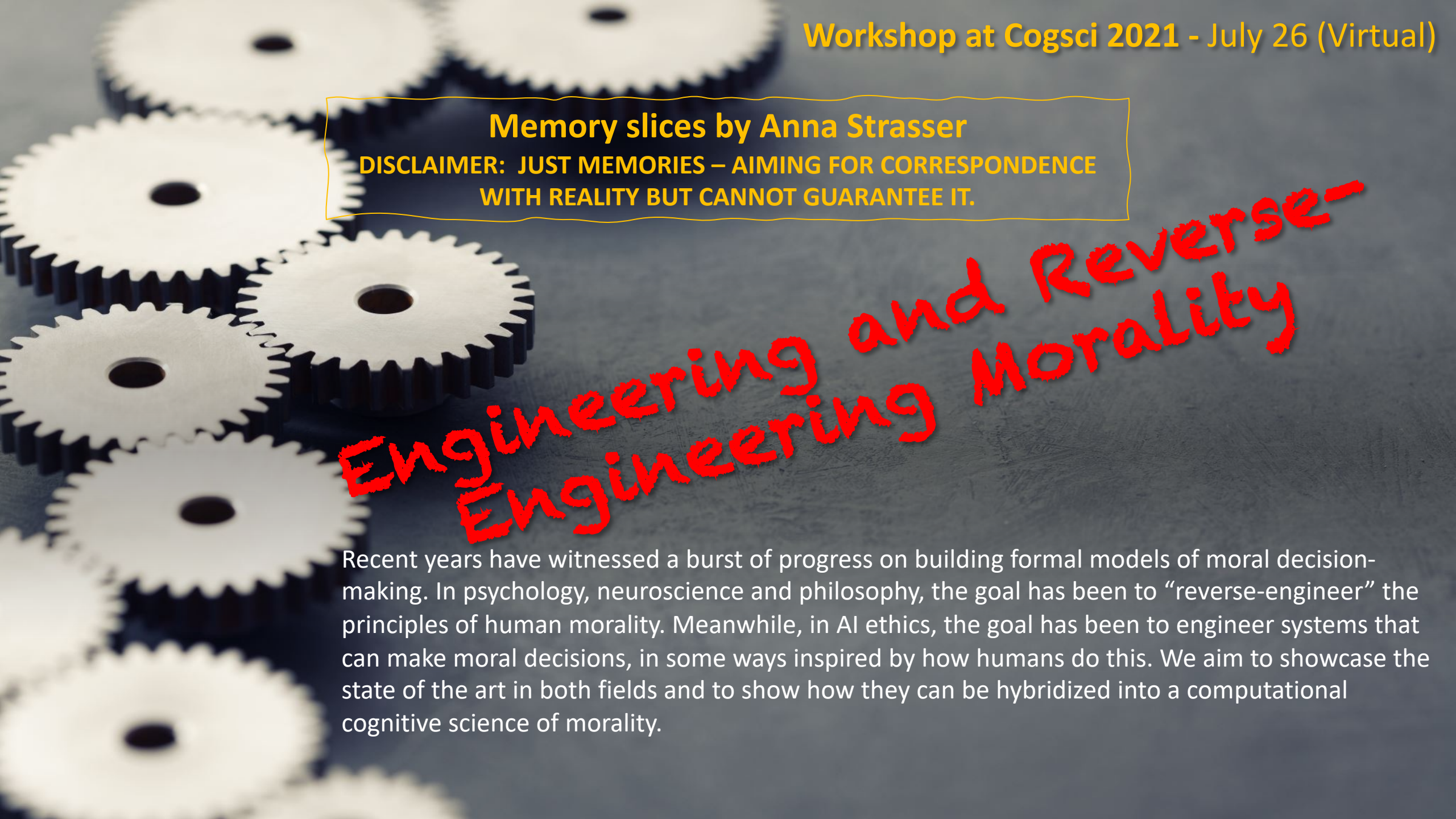


Workshop at Cogsci 2021 - July 26 (Virtual)

Memory slices by Anna Strasser

DISCLAIMER: JUST MEMORIES – AIMING FOR CORRESPONDENCE WITH REALITY BUT CANNOT GUARANTEE IT.

Engineering and Reverse-Engineering Morality

Recent years have witnessed a burst of progress on building formal models of moral decision-making. In psychology, neuroscience and philosophy, the goal has been to “reverse-engineer” the principles of human morality. Meanwhile, in AI ethics, the goal has been to engineer systems that can make moral decisions, in some ways inspired by how humans do this. We aim to showcase the state of the art in both fields and to show how they can be hybridized into a computational cognitive science of morality.

FIRST SESSION, “REVERSE-ENGINEERING THE MORALITY OF HUMANS”

- focus on the ways that human moral judgment can be studied using a reverse-engineering approach
- research using computational methods including computational cognitive modeling, rational analysis, and game theory

- Jean-Baptiste André & Nicolas Baumard
- Shaun Nichols
- Sydney Levine, Josh Tenenbaum, Fiery Cushman
- Gillian Hadfiled
- *group discussion of morning sessions*

Jean-Baptiste André & Nicolas Baumard

- moral computations considering comprehensive consequences

Input of the moral system: a game defined by
(i) a vector of players' opportunity costs, C
(ii) a set of feasible outcome, Ω

Output of the moral system: a course of action for each player, given by the **Nash bargaining solution** of the bargaining problem (Ω, C)

4 IMPLICATIONS

ECTM accounts for the

1. logic of merit
2. universalization principle
3. apparent variability of morality
4. intuitions in moral dilemmas


Jean-Baptiste André

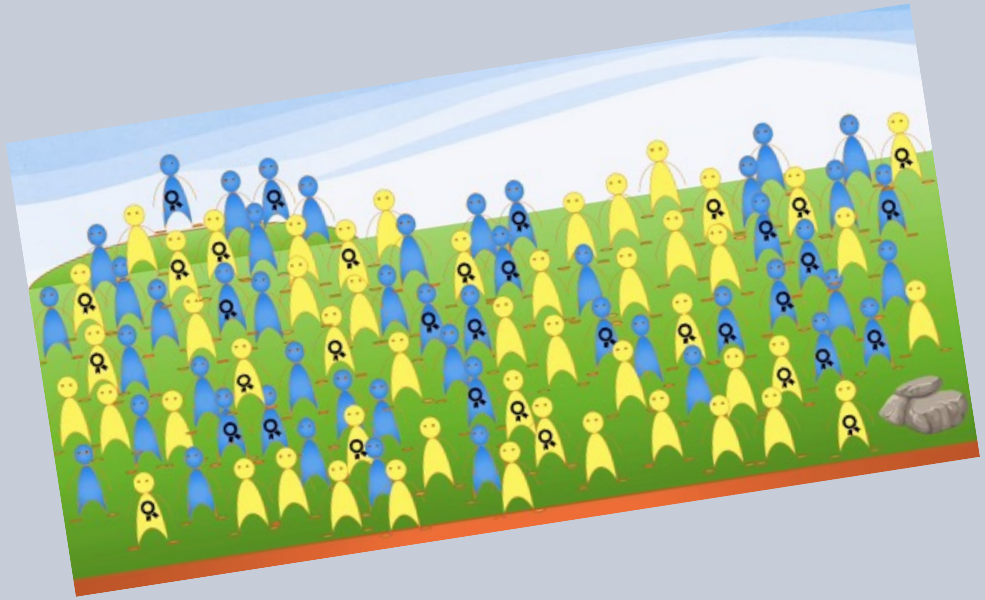

Nicolas Baumard

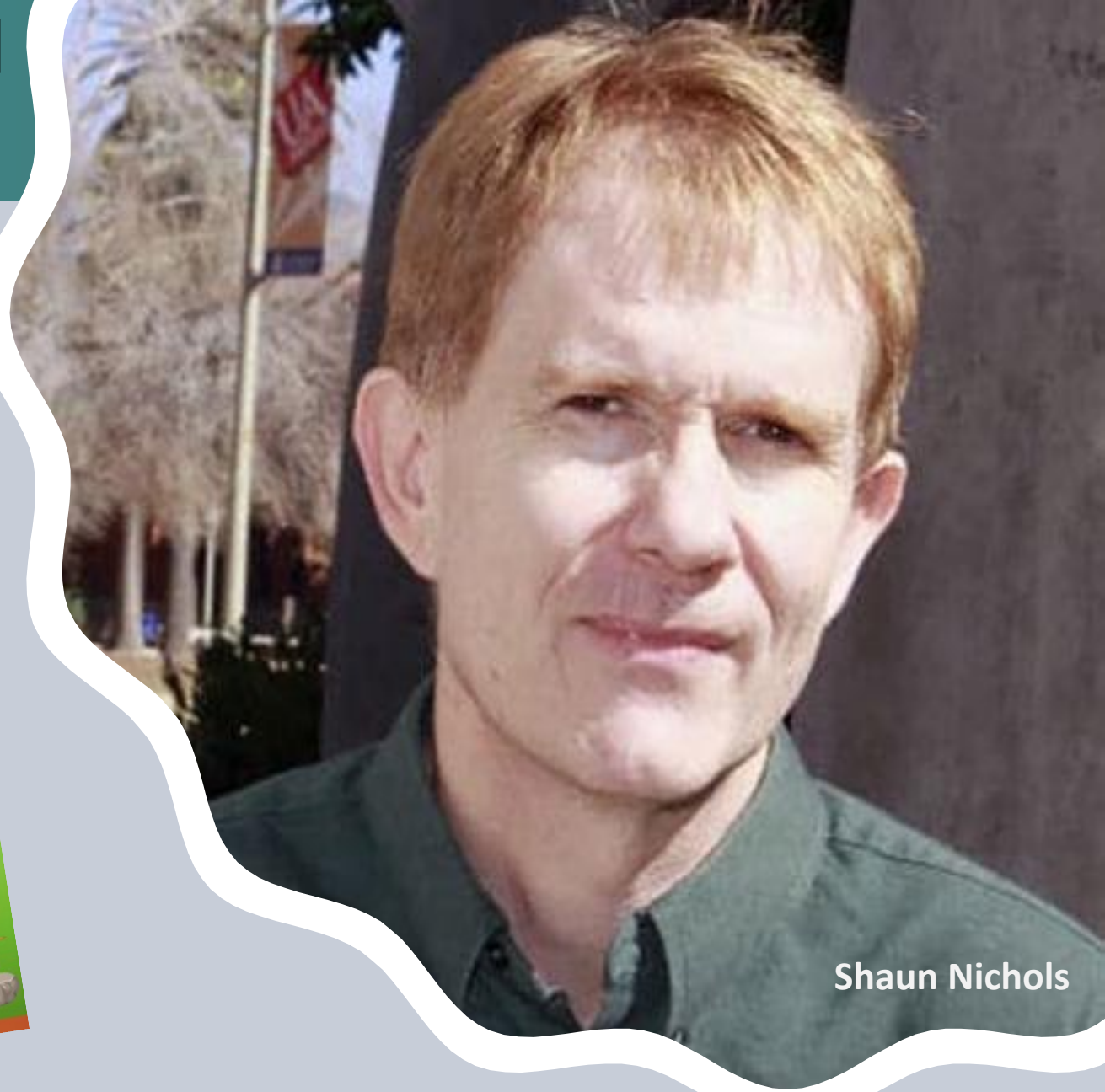
Reverse engineering parochial morality

tribalism

- explained by in-group / outgroup & automatic group bias hypothesis
- can also be explained by rational learning
  - decision between inclusive vs. parochial norms depends on the evidence presented

→ **reduce sampling errors**



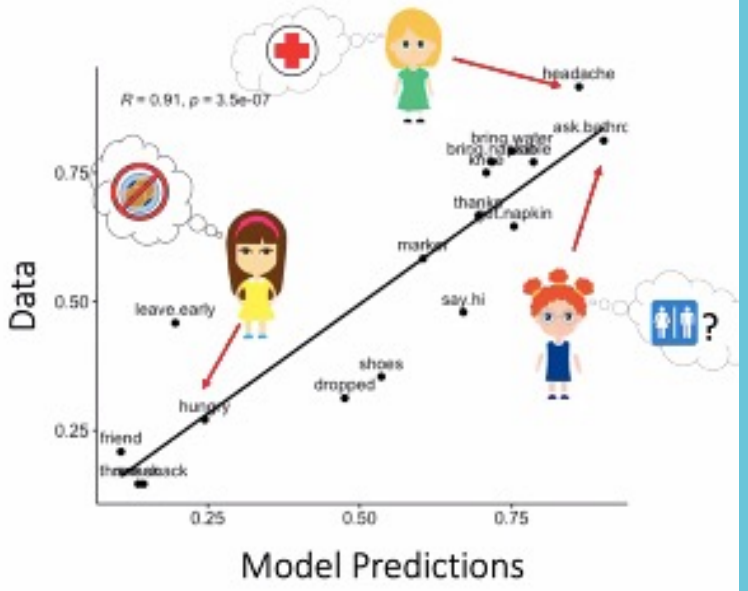


Shaun Nichols

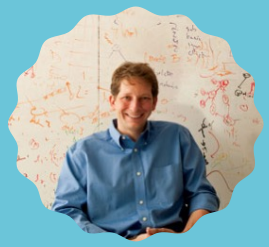
with Scott Partington and Tamar Kushnir

How does moral cognition work

- FLEXIBILITY !!!



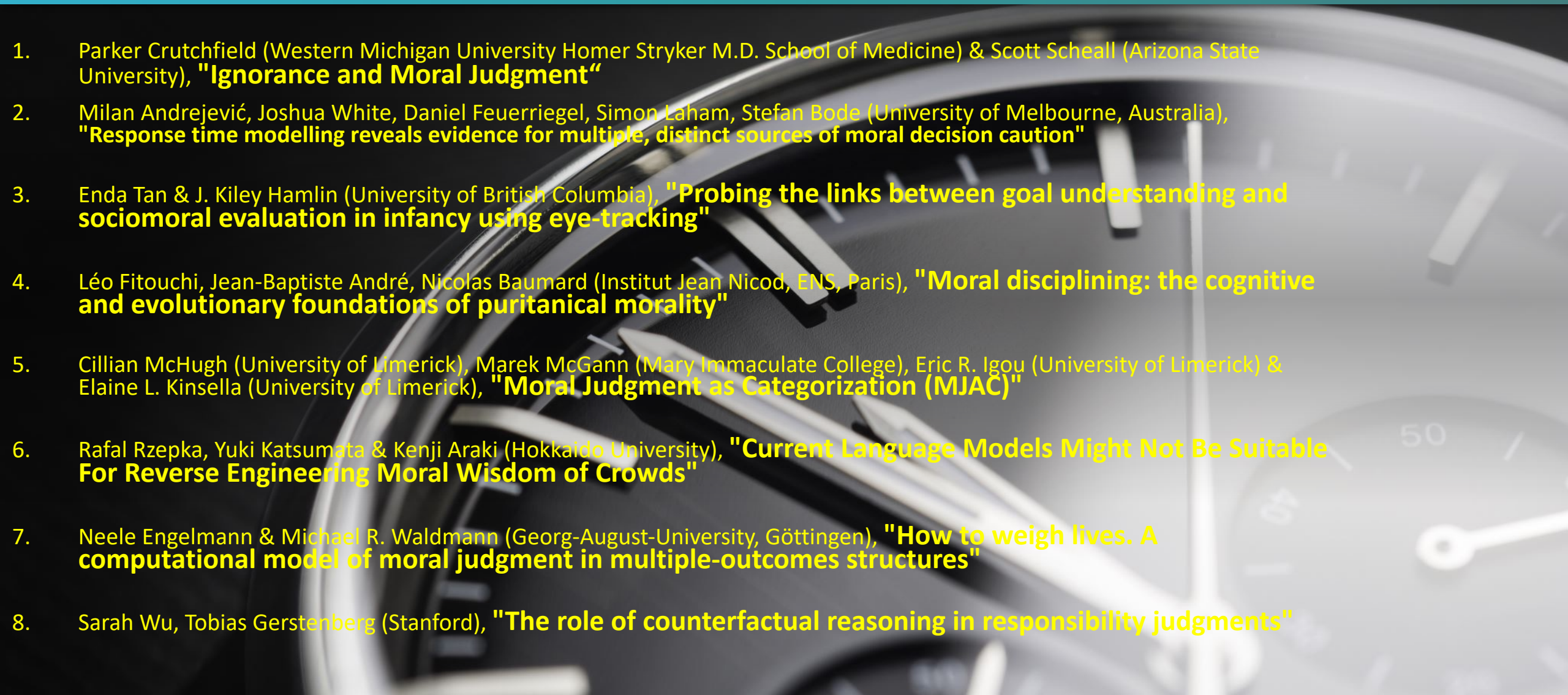
Sydney Levine



Josh Tenenbaum, Fiery Cushman

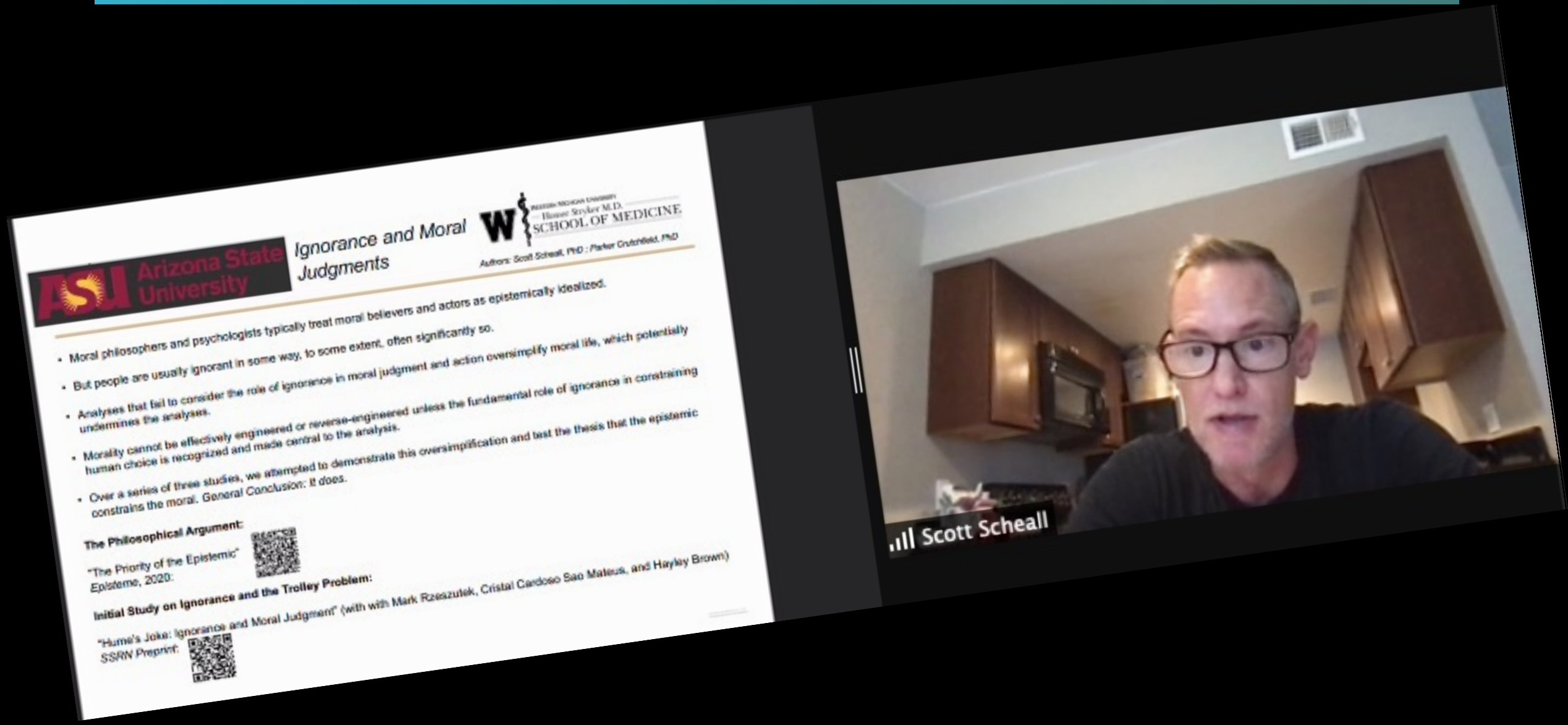
Takeaway: moral rules (+ their flexibility) is a resource-rational solution to the problem of inter-dependent rational choice.

90-second advertisements for posters



1. Parker Crutchfield (Western Michigan University Homer Stryker M.D. School of Medicine) & Scott Scheall (Arizona State University), **"Ignorance and Moral Judgment“**
2. Milan Andrejević, Joshua White, Daniel Feuerriegel, Simon Laham, Stefan Bode (University of Melbourne, Australia), **"Response time modelling reveals evidence for multiple, distinct sources of moral decision caution"**
3. Enda Tan & J. Kiley Hamlin (University of British Columbia), **"Probing the links between goal understanding and sociomoral evaluation in infancy using eye-tracking"**
4. Léo Fitouchi, Jean-Baptiste André, Nicolas Baumard (Institut Jean Nicod, ENS, Paris), **"Moral disciplining: the cognitive and evolutionary foundations of puritanical morality"**
5. Cillian McHugh (University of Limerick), Marek McGann (Mary Immaculate College), Eric R. Igou (University of Limerick) & Elaine L. Kinsella (University of Limerick), **"Moral Judgment as Categorization (MJAC)"**
6. Rafal Rzepka, Yuki Katsumata & Kenji Araki (Hokkaido University), **"Current Language Models Might Not Be Suitable For Reverse Engineering Moral Wisdom of Crowds"**
7. Neele Engelmann & Michael R. Waldmann (Georg-August-University, Göttingen), **"How to weigh lives. A computational model of moral judgment in multiple-outcomes structures"**
8. Sarah Wu, Tobias Gerstenberg (Stanford), **"The role of counterfactual reasoning in responsibility judgments"**

1. Ignorance and Moral Judgment



3. Probing the links between goal understanding and sociomoral evaluation in infancy using eye-tracking

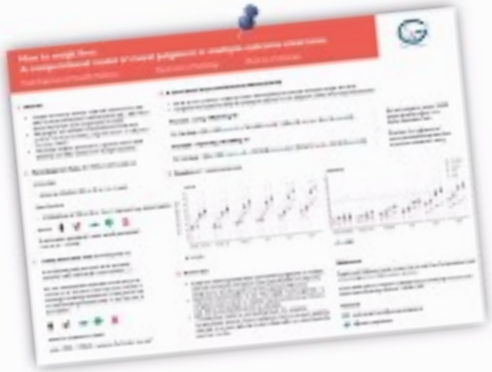
https://cic.psych.ubc.ca/research-publications/publications/

7. How to weigh lives. A computational model of moral judgment in multiple-outcomes structures

How to weigh lives. A computational model of moral judgment in multiple-outcome structures.

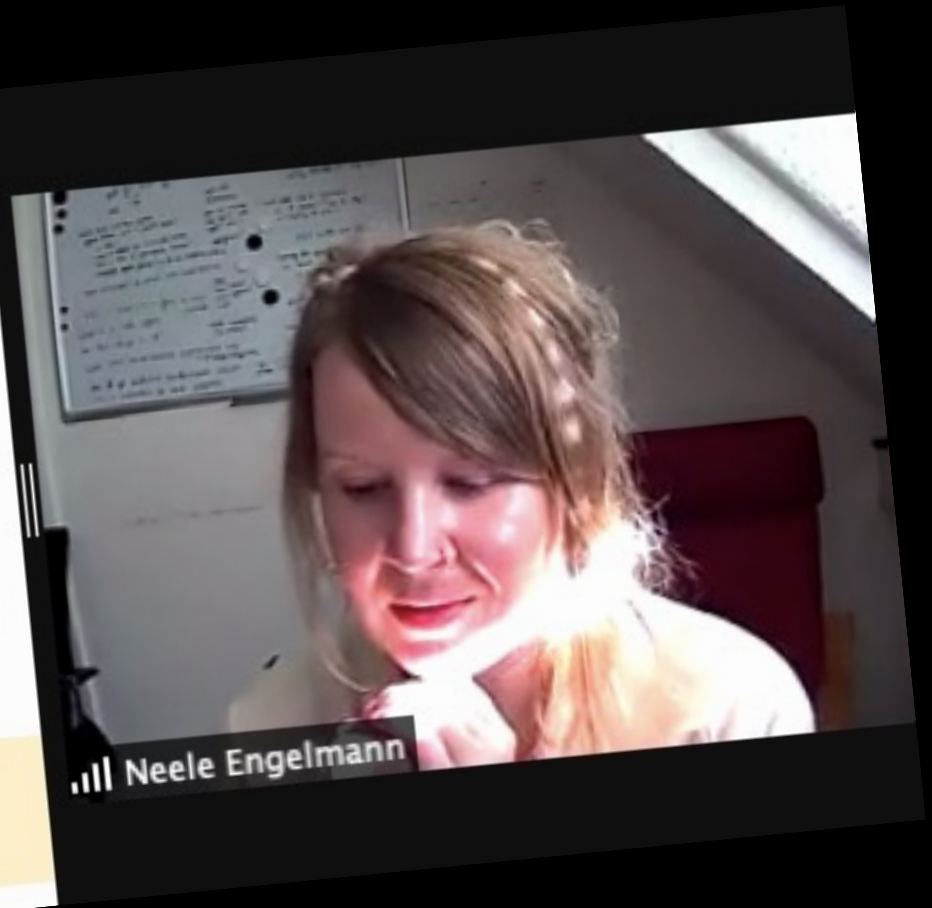
Neele Engelmann & Michael R. Waldmann, Department of Psychology, University of Göttingen

- People readily make moral judgments about all kinds of outcome trade-offs in moral dilemmas (see e.g., Awad et al., 2018)
- All overarching theoretical frameworks of moral psychology acknowledge that reasoning about consequences is one important component of arriving at a moral judgment about an action (Greene, 2001, 2004; Cushman, 2013; Crockett, 2013; Mikhail, 2007, 2011)



But how do we decide whether the consequences of acting outweigh the consequences of not acting in a moral scenario?

- We propose and evaluate a computational model that predicts moral permissibility judgments based on people's subjective utilities of consequences|action vs. consequences|inaction
- The model predicted people's moral judgments well in both dilemmas and other actions with multiple outcomes
- Our model could serve as one building block of a complete computational account of human moral judgment – the part that deals with evaluating consequences
- Such a complete computational account would be desirable in its own right, but also indispensable as a benchmark for designing machine morality – which aspects should machines imitate and which aspects should be altered!



second session

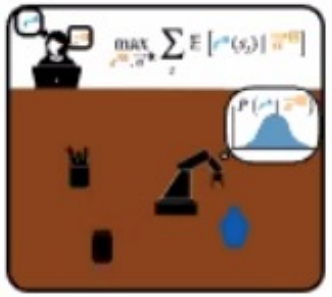
"Learning from Humans to Build Moral AI"

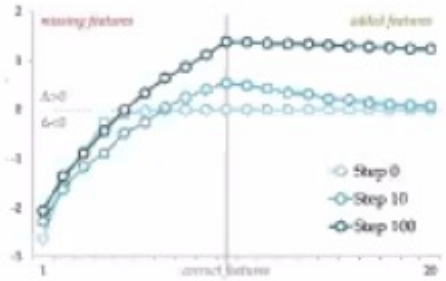
- showcase a series of proposals for building ethical AI that draw insights from cognitive science
- how human cognition navigates the complex moral world as a starting place to generate engineering solutions to similar problems

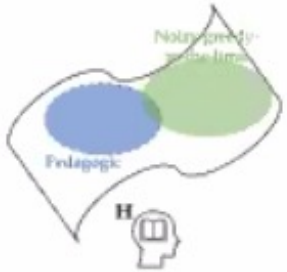
- Dylan Hadfield-Menell & Stuart Russell
- Julia Haas
- Alison Gopnik
- Peter Railton
- Henry Shevlin
- group discussion of afternoon talks
- Sholei Croom facilitates general discussion with questions for participants
- additional general discussion on any topic related to the workshop

Dylan Hadfield-Menell & Stuart Russell

Takeaways







Uncertainty and cooperation are <u>crucial</u> modeling components for learning normative properties of the world from people

- Too many features slows down learning
- Too few leads to persistent misalignment

- Modeling (moral) learning matters
- Imitation equilibria are fairly robust in the limit
- Pedagogic equilibria are efficient, but brittle

Artificial moral cognition

HUMANS

- moral cognition involves
  - reason (May (2018) / emotion (Prinz 2016) / hybrid (Mallon & Nichols 2011)

→ MORAL VALUATIONISM

valuation guides cognitive selection between competing moral states of affairs

e.g.: agent attributes **subjective reward**

- to the act or outcome of buying fair trade coffee OR
- to the determinants of the choice commodity OR
- to the idea of fair-trade practices themselves

AI

aim: a model covering a large portion of human cognition (Allen & Wallach 2012)

Julia Haas

The biology, intelligence and computational basis of care

the social contract oriented approaches tend to ignore care as an important factor

one should consider tradeoffs

- exploitation (utility)
- exploration
- **care / teaching**

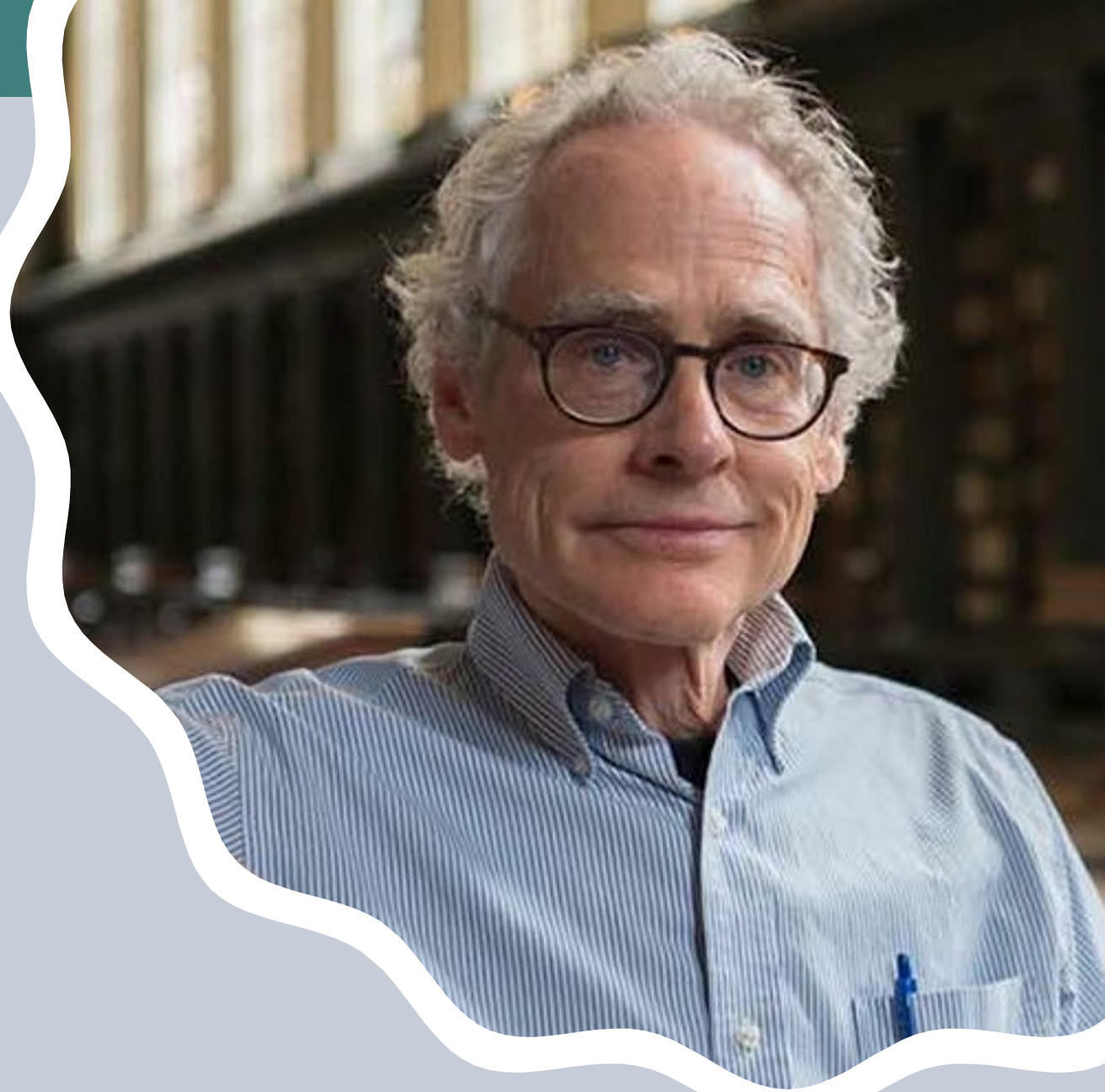
→ **explain the foundational fact of caregiving altruism**

Alison Gopnik

21:15 – Peter Railton

Ethical purport

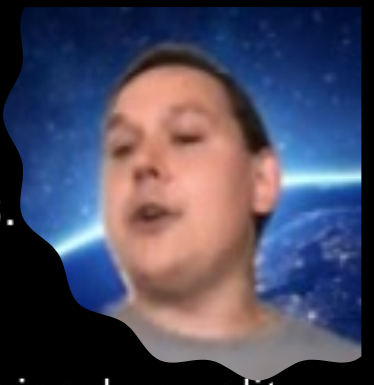
- Ethical assessments purport to meet criteria of:
  - (a) impartiality
  - (b) generality
  - (c) objectivity and rationality
  - (d) accuracy in representing agents, actions, and outcomes
  - (e) motivation
  - (f) independence from arbitrary authority or sanction,
  - (g) intrinsic weight is given to harms and benefits to others as well as the self
- Moral judgments can be criticized if they can be shown not to meet such criteria.



Peter Railton

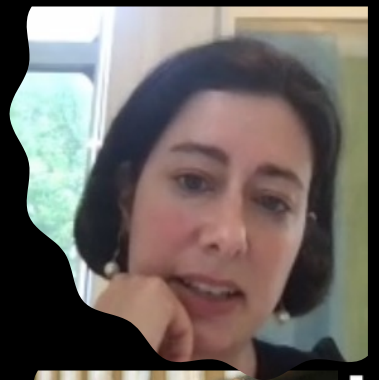
summarizing the 4 talks

Henry Shevlin,



Dylan: showed how cooperation and uncertainty help optimize outcomes.

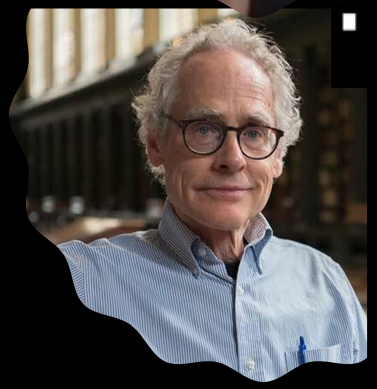
Essential if AI is going to do more than feed back our superficial desires.



Julia: A foundation to enable AI to learn contextual and fine-grained morality.

Critical for AI to go beyond laws and principles that underdetermine morality.

Alison: Brought out the centrality that love and relationships of care and dependency have in human affairs.



- Suggests ways in which these could help create a foundation for AI morality.

Peter: A compelling account of interrelations between communication, social cognition, language and morality.

Provides the groundwork for rewarding and ethical human-AI interactions.

10 questions

(1) Alison & Peter: the 'folie à deux' problem for AI relationships.

(2) Alison & Peter: Reciprocity and dependence

Relationships are a two-way street, but does this make sense for AI?

3. **Peter & Julia: Persistence of 'hot' moral disagreement.**

Should we expect 'moral pillarization' in AI?

(4) Julia & Dylan: risk of decline of human moral decision-making?

Two concerns: both moral de-skilling and intrinsic value worry.

(5) Julia & Dylan: what makes valuation/reward function specifically moral?

How could an AI grasp the difference between aesthetic & moral norms?

(6) Julia & Peter: would moral AI truly be acting for moral reasons?

(7) Dylan: how far should automated 'preference' detection go?

A good friend or therapist might help us discover surprising or socially complex preferences (e.g., someone unsure about sexual or gender identity).

(8) Dylan: how to learn preferences across 'transformational' boundaries?

Should an algorithm have the option to nudge someone towards Mormonism? How about meditation?

(9) All panelists: what's the relationship between being a moral decision-maker and being a member of a moral community?

(10) All panelists: What would moral innovation mean for an AI? How could we distinguish this from malfunction?

thanks a lot for organizing this inspiring workshop

Organizers

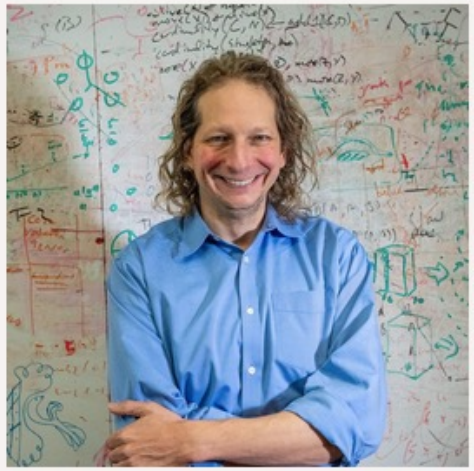
Sydney Levine

Harvard, Psychology

MIT, Brain and Cognitive Sciences

Fiery Cushman

Harvard, Psychology



Joshua Tenenbaum

MIT, Brain and Cognitive Sciences

Iyad Rahwan

Max Planck Institute for Human Development