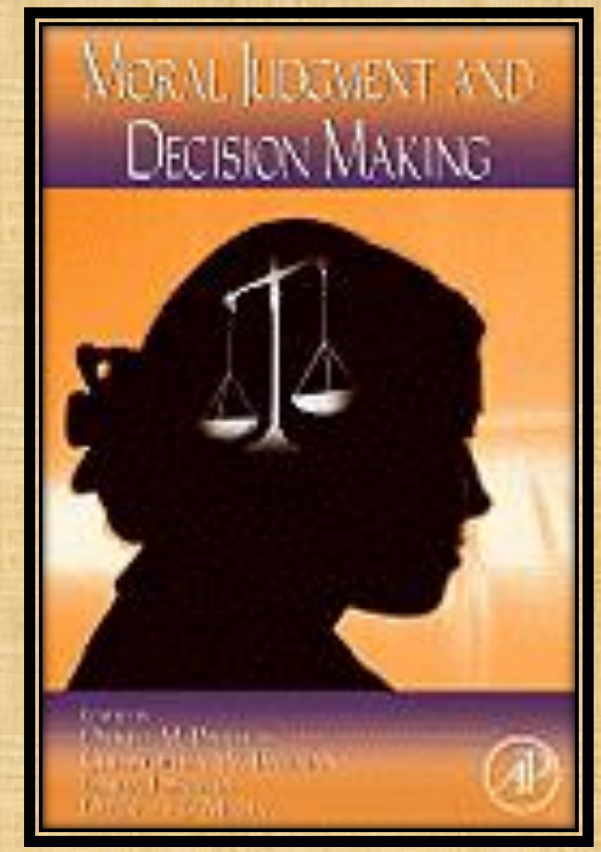
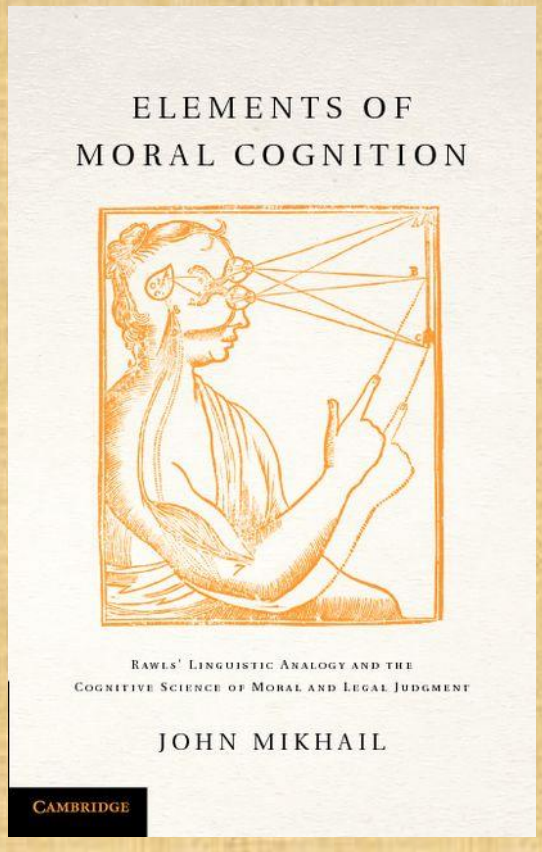
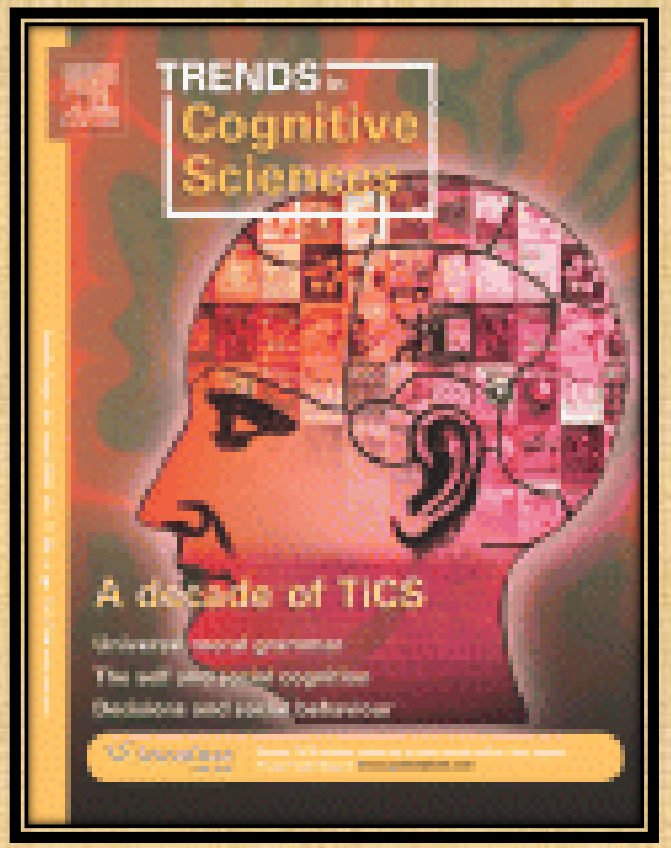
Machine Learning and Human Cognition: Some Parallels

University of Pennsylvania Law School
February 20, 2017



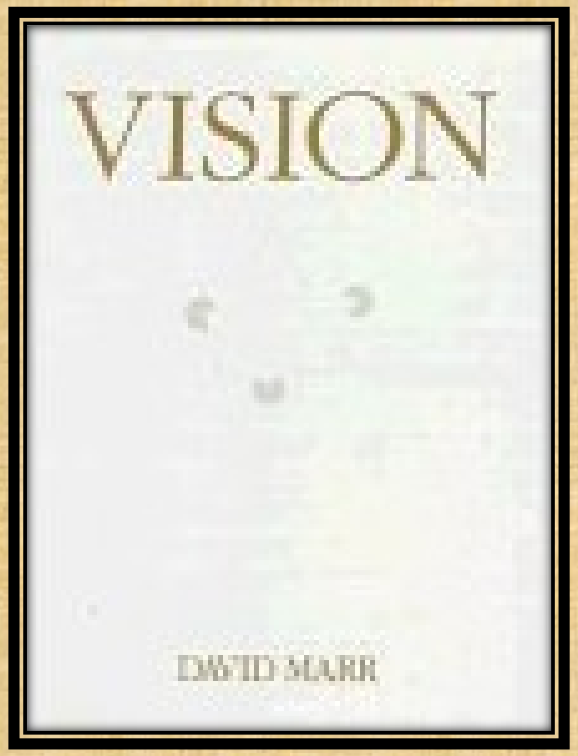




"Enlightenment Rationalism"

- Methodological naturalism
- Moral judgment:
  - Computational
  - Unconscious
  - Non-introspectible
  - Innate
  - Species property
- Domain-specific Darwinian algorithms

Importance of Computational Theory



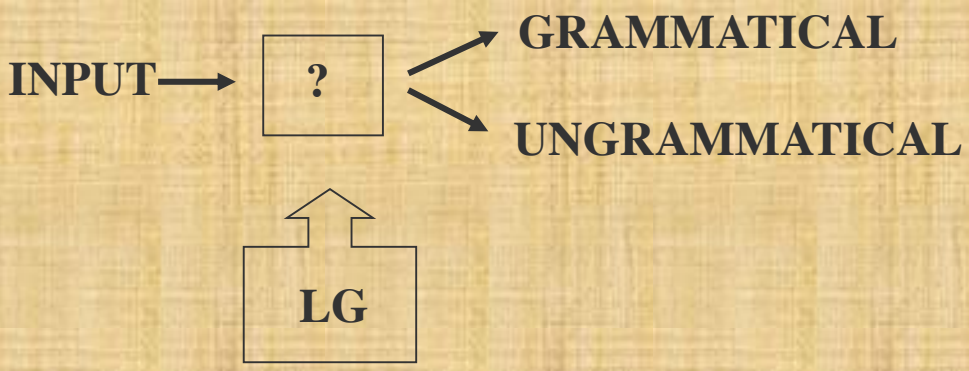
"Although algorithms and mechanisms are empirically more accessible, it is the top level, the level of computational theory, that is critically important from an information-processing point of view…one has to exercise extreme caution in making inferences from neurophysiological findings about the algorithms and representations being used, particularly until one has a clear idea about what information needs to be represented and what processes need to be implemented."

Marr's Three Levels

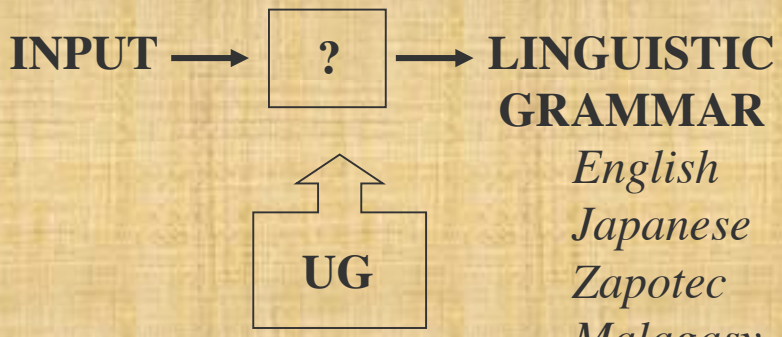
- *Computational theory*
- *Representation & algorithm*
- *Hardware implementation*

Language

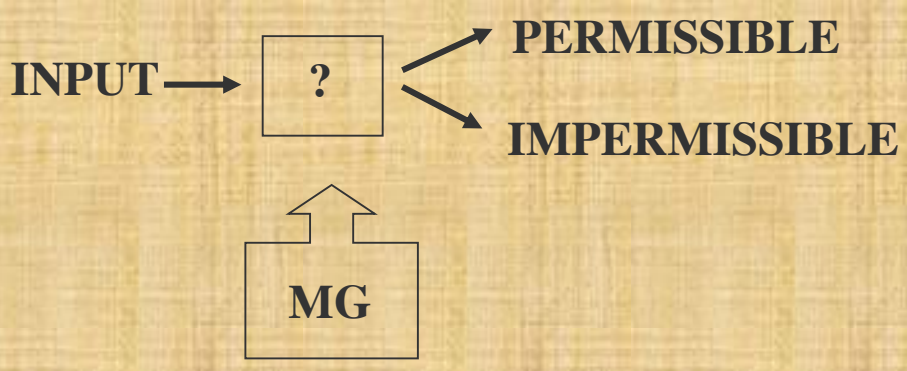
Perception

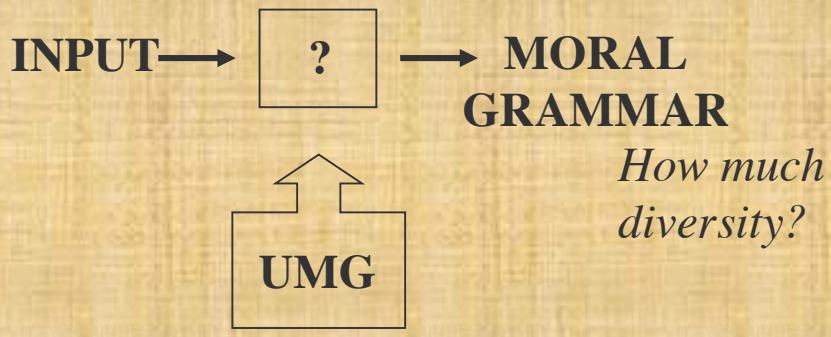


Acquisition

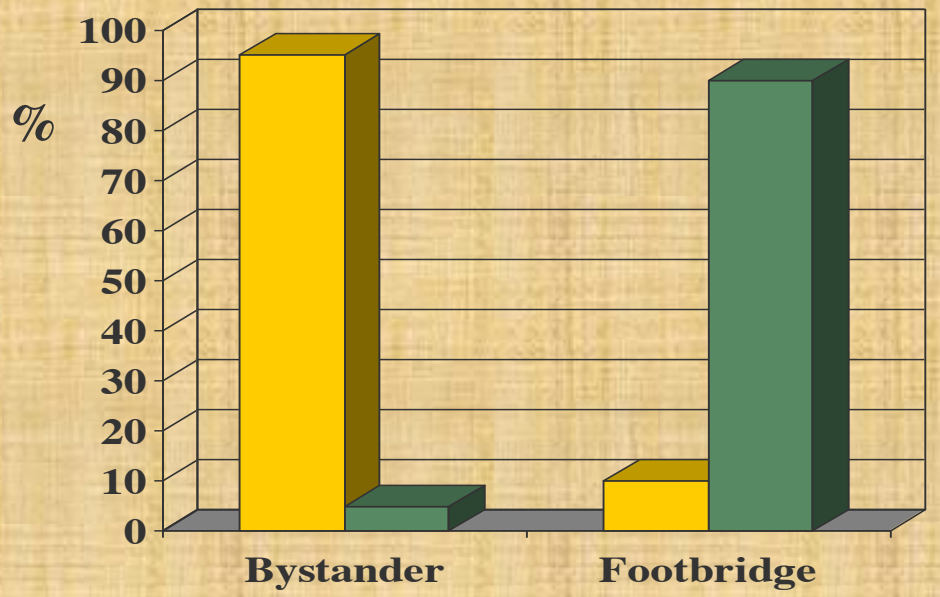


Morality



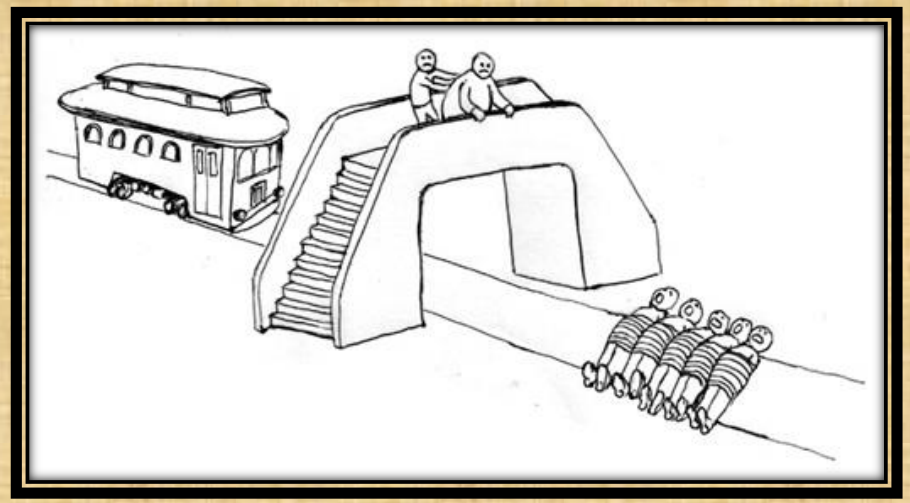


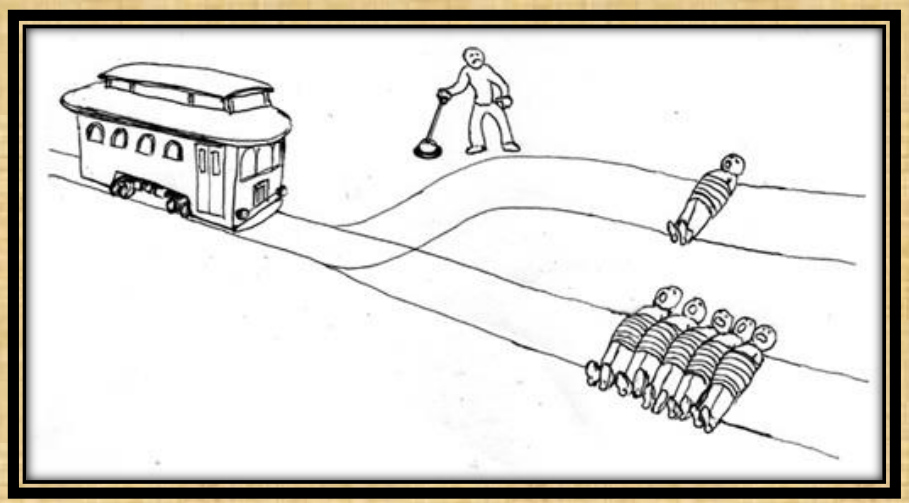
Trolley Problems



X^2 (1, N=40) = 29.0, $p < .001$

Mikhail, Sorrentino & Spelke 1998; Mikhail 2000, 2002



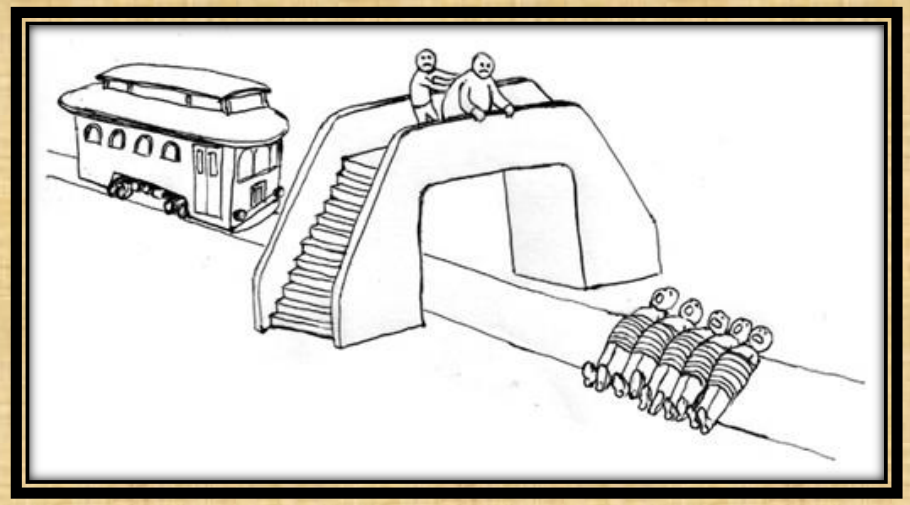


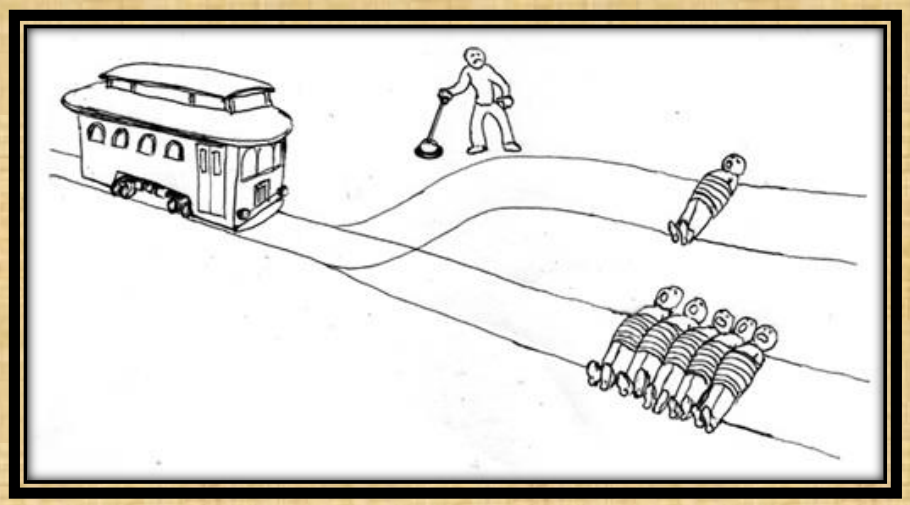
Trolley Problems

"Very odd. I don't know why I chose differently in the second scenario. The end result is the same. I just chose my gut response - and now am intrigued with how to reconcile them."

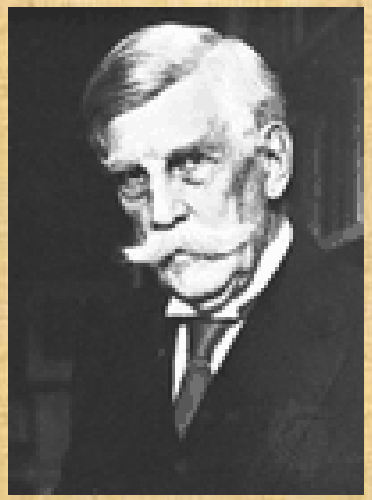
"My reaction is intuitive and I realize not logically justifiable."

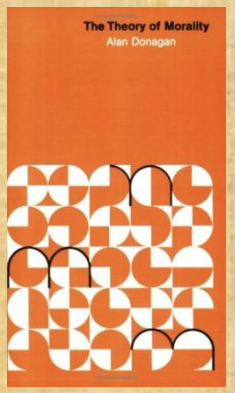
"It's amazing that I would not throw a person but would throw a switch to kill a person."

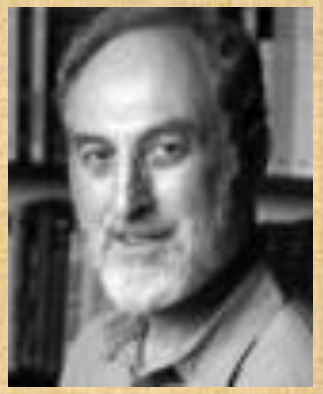




Unconscious representations







homicide at $t^{(+n+p)}$

Side Effects

preventing train from killing men at $t^{(+n+o)}$

End

battery at $t^{(+n)}$

causing train to hit man at $t^{(+n)}$

battery at $t^{(0)}$

throwing man at $t^{(0)}$

battery at $t^{(-m)}$

touching man at $t^{(-m)}$

Means

Footbridge

homicide at $t^{(+n+o+p)}$

battery at $t^{(+n+o)}$

causing train to hit man at $t^{(+n+o)}$

Side Effects

preventing train from killing men at $t^{(+n)}$

End

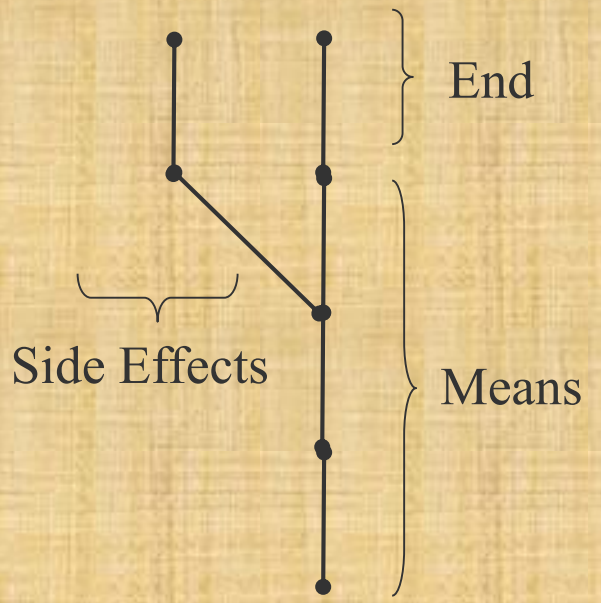
turning train at $t^{(+n)}$

throwing switch at $t^{(0)}$

Means

Bystander

Common legal analysis



- Laws of war
- Torture
- Equal Protection
- Establishment Clause
- Physician-Assisted Suicide
- Incitement
- Hearsay
- Taxation (charitable gift)
- Necessity
- Negligence