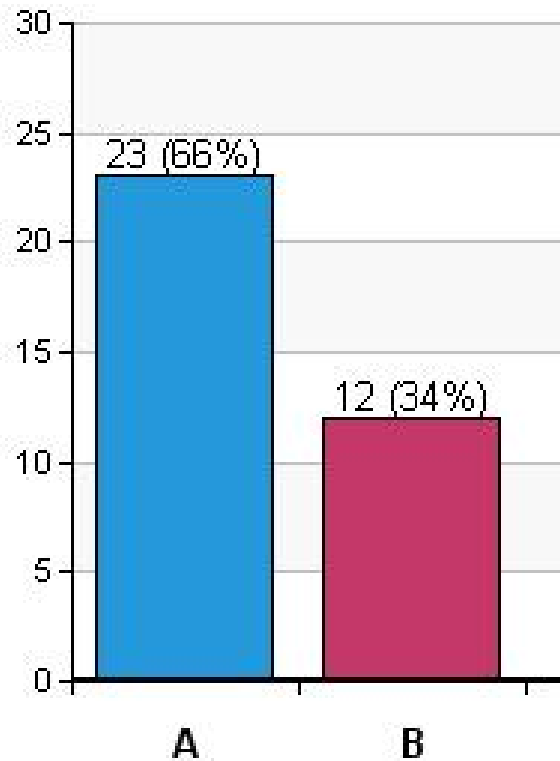
A small toolbar with icons for back, forward, search, and other presentation controls, with a dropdown menu showing 'A-E'.

### Question 3: Point of view

- When I first answered the question about the rule a self-driving car should follow, I implicitly thought of myself as *in* the car, rather than in the group of five people on the road ahead.
- (A) Yes
- (B) No

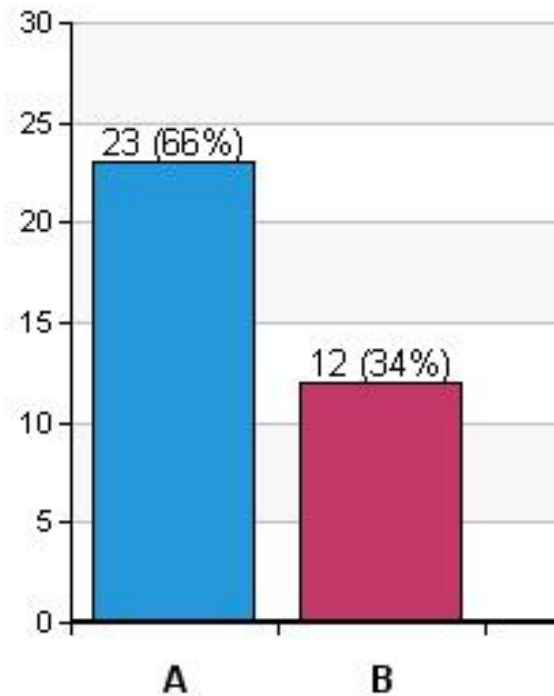
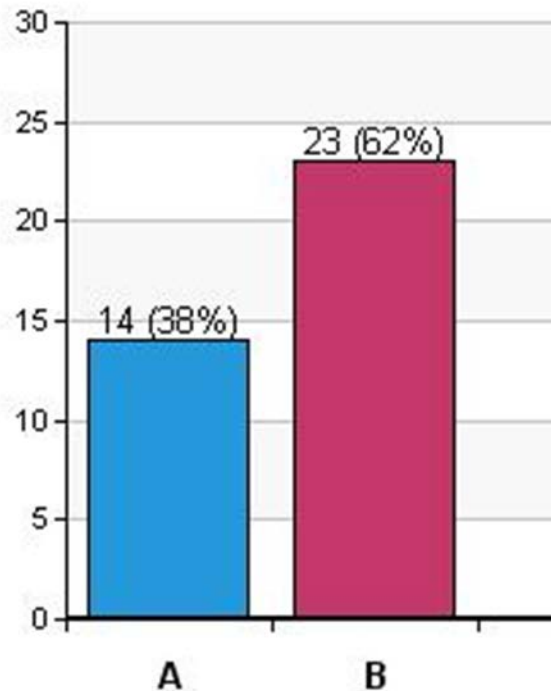
**At first took the point of view of the rider in the self-driving car**

A = yes B = no



# Self-driving verdict and simulated point of view in case of swerving into wall

A = swerve B = don't swerve    A = rider B = pedestrian viewpoint





## Question 13 Redux: Self-driving vehicle, II

- Should self-driving vehicles be programmed such that, if they approach a collection of five or more pedestrians on the street ahead, they will swerve to the side even when this causes the car to hit a wall and kill the person in the car?
- (A) Yes
- (B) No

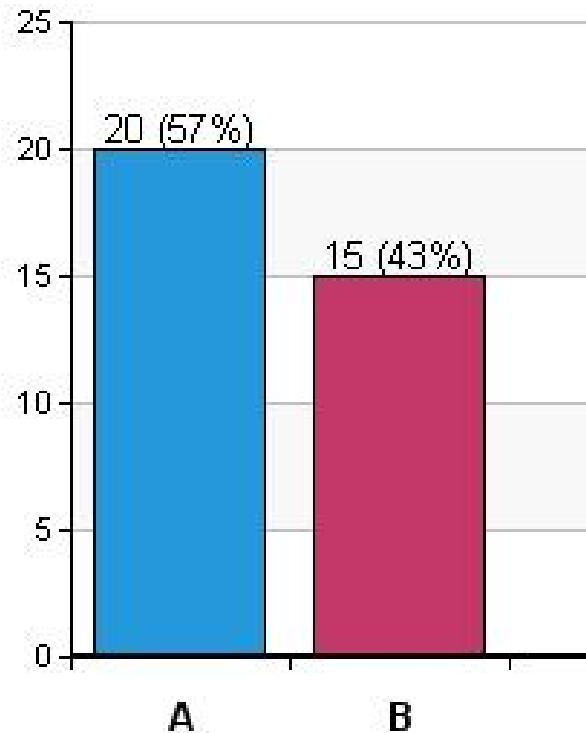


## **A week later ...**

- ... in the absence of further discussion of the self-driving car problem in lecture ...

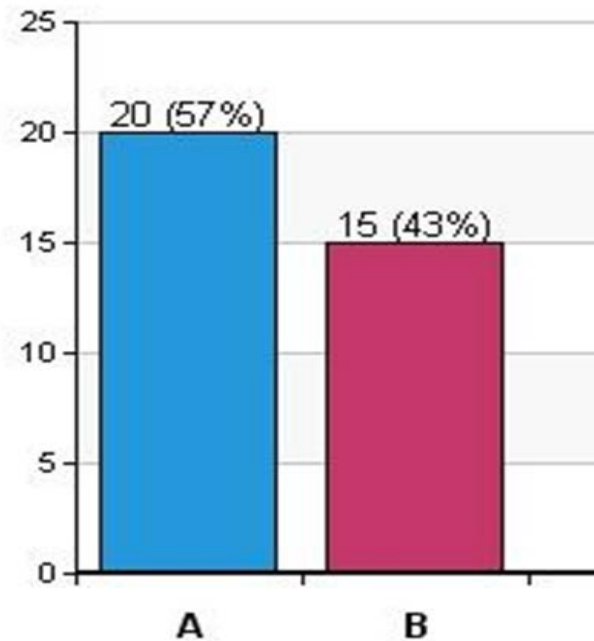
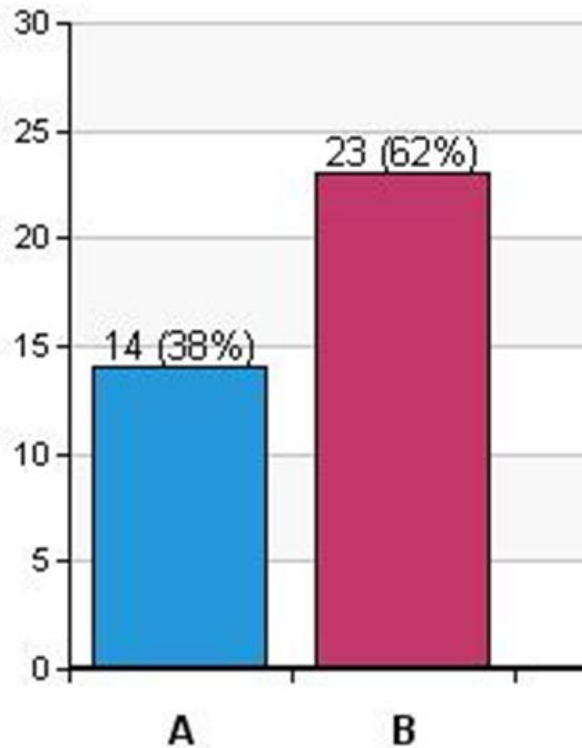
# Self-driving redux: Should self-driving cars in the US be programmed to veer into a wall, killing car occupant but saving five pedestrians?

A = yes B = no

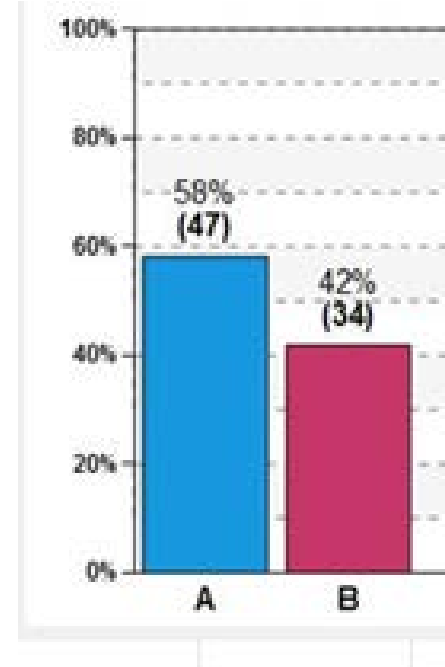
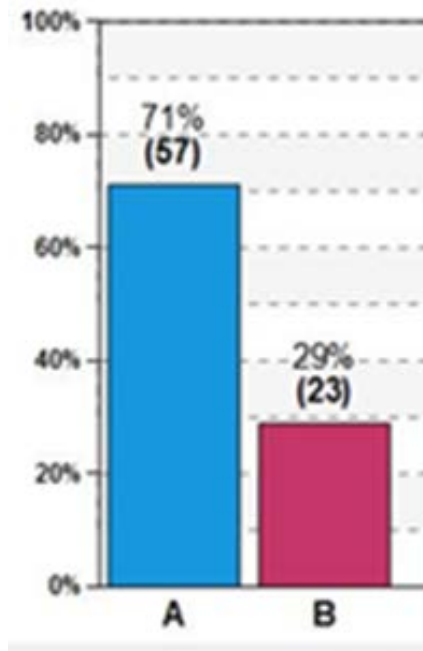


# Reversal of judgment on swerving the self-driving car into a wall, killing the occupant but saving five pedestrians

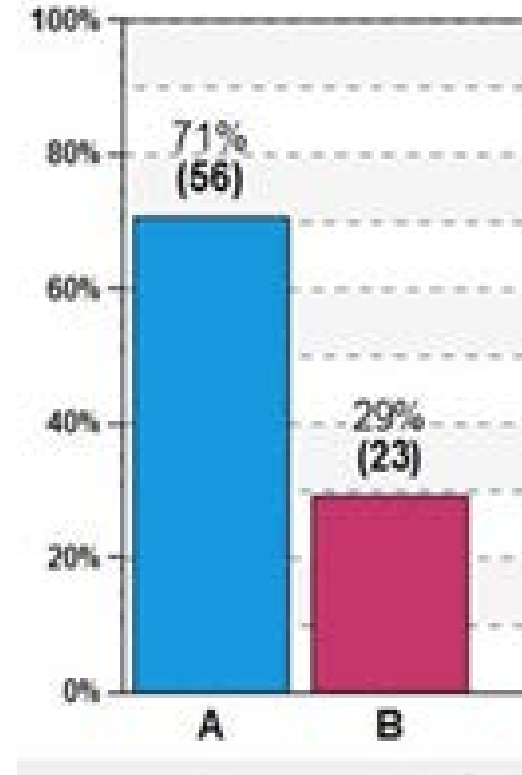
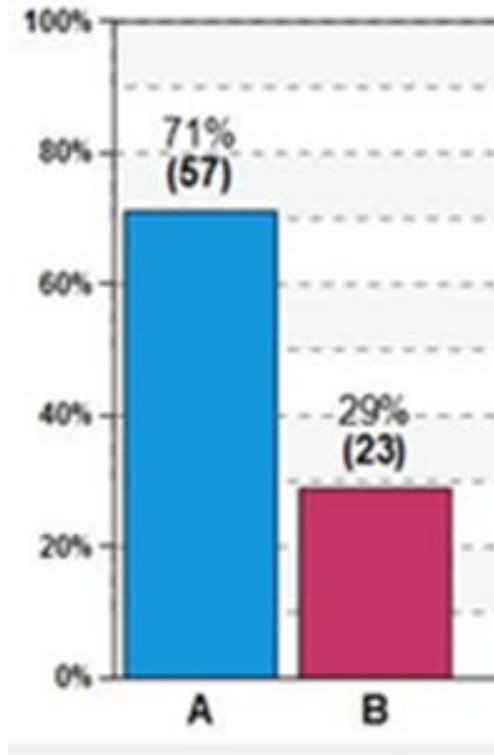
A = yes B = no



# Initial asymmetry, pedestrian vs. occupant

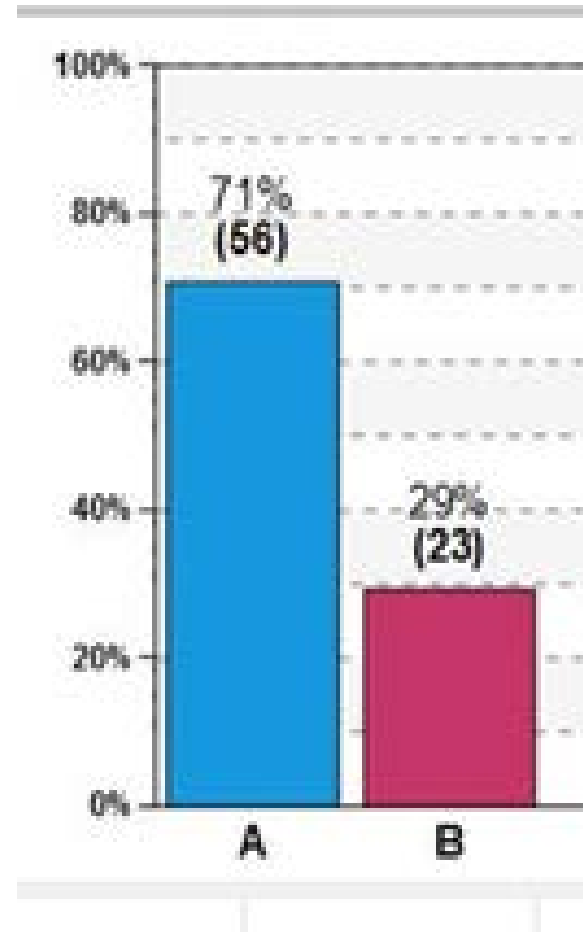
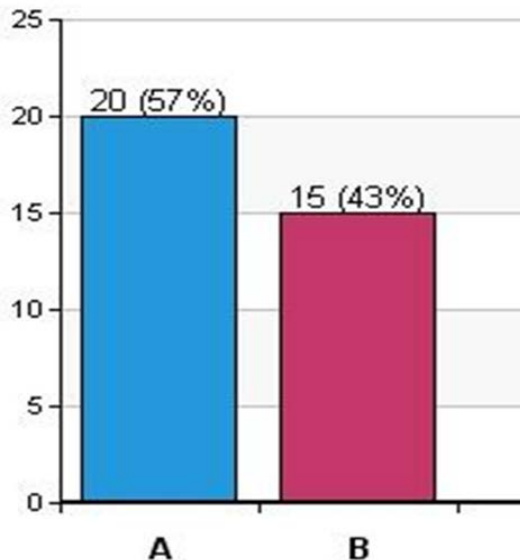


# Final symmetry: first week pedestrian victim vs. third week occupant victim



**Should self-driving cars in the US be programmed to veer into a wall, killing car occupant but saving five pedestrians?**

A = yes B = no



# Removing the human agent

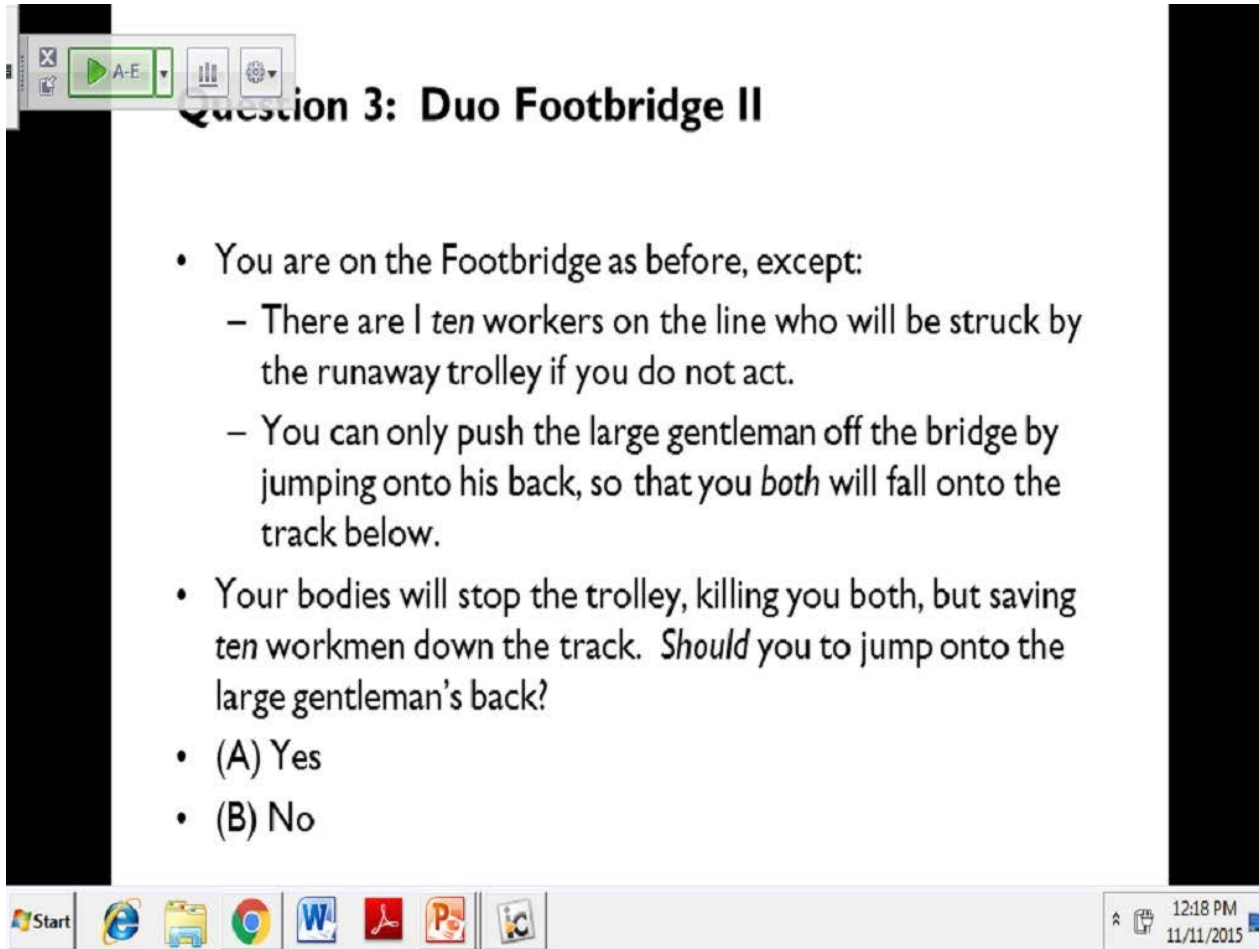
- ... and displacing the decision-making to the society as a whole, in anticipation of possible situations, may remove what appears to be a “trolley-like” asymmetry in judgment.

# Can we remove dubious character from Footbridge?

- One prediction: changing Footbridge-like scenarios in a way that would involve different cognitive and motivational attitudes, as in Bus, would produce a difference in intuitive evaluation of the appropriateness of the act.



# Duo Footbridge

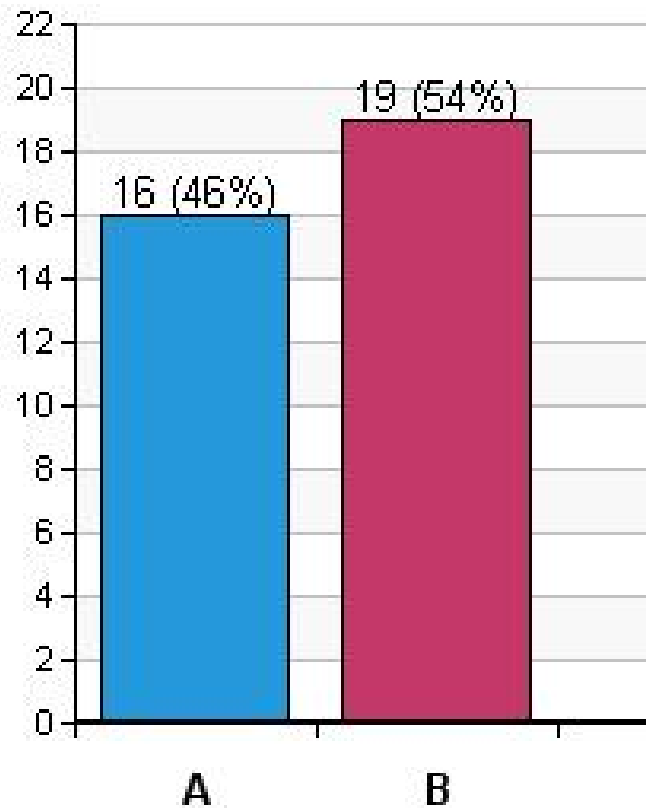


**Question 3: Duo Footbridge II**

- 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

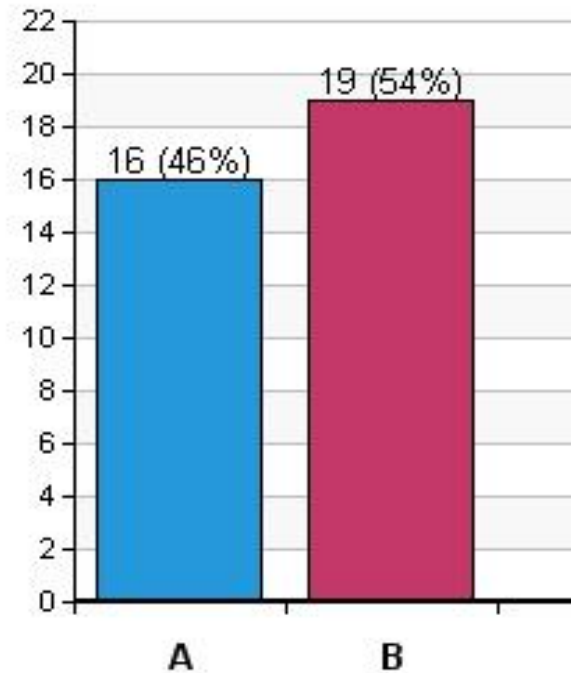
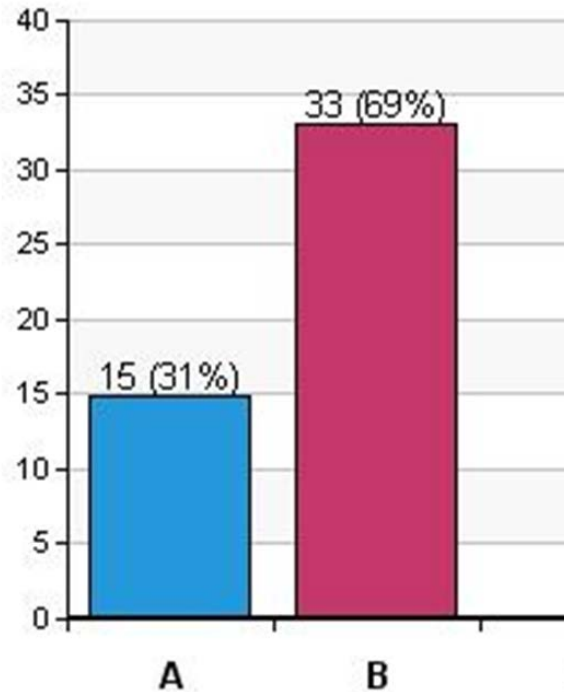
**Should you jump on the back of the large gentleman, so that you  
*both* block the trolley, saving ten lives?**

A = yes   B = no

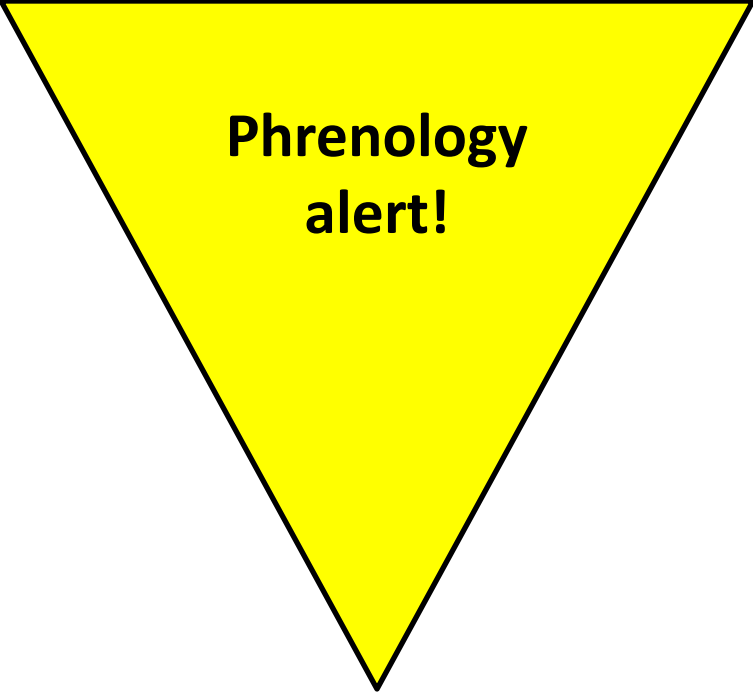


# Footbridge vs. Duo Footbridge

A = push/jump    B = don't push/don't jump



## **(7) But what about the neuroscience evidence?**

A large yellow equilateral triangle pointing downwards, centered on the page. It has a black outline.

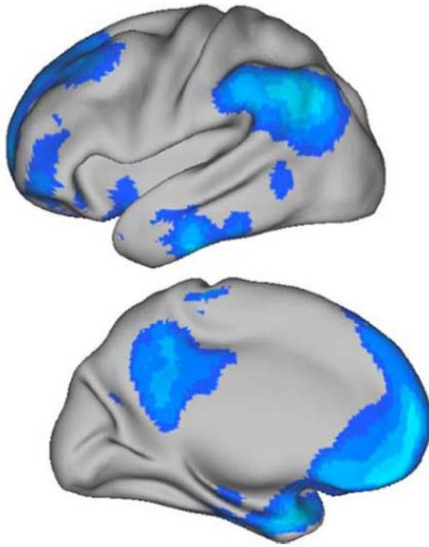
**Phrenology  
alert!**

(Hagman *et al.*, 2008)



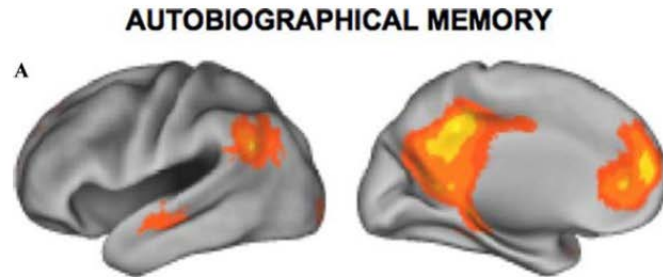
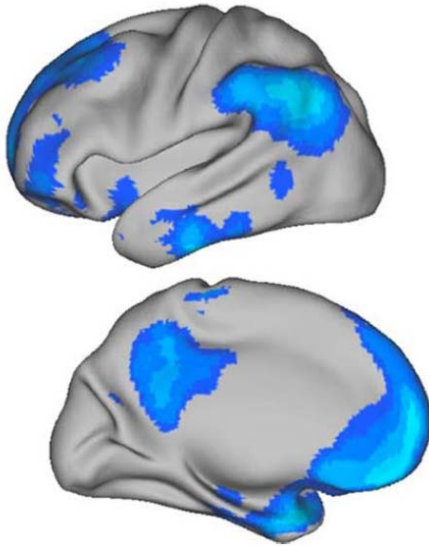
# Default network

(Buckner *et al.*, 2008)



# Default network

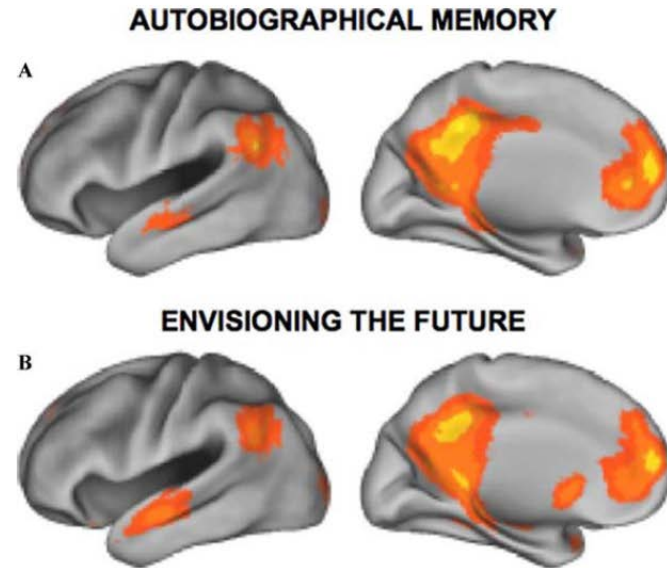
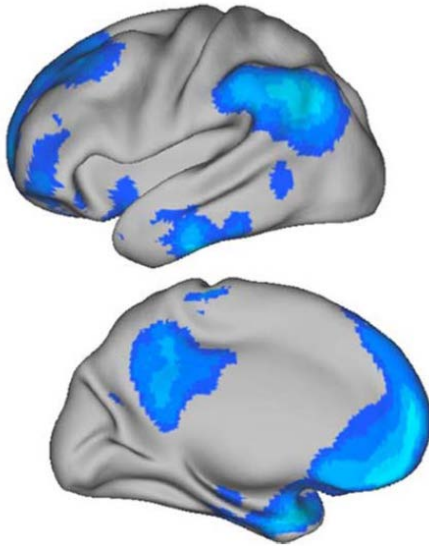
(Buckner *et al.*, 2008)





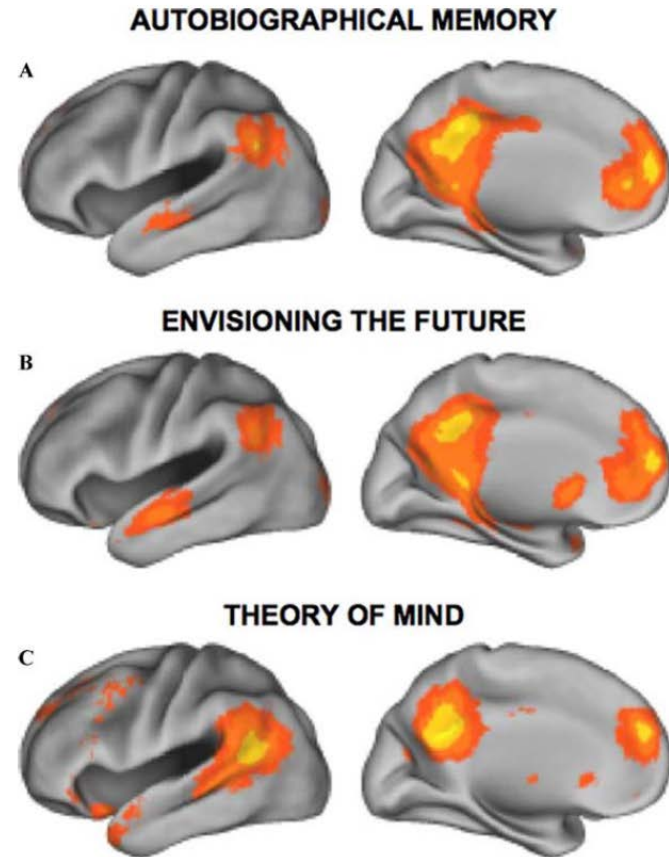
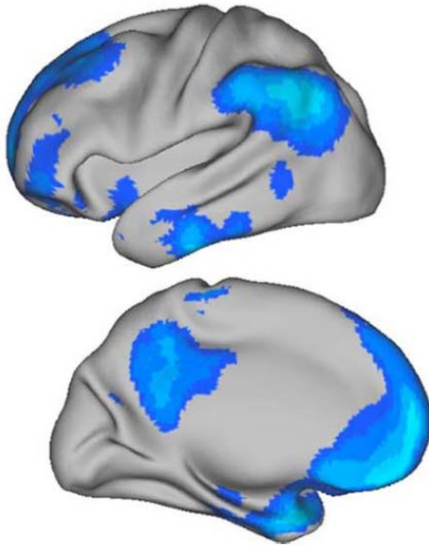
# Default network

(Buckner *et al.*, 2008)



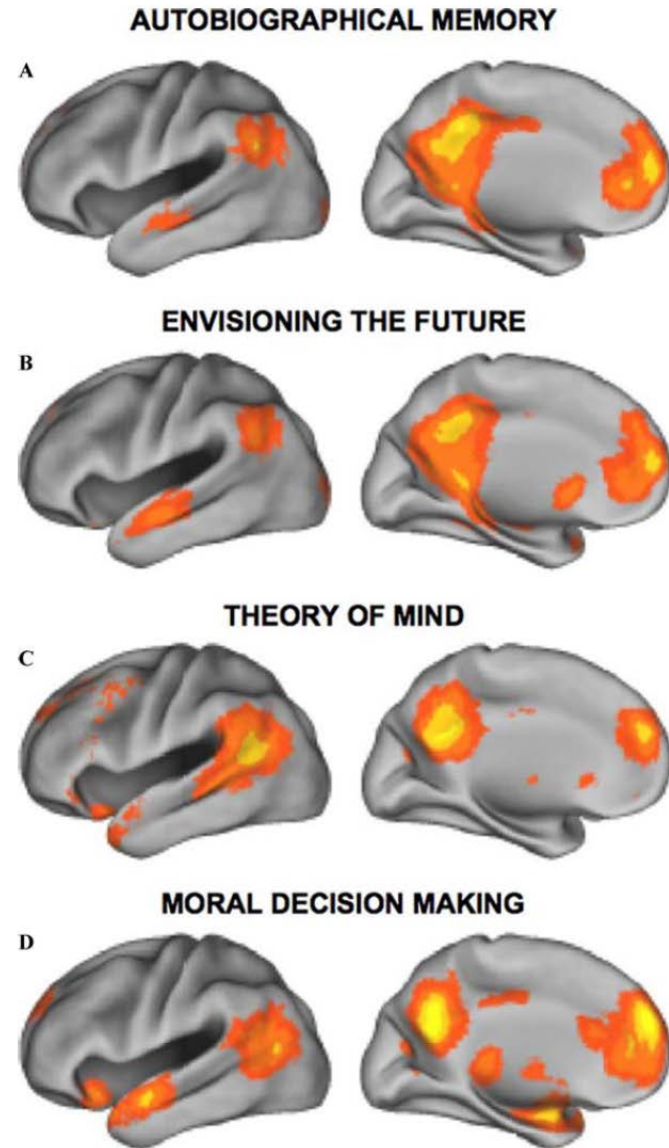
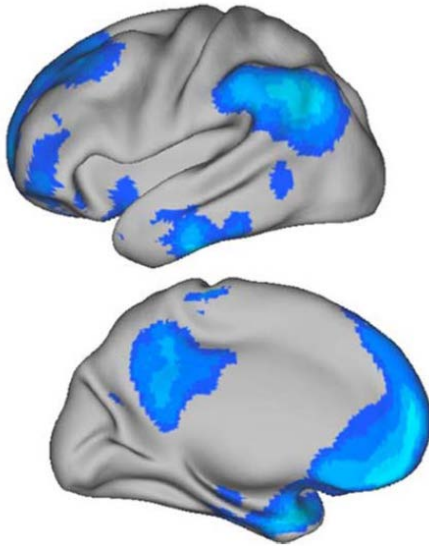
# Default network

(Buckner *et al.*, 2008)



# Default network

(Buckner *et al.*, 2008)



# A more unified picture of evaluation and action

- These large-scale, functionally-integrated, highly general brain networks that recruit information widely to construct models that permit prospective simulation of actions and outcomes recruit information widely (Buckner & Carroll, 2006; Hassabis & Maguire, 2009; Moll, *et al.*, 2005; Shenhav & Greene, 2010).
  - These models guide decision and action *generally*, integrating *evaluative* and *causal* information to yield expected values for actions and outcomes, and these expectations then can promote learning through discrepancy reduction (Buckner *et al.*, 2008; Daw *et al.*, 2016; Seligman *et al.*, 2016).

## **(7) Normative interest**

## (7) Normative interest

- Perhaps the debunking of moral intuition by “dual-process” theories is unsuccessful—the patterns of judgments we observe seem consistent with individuals making intuitive moral judgments on morally-relevant grounds.
- What might a *descriptively adequate* account of moral intuitions look like, in light of the evidence we’ve discussed?

## Suppose that one were seeking to develop ...

- Ordinary moral intuitions may embody a good deal of commonsense moral understanding.
  - Such an understanding involves a capacity to model, simulate, and situations, agents, and actions,
  - ... and to pose and answer evaluative questions about the kinds of character that might conduce to certain actions, and whether these underlying psychological characteristics would generally be morally good for people to have.
    - This would be a form of *indirect* assessment of individual *actions* in terms of more general models of *agents*.

## **This could point in the direction ...**

- ... of the empirical adequacy of a virtue theory, or of a characterological consequentialism.



# Beyond trolleyology

- We've seen that trolley problems can be diagnostic in understanding our implicit moral competency,
  - ... though we need to handle them with the kind of care any diagnostic method needs
- We can't simply "read off" moral conclusions. We should look at characteristics of given examples and explanations of responses:
  - What an example might or might not throw into salience or evoke
  - What limitations this might have in terms of responsiveness to morally relevant considerations