Moral Psychology and Ethical Theory: 50 Years of SPP

Society for Philosophy and Psychology
50th Anniversary Session
Purdue University

Peter Railton
June 2024

Late to the party …

Late to the party …

- Although many of the most important figures in ethics saw their philosophical work as intimately tied to the empirical study of mind and society—Aristotle, Hobbes, Rousseau, Mill, James, Dewey, even Kant—the 20th century saw an unprecedented effort by philosophers to put ethical theory on an independent, *a priori* foundation, grounded in conceptual analysis.
  - Moore and others denounced psychologizing in ethics as committing the "naturalistic fallacy" of thinking that an *ought* could be derived from an *is*, when conceptual analysis shows that they are categorically distinct.

Late to the party …

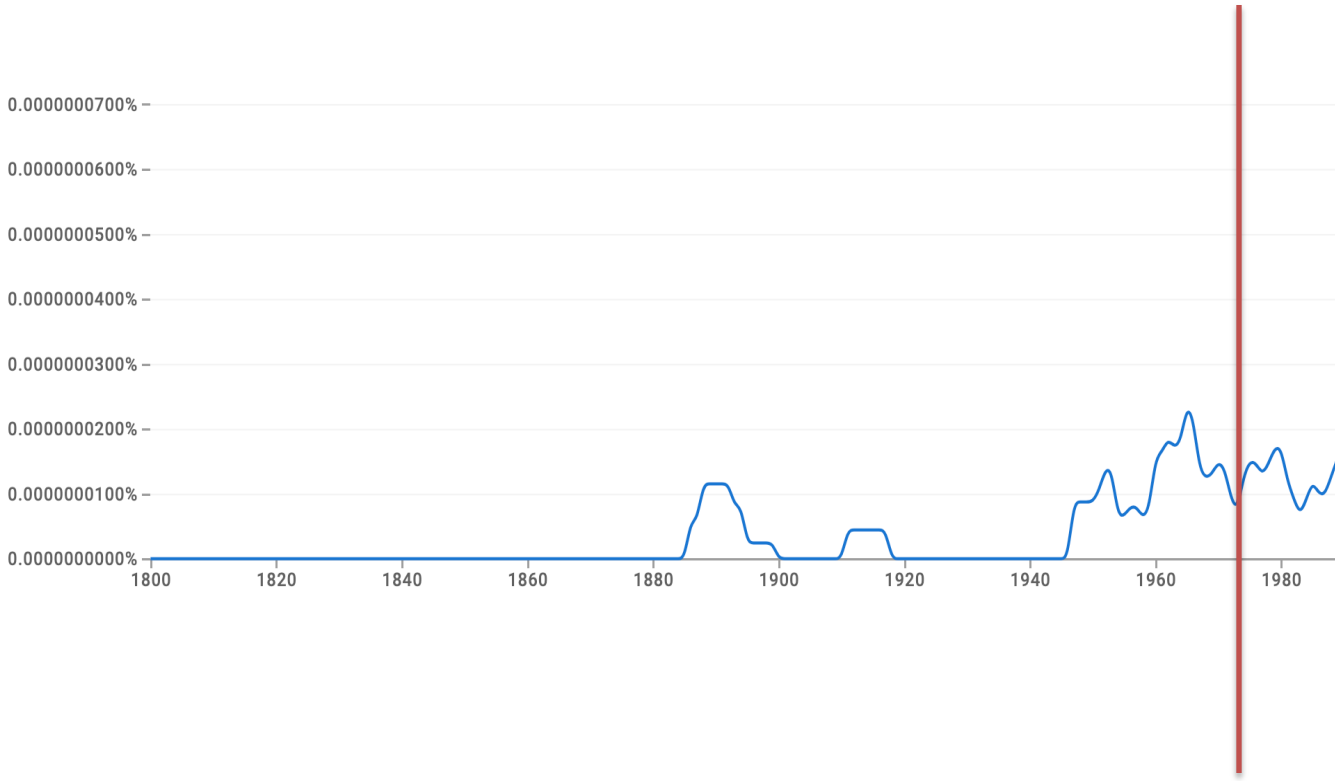
- But Moore also recognized the **supervenience** of the normative upon the non-normative:
  - No two situations, actions, persons, or practices could be different *only* in their normative features while being the same in every non-normative respect.

Bringing psychology to bear on ethics

- For example,
  - ***is* does not imply *ought***,
- but
  - ***ought* implies *can***.
- That is, one cannot be normatively *required* to do that which is impossible to do, whether in epistemology or ethics.
  - This enables empirical considerations to bear directly upon the tenability of moral assessments and ethical theories.
- But still, crossing this bridge came late …

1800-1990 – ‘moral psychology’

(Google Ngram)



Moral psychology in psychology, 1974-1999

- An informal check of two leading generalist psychology journals for the period of the first 25 years of SPP found:
  - Topics of articles in *Psychological Review*:
    - Memory: 134
    - Perception: 91
    - Morality: 2
  - In *Psychological Science*:
    - Memory: 122
    - Perception: 76
    - Morality: 6
- Looking at SPP itself …

moral, perception, cognition, consciousness
(SPP Program word counts)

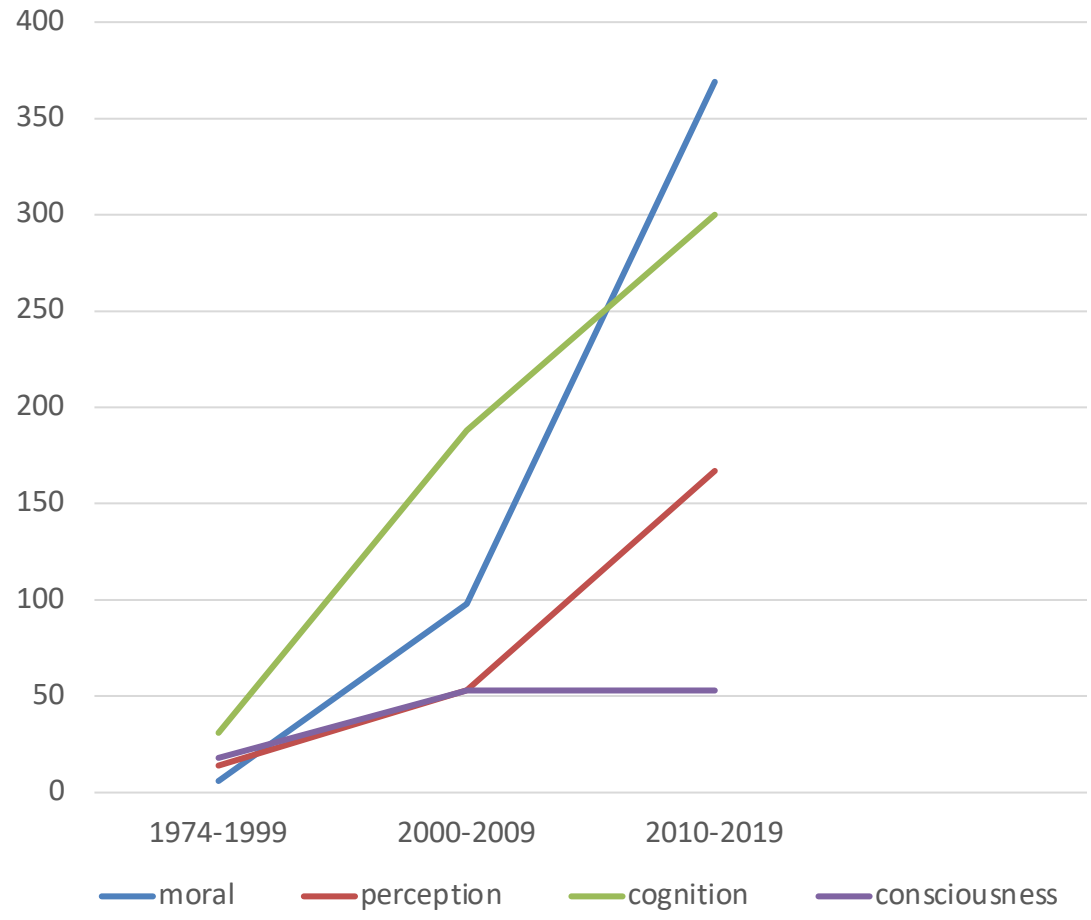
	moral	perception	cognition	consciousness
1974-1999 (incomplete)	6	14	31	18
2000-2009	98	53	188	53
2010-2019	369	167	300	53
2020-2024 (incomplete)	89	28	23	13
Total 1974-2024 (incomplete)	562	262	542	137

moral, perception, cognition, consciousness
(SPP Program word counts)

	Moral	perception	cognition	consciousness
1974-1999 (incomplete)*	6	14	31	18
2000-2009	98	53	188	53
2010-2019	369	167	300	53
2020-2024 (incomplete)	89	28	23	13
Total 1974-2024 (incomplete)	562	262	542	137

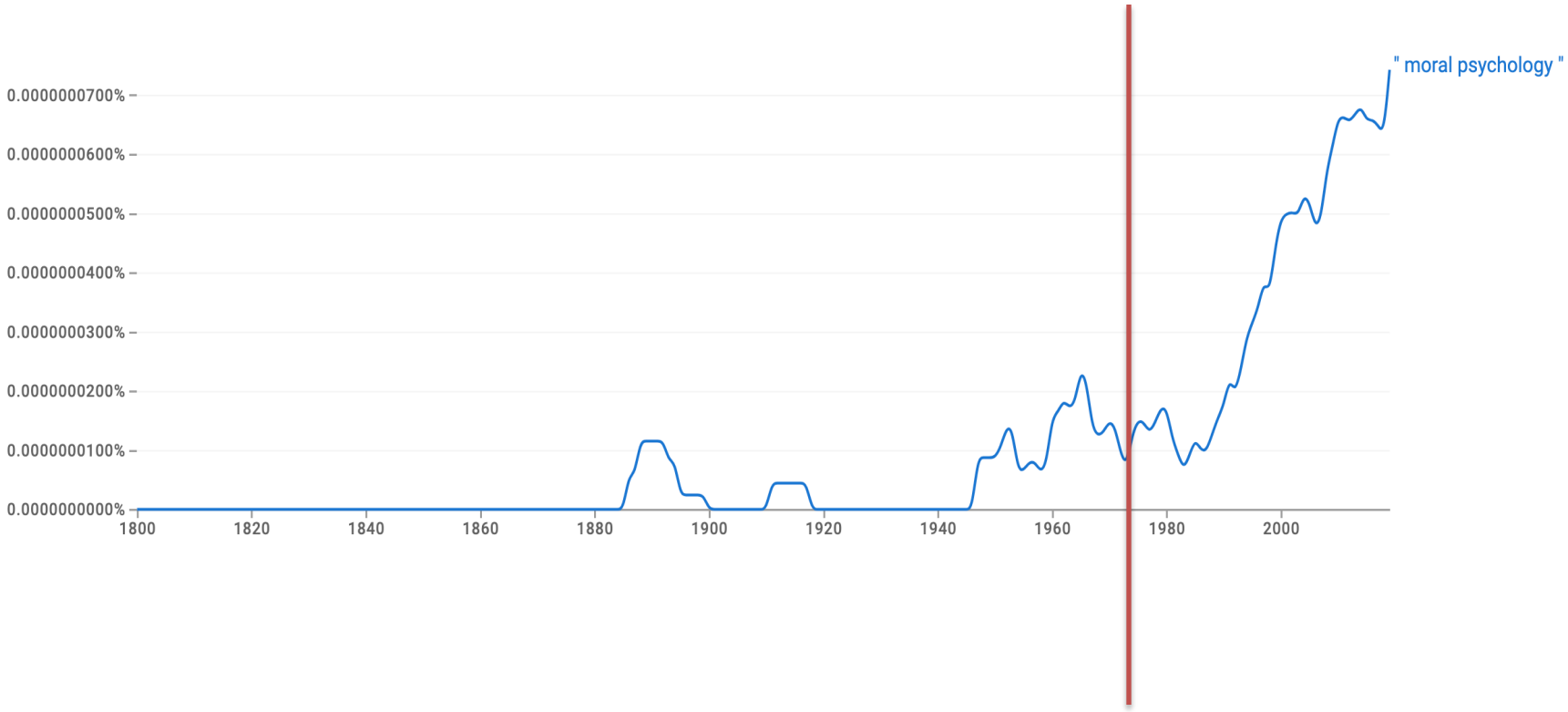
*In an analysis that had access to programs for more of the years between 1974 and 1999, David Chalmers (2024, SPP Special Session) did not find *moral* (or *ethical, value, judgment,* or *norm*) among the top 17 topics. (See also Appendix 1, below.)

moral, cognition, perception, consciousness



And more generally: 1800-2020 – 'moral psychology'

(Google Ngram)



Underwriting or undermining?

- Since *ought* implies *can*, psychological evidence can be produced that could either *underwrite* or *undermine* ordinary moral thought or ethical theory.
  - Which will it be?
- Morality—with its purport of impartiality, objectivity, universality, demands of obligation, and claims of virtue, etc., and the seeming lack of direct observation—seems to be an odd thing to find in the natural and social world, and so can be an attractive target for debunking.

Undermining?

- So it is unsurprising that once discussion of morality began to take hold in SPP, we see:
  - Evolutionary arguments to undermine the possibility of moral motivation or cognition, suggesting an Error Theory.
  - Evidence that folk moral notions such as free agency and responsibility are incoherent and cannot be satisfied, to undermine moral claims that depend upon them.
  - The Situationist critique of determinate character traits, to undermine certain forms of Virtue Theory.
  - Non-moral explanations of the source of seeming contradictions in ordinary moral evaluations.

Underwriting?

- However, there are a number of areas where, I believe, ethical theory and moral psychology have not been on a collision course.
- Rather, what initially appeared to be an area of sharp conflict emerged, over a shared, 50-year evolution in both fields, as a potential source of underwriting:
  - (1) The connection between moral judgment and action
  - (2) The nature and status of moral intuitions
  - (3) The role of learning in moral development and competence

Unsettling?

- (4) These three areas also lead us naturally into a fourth, which was until recently a highly speculative realm of inquiry in ethical theory and moral psychology: the realm of artificial systems and agents and our interactions with them.
  - Things here are very uncertain, to say the least, but it is none too early to start readying ourselves as philosophers and psychologists for new opportunities and perils.

(1) Moral judgment and action

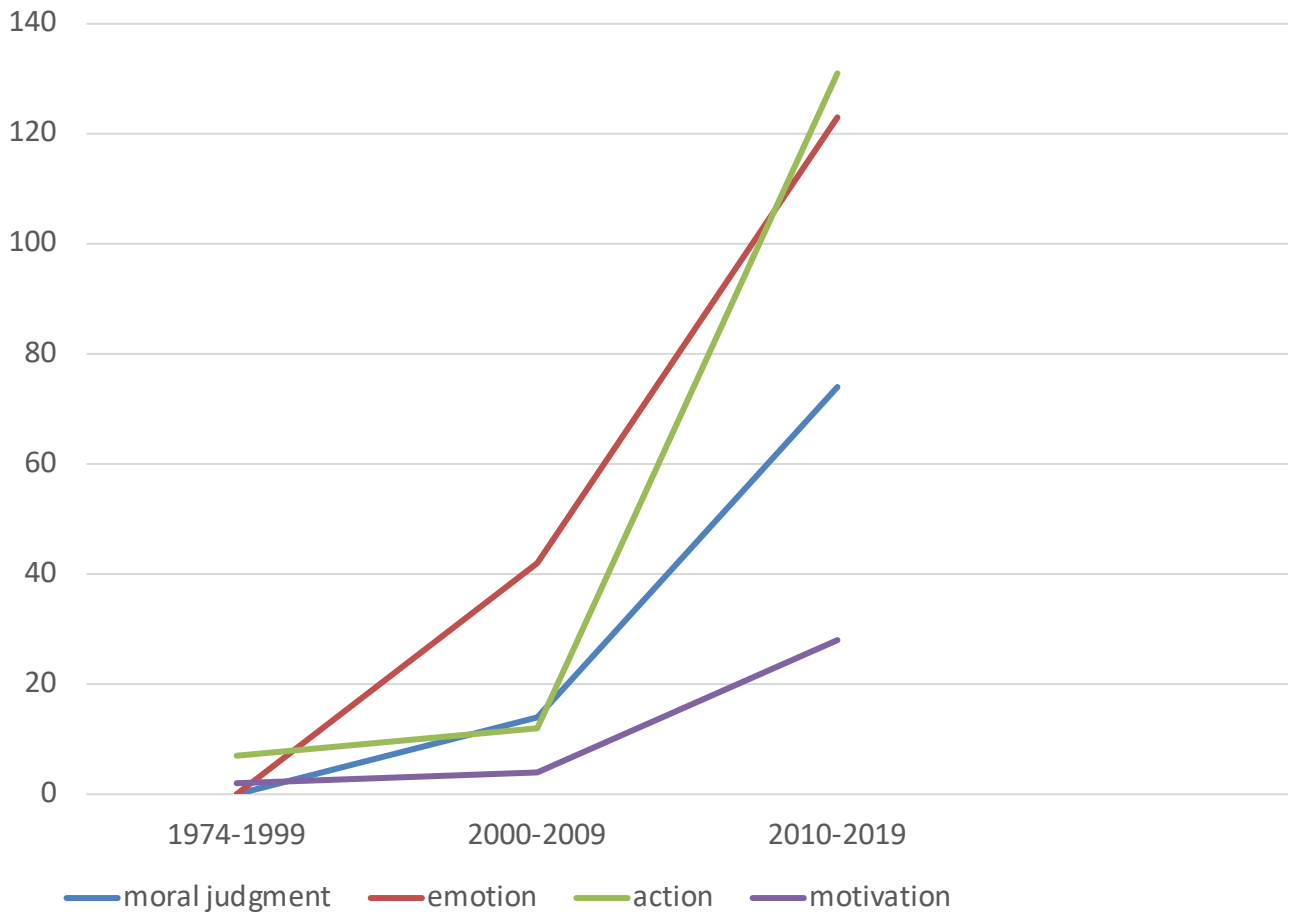
"There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy." (*Hamlet*)

Moral judgment, motivation, emotion, action
(SPP program word counts)

	moral judgment	motivation	emotion	action
1974-1999 (incomplete)*	0	2	0	7
2000-2009	14	4	42	32
2010-2019	74	28	123	131
2020-2024 (incomplete)	8	6	11	26
Total 1974-2024 (incomplete)	96	40	176	196

*In an analysis that had access to programs for more of the years between 1974 and 1999, David Chalmers (2024, SPP Special Session) did not find *moral, judgment, emotion, or action* among the top 17 topics. (See also Appendix 1, below.)

moral judgment, motivation, emotion, action



(1) Moral judgment, motivation, and action

- As the number of papers on moral subjects grew at SPP, a key focus was on the nature and status of moral judgment, for example:
  - What kind of mental state do moral judgments express?
  - In what sense, if any, are moral judgments cognitive or potentially objective?

- Partly these questions reflect a long-standing dispute within Ethical Theory.
  - Since Aristotle, forming a moral judgment has been seen as intimately connected to actual *action-guidance*

"Direction of fit"

- This, however, poses a challenge to *cognitivist* views of moral judgment.
  - Cognitions are representational, capable of truth or falsity, or accuracy or inaccuracy.
    - They have "mind-to-world" direction of fit.
  - But representational states alone do not suffice for action, since some form of motivation is needed. Motivations, however, are not capable of truth or falsity, accuracy or inaccuracy—their role in action is that of goal-setting.
    - They have "world-to-mind" direction of fit.

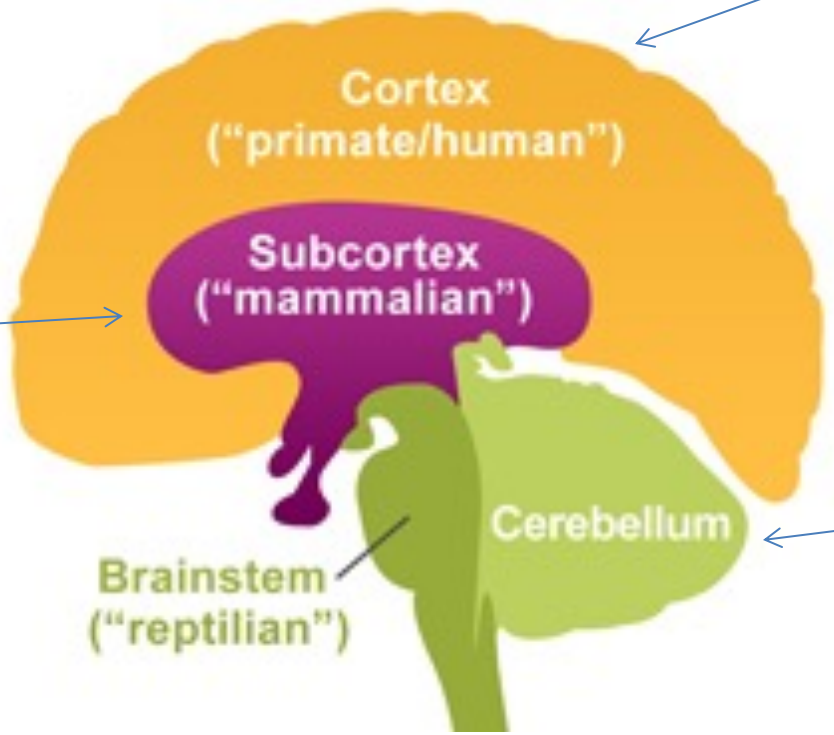
Emotion, cognition, and rational guidance

- By the midpoint of SPP, however, we find growing philosophical and psychological debates over the nature of emotion, and, increasingly, the relation of emotion to cognition and morality.
- In particular, while some philosophers and psychologists insisted on the non-cognitive nature of emotion, others increasingly questioned the distinctness of cognitive and emotive states and processing.
  - Affective responses—both aroused and default—present situations as having certain significant features *and* have a coordinated effect on physiology, attention, thought, motivation, action-readiness, etc., in ways that orient or reorient the individual toward these significant features.

Cognition, emotion, valuation, and rational guidance

- Many affective states can be spoken of as more or less accurate, warranted, rational, well-calibrated, etc., in how they present the world, and play a central role in learning.
  - That is, they have some ”mind-to-world” direction of fit,
- At the same time, they can also provide motivating action-guidance.
  - That is, they have some ”world-to-mind” direction of fit.
- The affective system is an *evaluative* system, and in rational agents value should be playing this sort of pivotal role.

The brain I was brought up on:

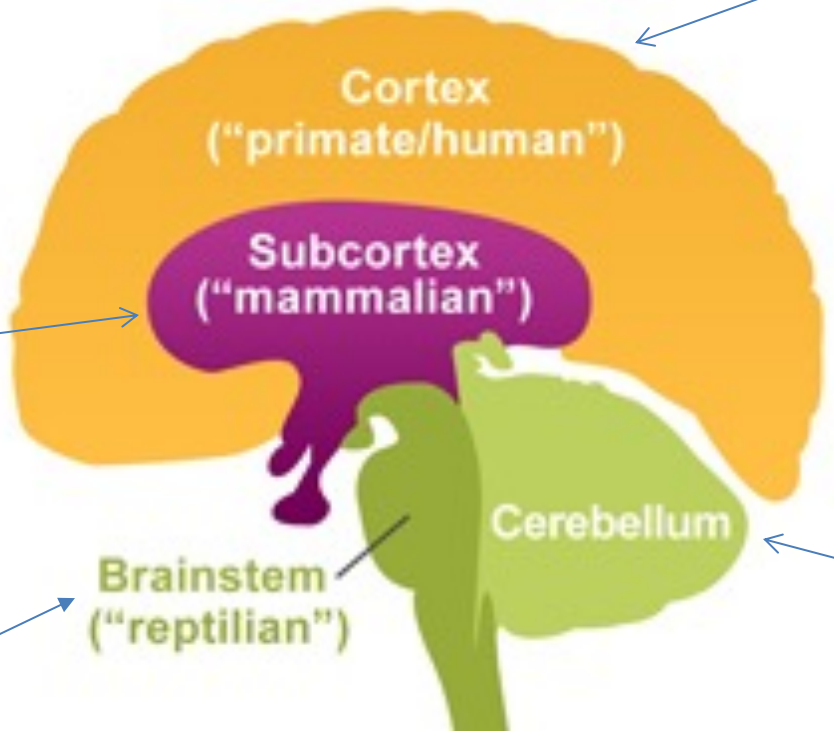


Cognitive System "Rational Thought"

Limbic System "Emotion"

Autonomic System

The brain we now think we have:



Cognitive and perceptual information processing; associative learning; declarative reasoning

Affect, evaluation, social guidance, deliberation, memory, reinforce-ment learning

Planning of motion and motor learning, emotion regulation

Bodily regulation

Visual processing and executive control
(Pessoa 2008)

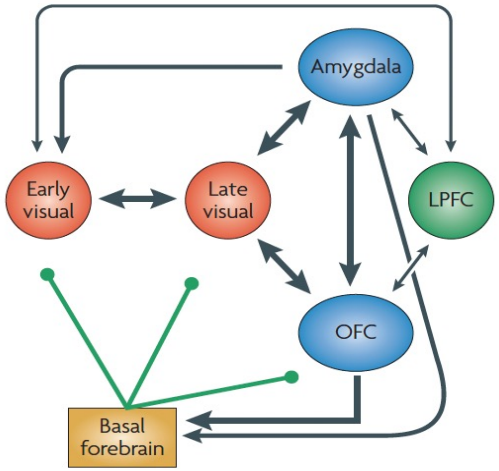


Figure 2 | **Circuit for the processing of visual information.** The affective component of a visual item is reflected at multiple processing stages, from early visual areas (including V1) to prefrontal sites. Diffuse, modulatory effects exerted by the basal forebrain are shown in green. Crucially, cognitive and emotional contributions cannot be separated. For instance, visual cortical responses reflecting an item's significance will be a result of simultaneous top-down modulation from fronto-parietal attentional regions (see lateral prefrontal cortex (LPFC)–early visual connections) and emotional modulation from the amygdala (see amygdala–early visual connections). In this manner, the cognitive or affective origin of the modulation is lost and the item's impact on behaviour is both cognitive and emotional. Several connections are not shown to simplify the diagram. Line thickness indicates approximate connection strength. OFC, orbitofrontal cortex.

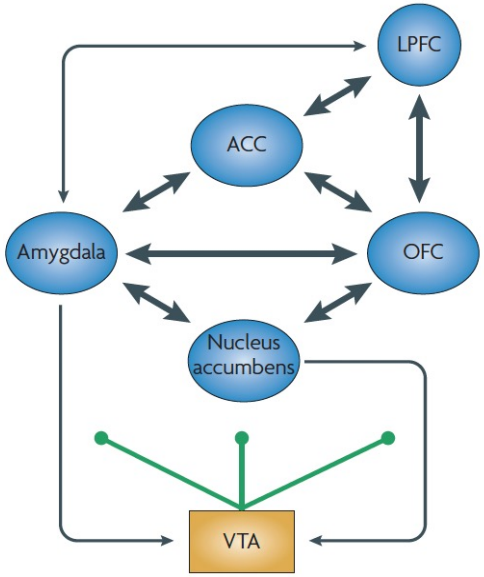


Figure 3 | **Circuit for executive control.** This extended control circuit contains traditional control areas, such as the anterior cingulate cortex (ACC) and the lateral prefrontal cortex (LPFC), in addition to other areas commonly linked to affect (amygdala) and motivation (nucleus accumbens). Diffuse, modulatory effects are shown in green and originate from dopamine-rich neurons from the ventral tegmental area (VTA). The circuit highlights the cognitive–affective nature of executive control, in contrast to more purely cognitive-control proposals. Several connections are not shown to simplify the diagram. Line thickness indicates approximate connection strength. OFC, orbitofrontal cortex.

A lesson?

- Beware tempting ways of thinking—reason vs. emotion, higher vs. lower, rational vs. animal, cognitive vs. affective, representational vs. non-representational—that prematurely close the possibilities one can conceive or consider.
  - Making distinctions is an inevitable starting point for theory development and application, but they should be a scaffold, not a cage.
- We may think we have swept out the whole of conceptual space *a priori* with our category schemes, but there may be "more things in heaven and earth than we have dreamed of in our philosophies,
  - … and the combined efforts of philosophers and psychologists helped us find them.

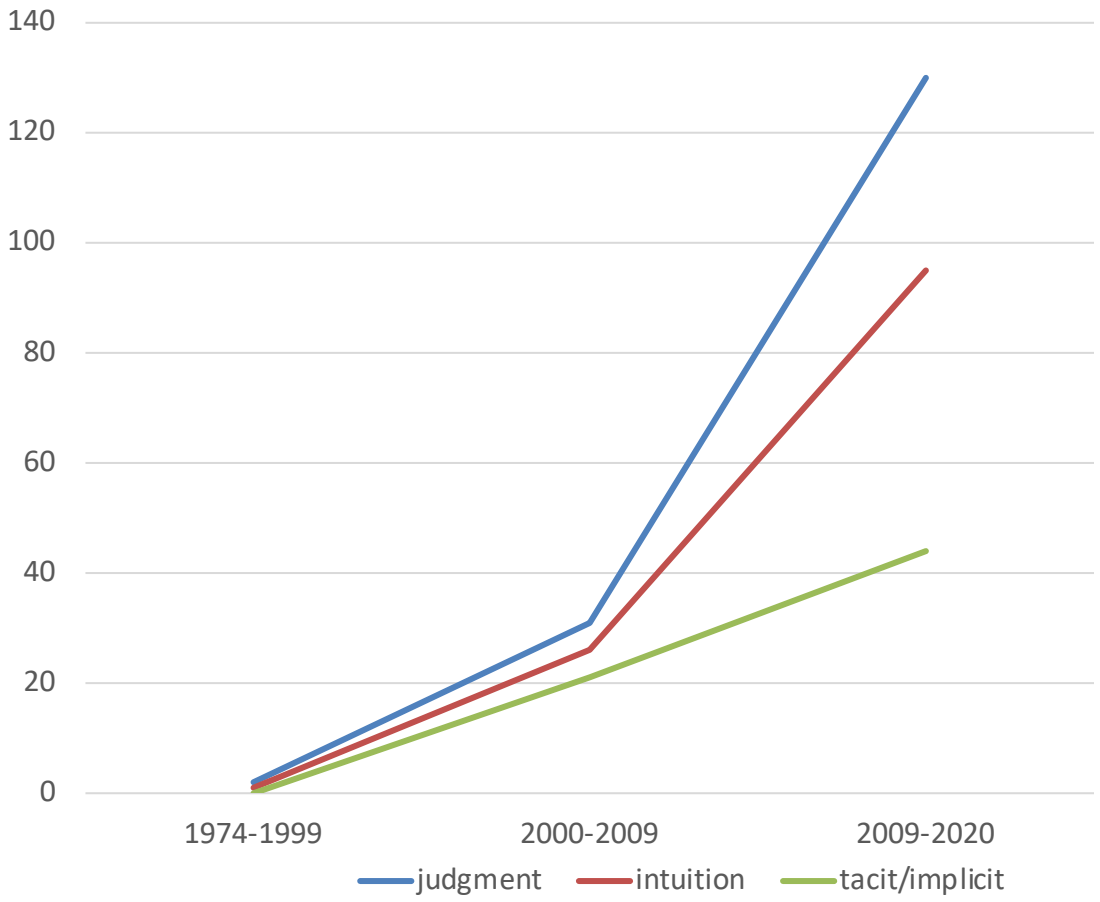
(2) The nature and standing of moral intuitions

judgment, intuition, tacit/implicit
(SPP program word counts)

	judgment	intuition	tacit/implicit
1974-1999 (incomplete)*	2	1	0
2000-2009	31	26	21
2010-2019	130	95	44
2020-2024 (incomplete)	27	7	9
Total 1974-2024 (incomplete)	190	129	74

*In an analysis that had access to programs for more of the years between 1974 and 1999, David Chalmers (2024, SPP Special Session) did not find *judgment, intuition,* or *tacit/implicit* among the top 17 topics. (See also Appendix 1, below.)

judgment, intuition, tacit/implicit



(2) The nature and standing of moral intuitions

- Intuitions are usually conceived of as:
  - explicit or implicit thoughts, feelings, or judgments that can non-deliberatively come to mind and guide perception, thought, feeling, and action.
  - while they might be the product of extensive experience or training, they typically can come to mind rapidly and spontaneously, though without much insight into their origins.
  - characteristically, they come to mind with a certain felt credibility—they "seem" right or apt, and it can feel like a mistake to ignore them altogether.

"Intuitions"

- If Aristotle is right, "intuitions" are indispensable in thought and judgment—unless some basic premises, patterns, values, or principles struck us as intuitively right, without requiring proof or demonstration, then our thinking could never get underway, or carry conviction (*Posterior Analytics*).
  - This does not make intuitions indubitable, just initially credible.

However, …

- Despite their seeming indispensability in commonsense thought and philosophical and psychological theorizing,
  - … just as SPP was being born, intuitions became subject to a sustained critical examination by Tversky, Kahneman, and others.
  - While they did not challenge the idea that intuitions are often useful and appropriate, they developed a picture of intuitive thought—"heuristics and biases"—that cast doubt on the extent of our reliance upon it.

Dual-system Psychology

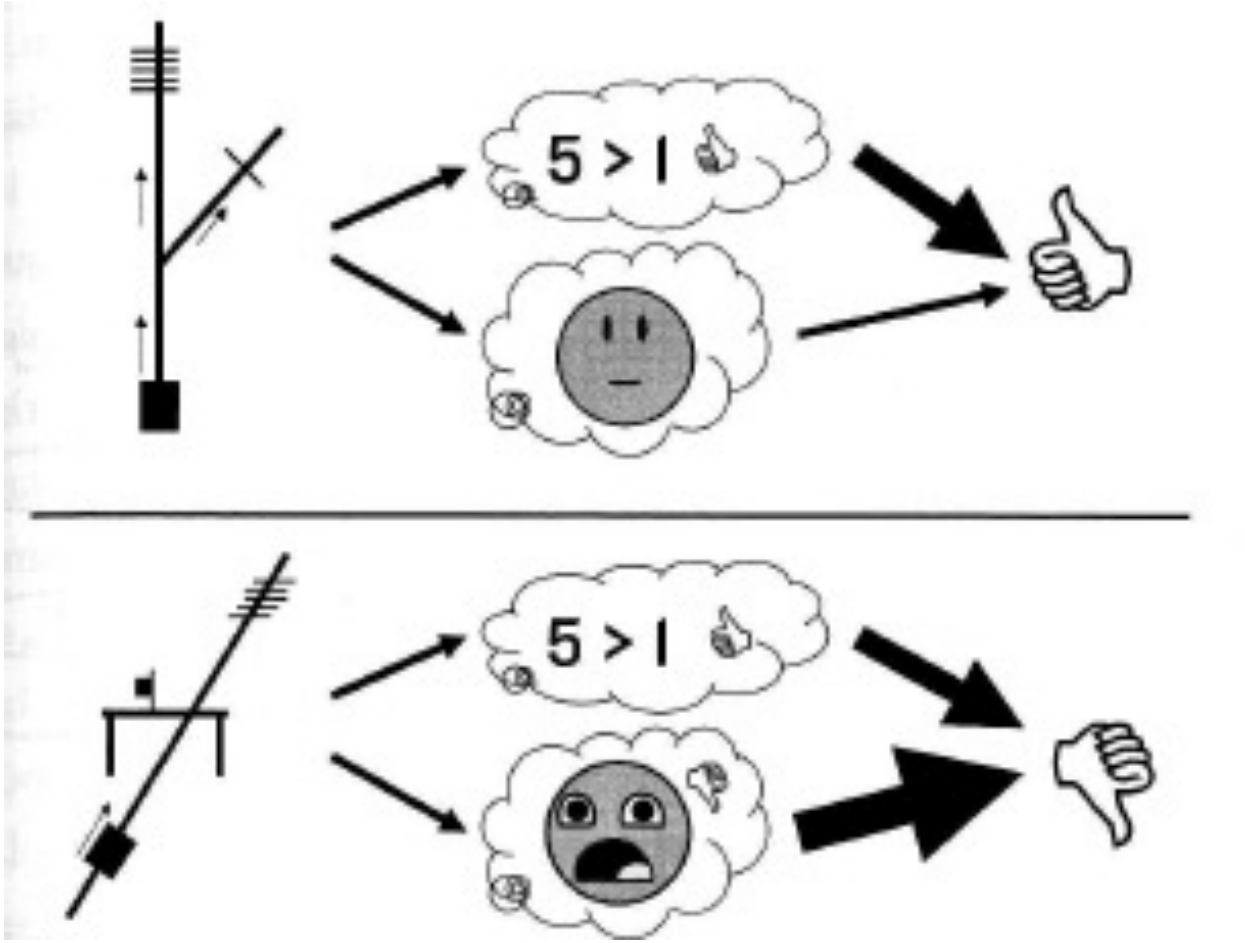
- "Dual system" or "dual process" psychology came into its own. This proved to be another highly attractive idea.
  - An intuitive, effortless, often affectively-charged System 1, innocent of statistics and capable of handling only a few things at once
  - A deliberative. effortful System 2, capable of statistical and logical reasoning, and balancing considerations rationally.

Dual processing moral psychology

- In the moral realm, dual processing theories offered a way of understanding seeming inconsistencies in moral judgment, "moral dumbfounding", and other worrying patterns in commonsense moral judgment.
  - The prime target for the critique of intuitions became the Trolley Problem, and here the critique promised to do *normative* as well as empirical work.

Dual-process trolleyology

(Greene, 2013)



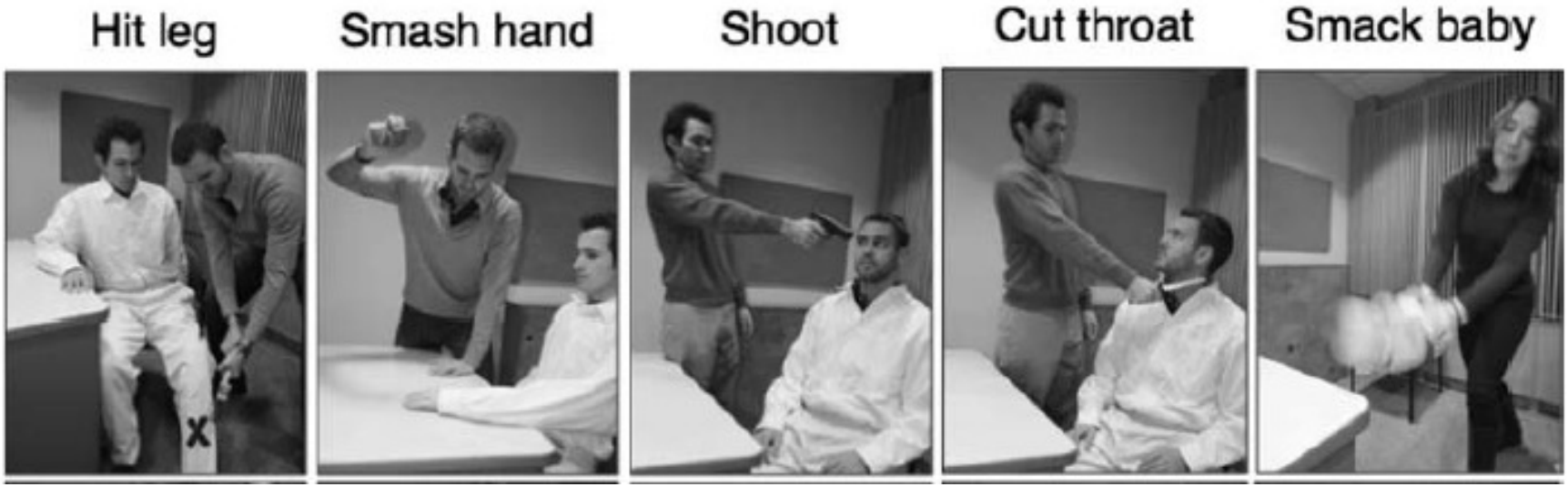
(3) The role of learning in moral development and competence

- At this time, moral judgment was frequently seen as modular, lending strength to the idea of an intuitive system fairly impermeable to being re-educated by our more deliberative capacities, hence the seeming intransigence of the puzzling Trolley Problem pattern of responses.
  - But that, too, was starting to change.

Learning-based moral psychology

- For reasons partly considered in the previous discussion of the dualism of cognition and emotion, the dual processing model of the mind has come under sustained criticism.
  - The mind does not appear to be assembled from two such systems, and while multiple processes are at work, they are not organized, and do not function, as the dual-system approach required.
- A shift is occurring toward a more integrated theory of mind and a different kind of duality that has firmer grounding in neuroscience—a duality in learning between model-free vs. model-based learning.

Model-free learning (Cushman *et al.*, 2012)

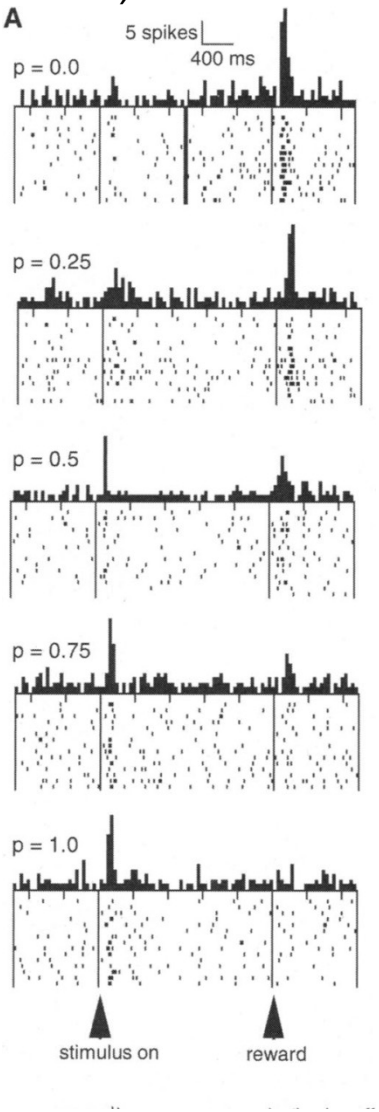


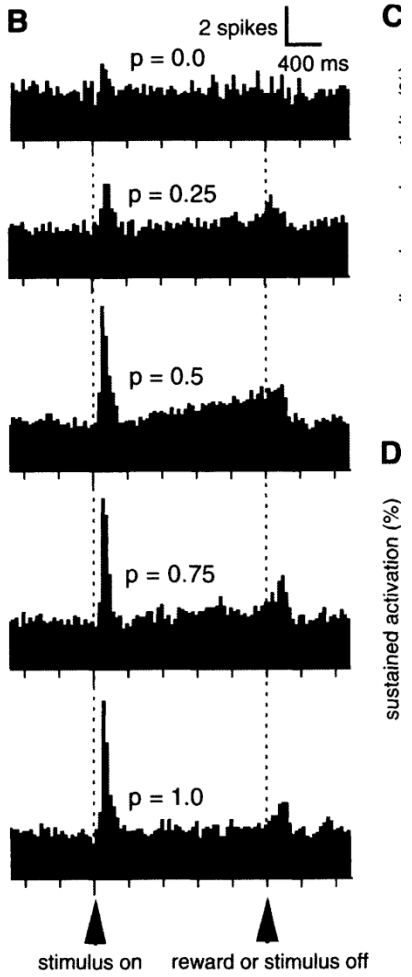
Foraging for information and value

- At the same time, detailed neuroscientific work on tasks like navigation and action-selection in the face of uncertainty and reward variability in rats and other animals began to suggest that our "inherited" intuitive system is actually statistically highly competent, and capable of constructing and operating with well-calibrated models of the environment.
  - Widely-observed near-optimal foraging behavior, which combines multiple variables and sources of value and risk, now seems to have an explanation in the capacity of foragers to learn and use such multivariate causal-evaluative models.

Expectation and risk

(Fiorillo *et al.*, 2003)

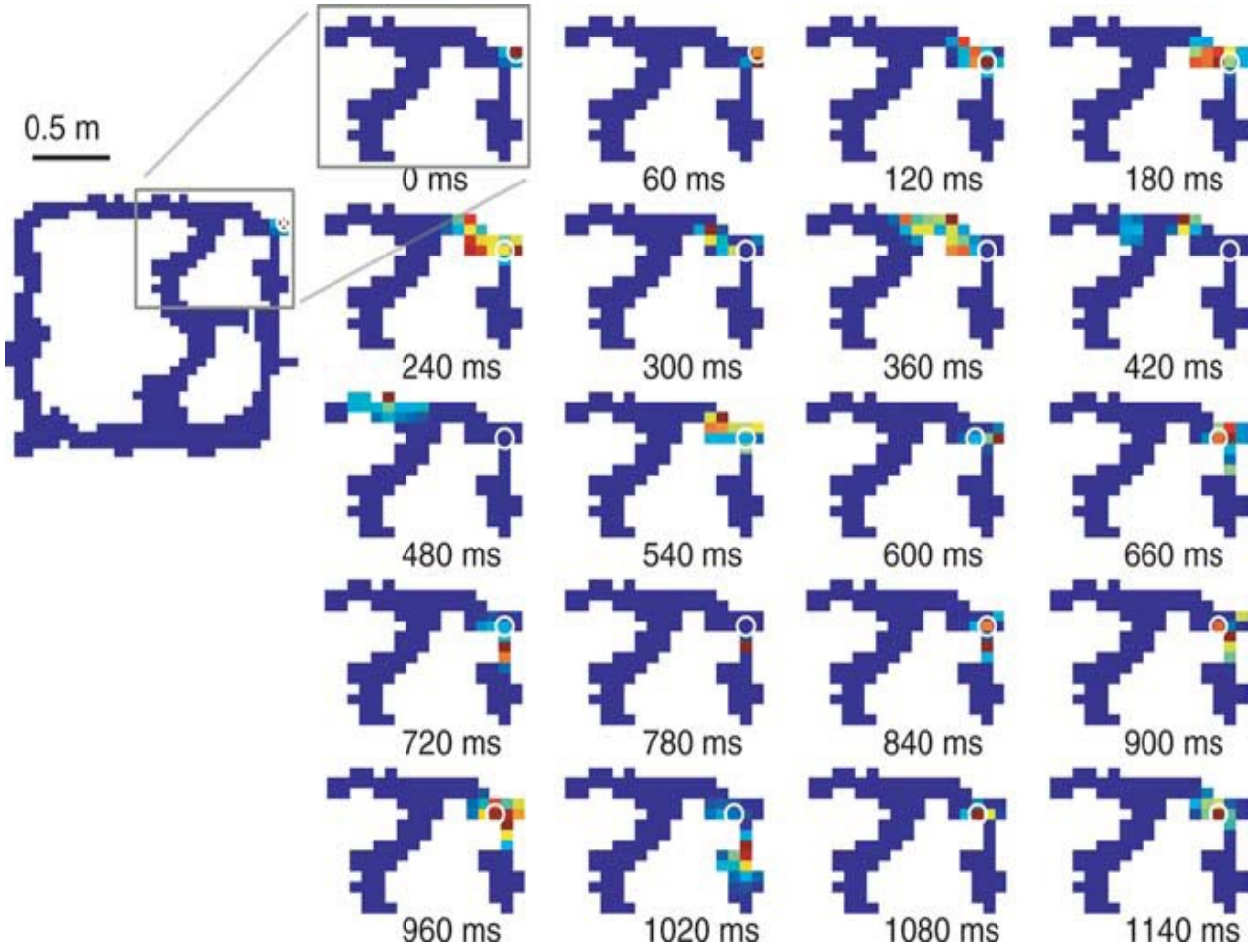




40

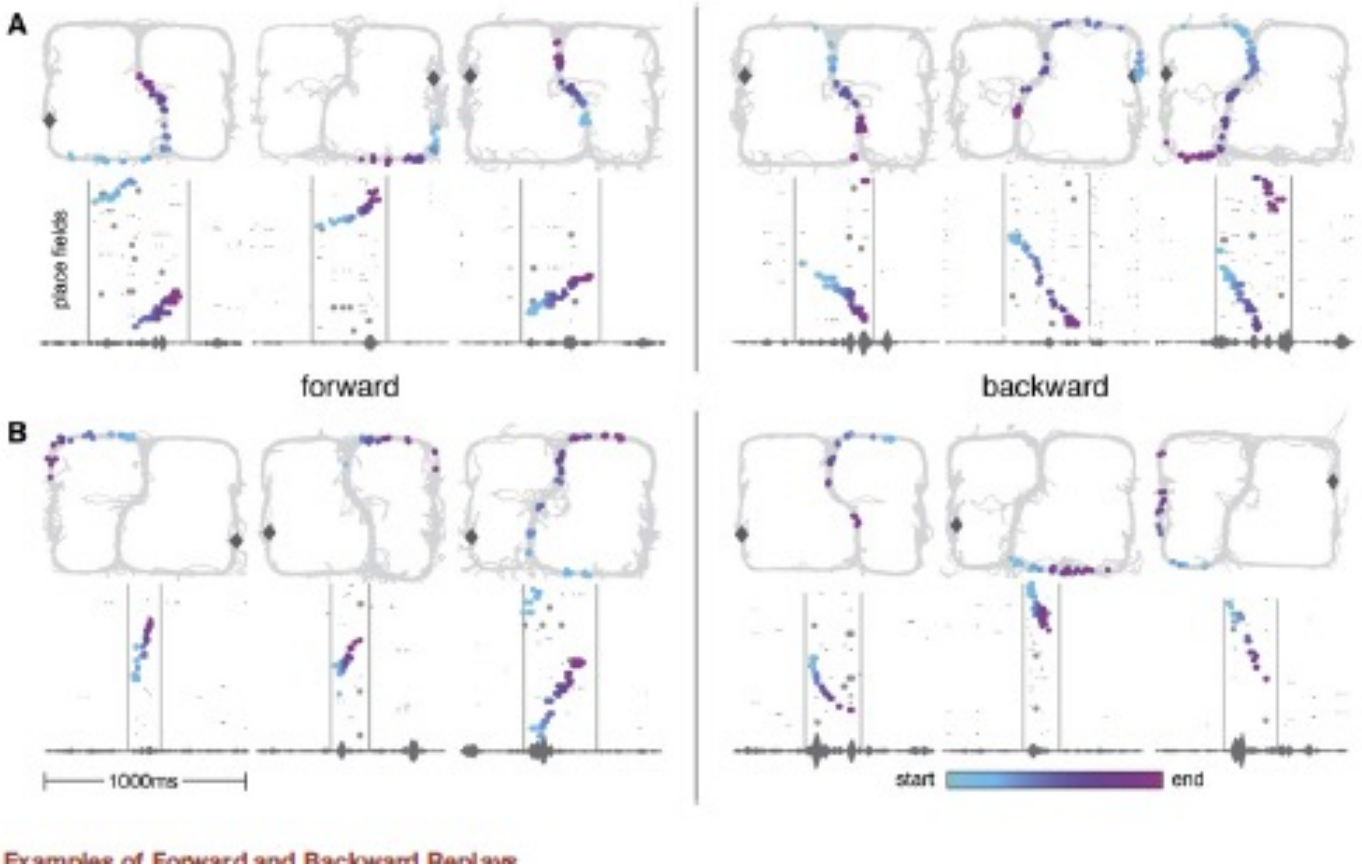
A rat following an evaluative representation

(Johnson & Redish, *J Neurosci* 2007)



41

Hippocampal construction of novel paths in sleep (Gupta *et al.*, 2010)



Mapping value

- Accurate mapping of the value structure of the environment is as central to an animal's success as accurate mapping the physical structure of the environment.
- For highly social animals, this should hold for navigating the social environment as well, giving rise to the possibility that the "intuitive" system we inherit is capable of accurate mapping of the evaluative structure of the social environment and the agents and actions in it.
- And developmental studies suggest that infants indeed start learning the expected value of third-party social interactions early on. This seems to proceed integrally with causal learning and theory of mind development.

Going back to the trolleys …

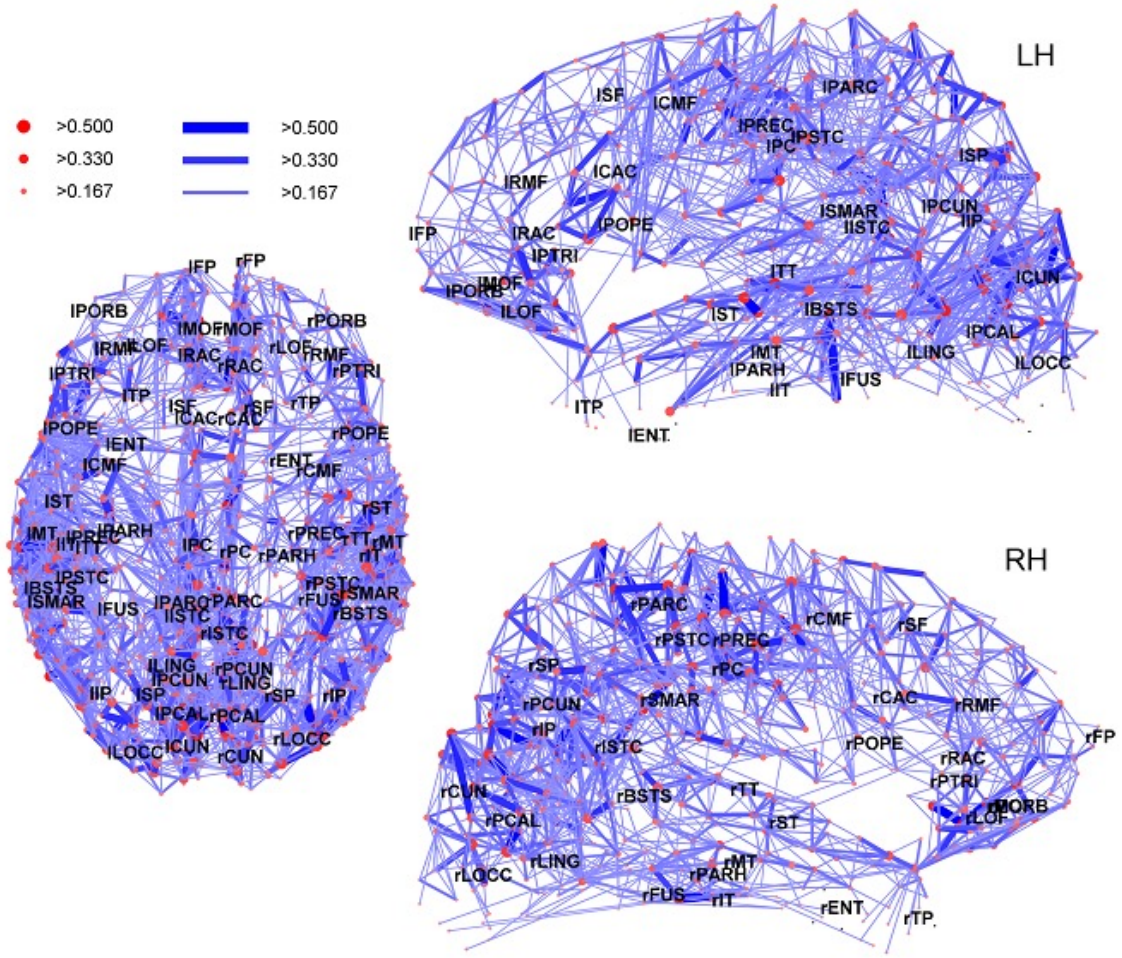
- Might evaluative *social* modeling play a role in explaining puzzling Trolley Problem intuitions?
  - About Footbridge: Numerous studies found that likelihood of giving push-like verdicts in Footbridge-like scenarios was correlated, not with impartial altruism, but with rating on psychopathy scales, egoism, disregard for ethical transgressions, and lack of perspective-taking.
- Perhaps intuitive reactions to the Footbridge scenario do reflect model-based learning—concerning agents in our environment, and which ones to trust or not.

What should we then expect to see …

- … as the infrastructure subserving moral judgment?

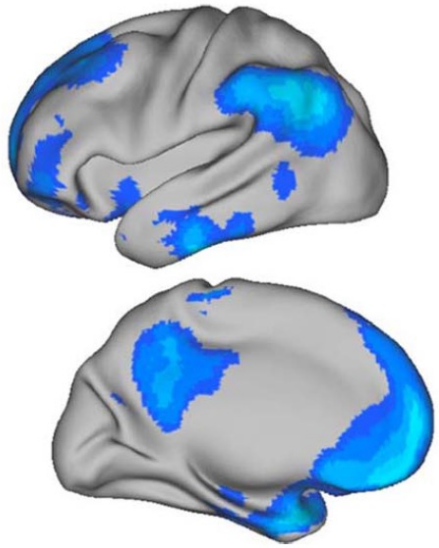
Connectomic view of the brain

(Hagman *et al.*, 2008)



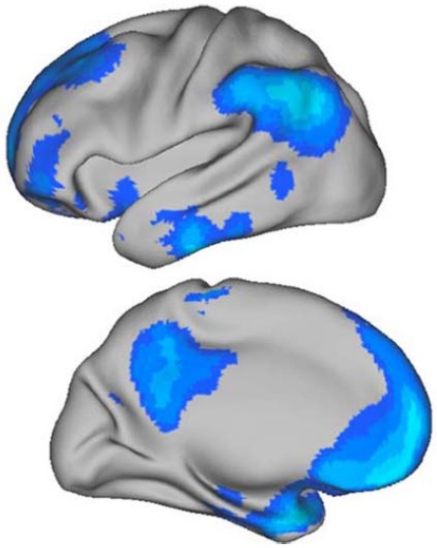
Default network

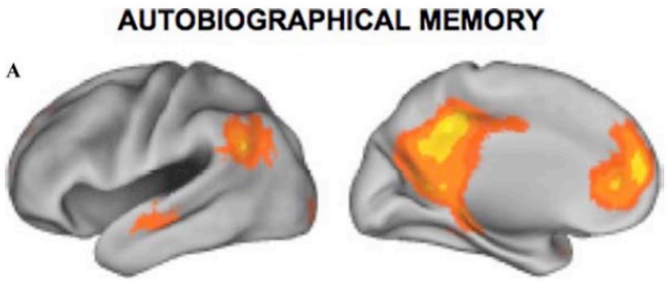
(Buckner *et al.*, 2008)



Default network

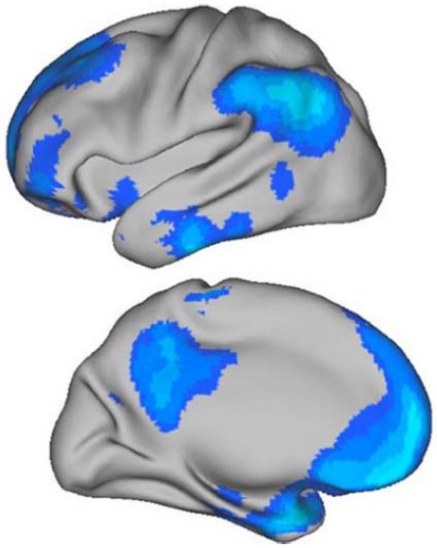
(Buckner *et al.*, 2008)

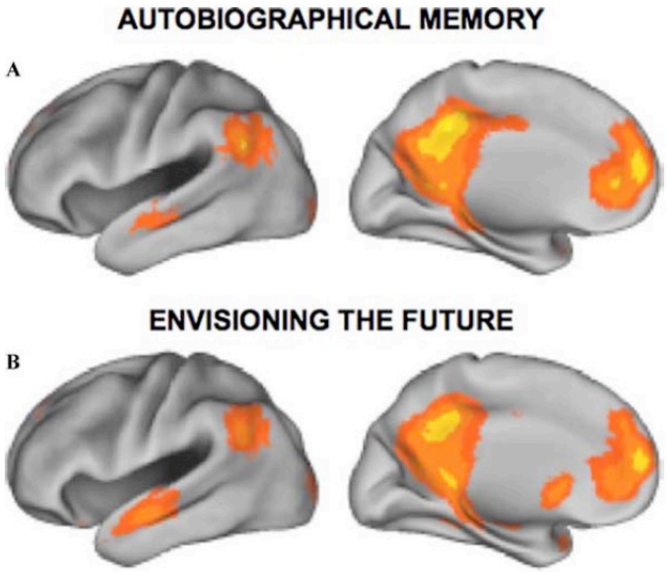




Default network

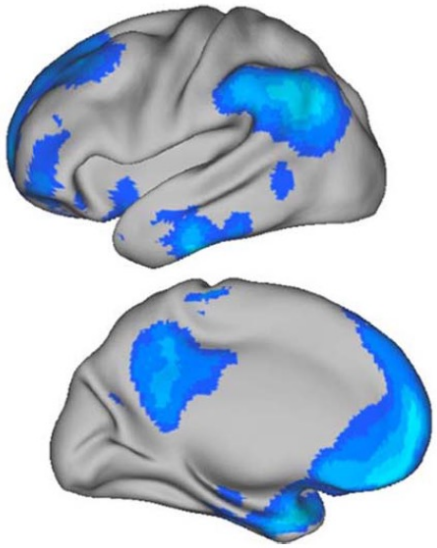
(Buckner *et al.*, 2008)

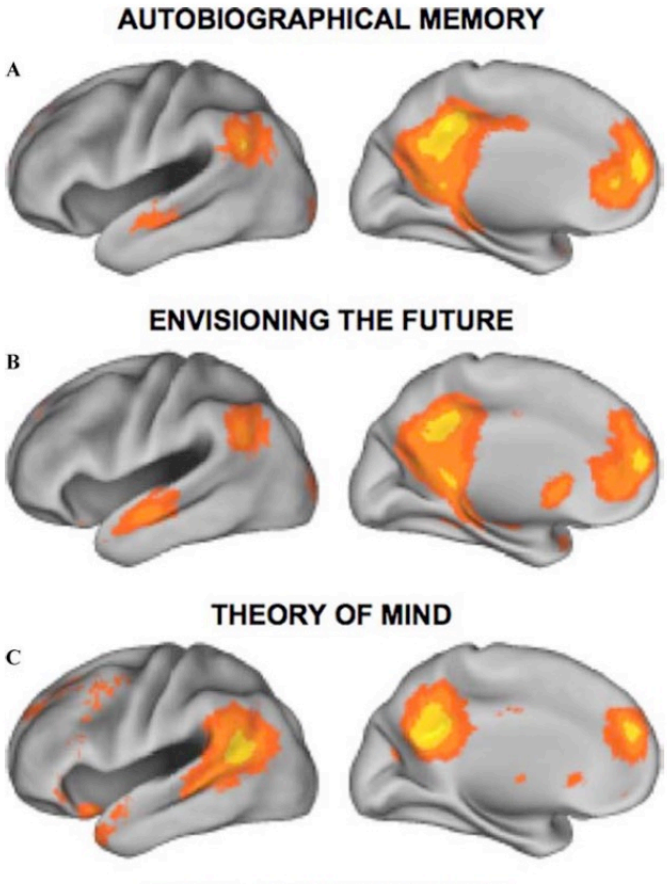




Default network

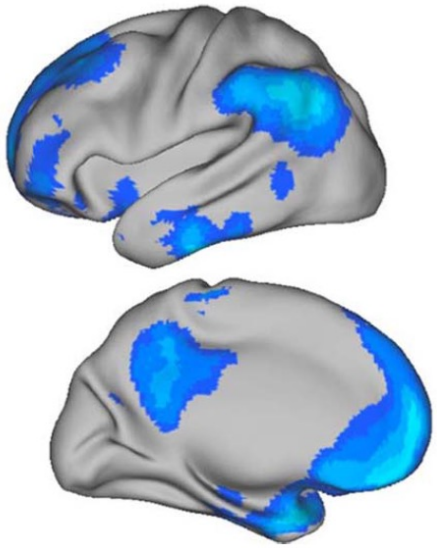
(Buckner *et al.*, 2008)

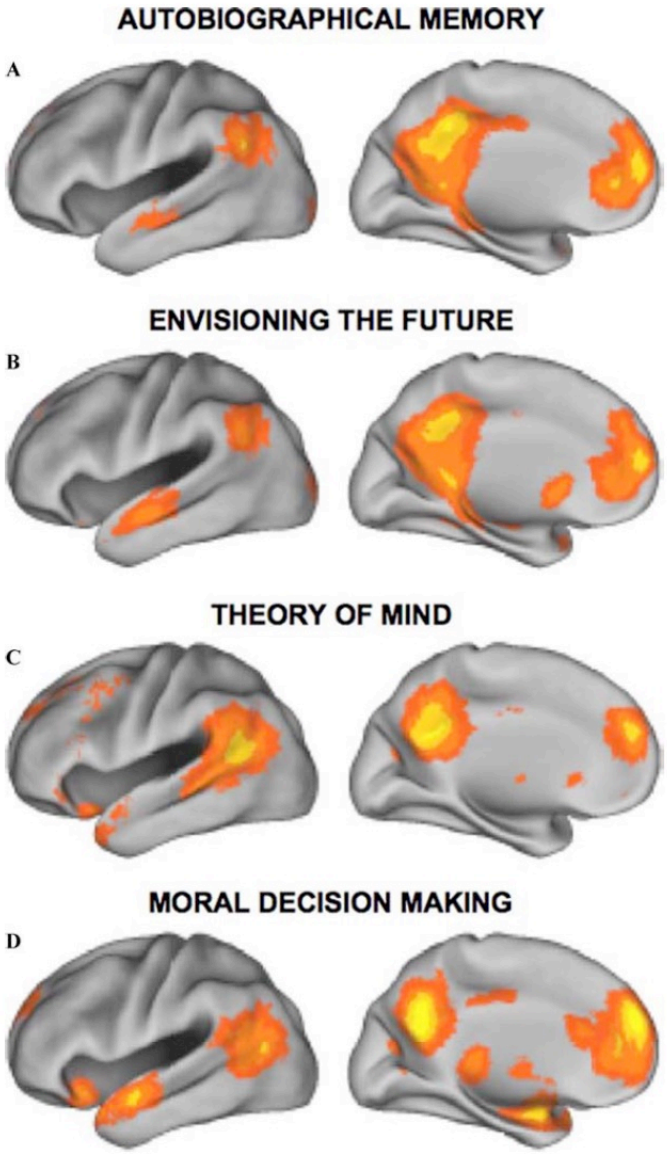




Default network

(Buckner *et al.*, 2008)





A lesson?

- Do not underestimate the intuitive learning and modeling capacities of animals and infants … or adult humans.
- While morality has often been seen as a singular exception in the natural world, tempting theorists to posit innate moral capacities or "modules" distinctive to humans,
  - … if we think of morality as a set of cognitive, affective, and practical capacities and practices that enable agents to live together better than they could live apart, we might be less tempted to think of morality as discontinuous with general intelligence, problem-solving, and learning, or confined to creatures with human psychology.
  - And that brings us to …

(4) Artificial moral psychology

- In the trends (1)-(3), we see a pattern that resembles the evolution of artificial intelligence

Intelligence and agency in machines

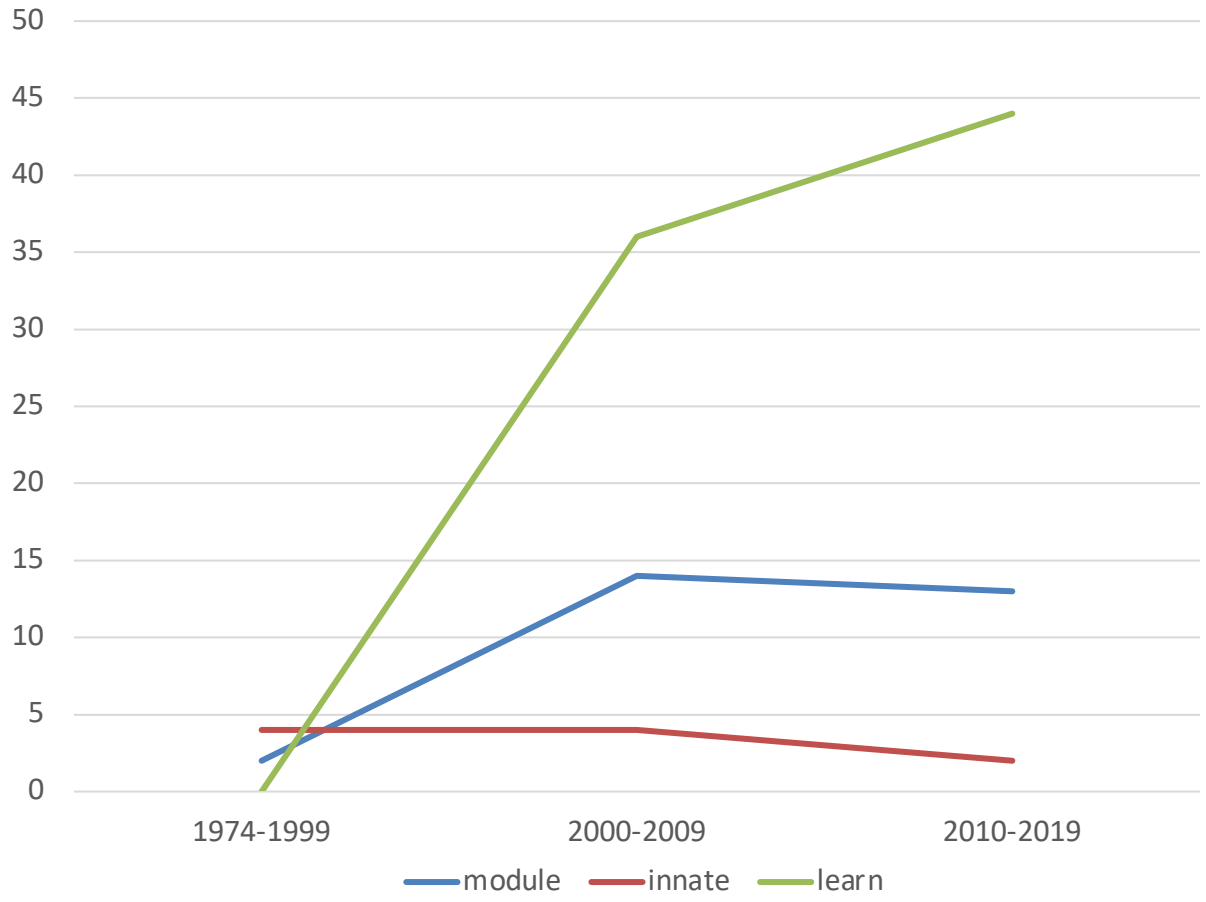
- Recent developments in "artificial intelligence" are based upon learning processes of kinds analogous to those of humans and animals—self-supervised pattern recognition and modeling, and reinforcement learning using simulation as well as feedback.
- When LLMs are trained on massive textual data, the competencies that emerge bear some of the marks of generalizable intelligence—spanning, e.g., object recognition, generative capacities in language and images, modeling physical processes, and motor control.
  - Without calling upon in-built modules or principles.

module, innate, learn
(SPP program word counts)

	module	innate	learn, learning
1974-1999 (incomplete)*	2	4	0
2000-2009	14	4	36
2010-2019	13	2	44
2020-2024 (incomplete)	0	1	25
Total 1974-2024 (incomplete)	29	11	105

*In an analysis that had access to programs for more of the years between 1974 and 1999, David Chalmers (2024, SPP Special Session) found *innate*, but not *learn/learning*, among the top 17 topics.

module, innate, learn



But …

- … don't artificial neural networks have only a shallow representation of language or the world, with well-known problems of confabulation, lack of groundedness?
  - Aren't they vast networks of artificial neurons connected by probabilistic weights?—There are no propositions, no rule-based reasoning, … .
- Yes, all that and more.

Living and learning with artificial agents?

- Yet these very systems will be giving us a new window into such phenomena as attention, perception, cognition, motivation, action, the nature of intuitive knowledge, and, above all, learning.
  - And, just as philosophers and psychologists have had to challenge preconceptions about emotion and "mere intuition", we may have to challenge preconceptions about what must be innate vs. what can be learned, or about how reasoning and causal inference can work, or about what are the conditions for stable, mutually-beneficial collaboration among intelligent agents, etc.

Intelligence is not holding a mirror to nature …

- … but an inventive capacity to use past experience to solve problems in novel circumstances, projecting beyond what we've learned, and recombining and generalizing information.
  - Projection and generalization, moreover, prime us for further, error-based learning, which can contribute to more robust generative capacities.
- It is this continuing, projection-based engagement with the physical and social world that serves to ground human thought and feeling—and that, of course, is what most existing LLMs lack:
  - Multi-modal, causally- and socially-engaged agents, interacting with a non-textual reality.

A proof of possibility?—Intelligent machines without pre-existing norms or institutions

- Multi-agent learning suggests that, under somewhat realistic conditions, artificially intelligent agents, too, can learn social-contract-like norms and practices. They can:
  - form communities of information sharing resistant to invasion by opportunistic agents (Köster *et al.* 2020),
  - converge on equilibrium solutions through repeated play (Marris *et al.* 2021) and cooperate effectively in tasks requiring teamwork (Balachandar *et al.* 2019)
  - develop a system of shared symbols for communication use where none existed before (Lazaridou *et al.* 2020)
  - solve certain problems of public goods and sustainable resource use (Perolat *et al.* 2017)

However, …

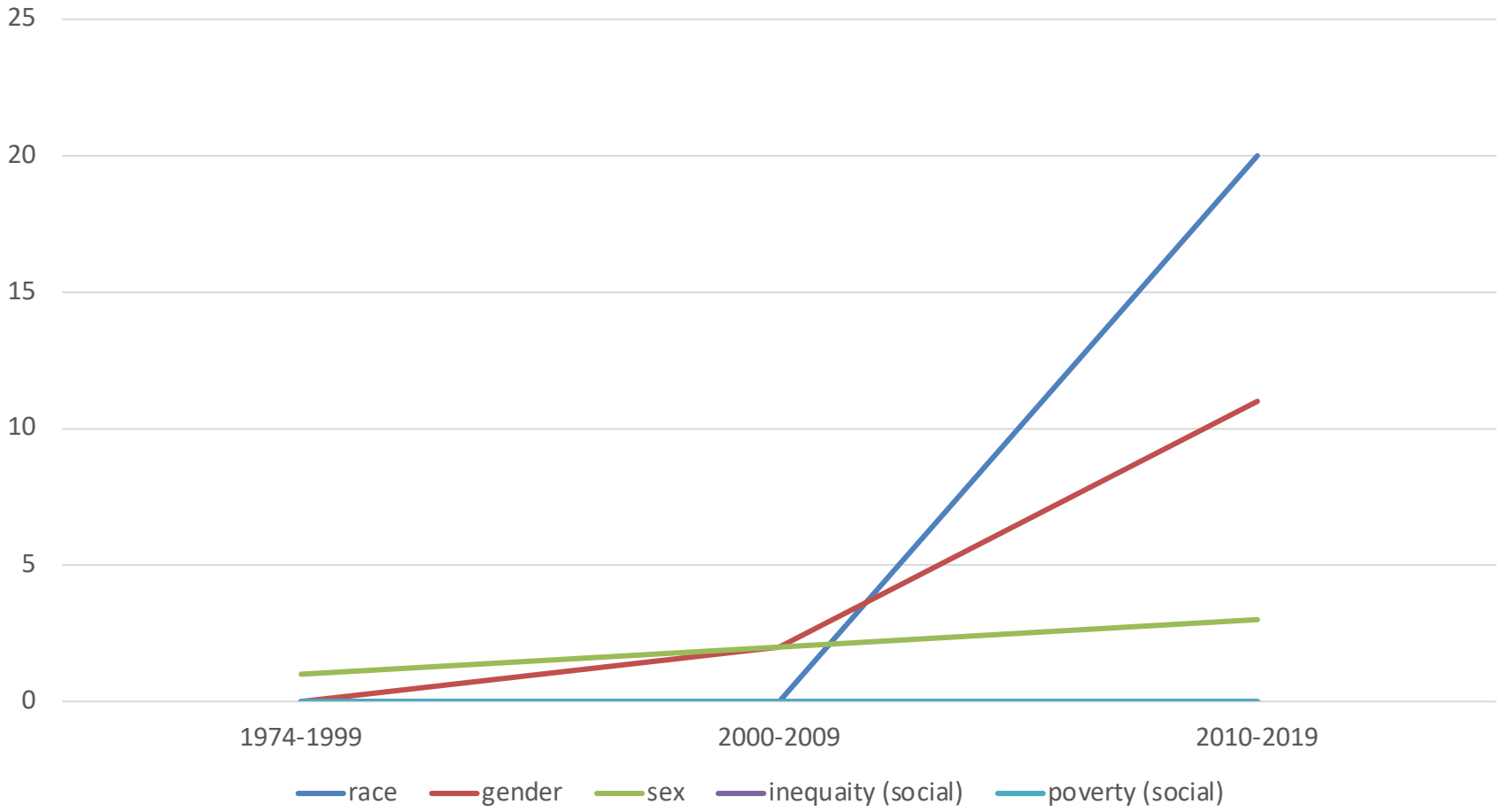
- Given the current distribution of resources and power in the world, and of the capacity to create and train the most powerful AI systems—largely within corporations and government security agencies—one can hardly be optimistic.
  - If SPP is to continue and flourish for another 50 years, human agents will need to learn massively as we interact with new classes of AI systems capable of increasing degrees of autonomous agency.
  - Especially, we will have to learn how to preserve scientific inquiry, democracy, autonomy, diversity, and fundamental human rights in the face of contemporary concentrations of resources and power.

Can SPP rise to the occasion?

	race	gender	sex	inequality (social)	poverty (social)
1974-1999 (incomplete)*	0	0	1	0	0
2000-2009	4	2	2	0	0
2010-2019	20	11	3	0	0
2020-2024 (incomplete)	4	10	1	1	0
Total 1974-2024 (incomplete)	28	23	7	1	0

*In an analysis that had access to programs for more of the years between 1974 and 1999, David Chalmers (2024, SPP Special Session) did not find *race, gender, sex, inequality*, or *poverty* among the top 17 topics.

race, gender, sex, inequality (social), poverty (social)



Track record

- The cases of *race*, *gender*, and *sex* give us some ground for optimism that SPP *can* rise to the occasion.
- And if not SPP … who?

Appendix 1:

From David Chalmer's SPP topics data (1985-1995)

- Occurrence of the term 'moral' in programs, reduced to session titles:

    - 1985 – 1*
    - 1986 – 0
    - 1987 – 0
    - 1988 – 0
    - 1990 – 0
    - 1991 – 0
    - 1992 – 0
    - 1993 – 1**
    - 1994 – 0
    - 1995 – 0

- * "Human Nature, Love, and Morality: The Possibility of Altruism"
- **Invited Symposium on Morality