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

**Input** → ? → LG

Acquisition

**Input** → ? → UG → LINGUISTIC GRAMMAR
  - English
  - Japanese
  - Zapotec
  - Malagasy
  - Arabic
  - ......

**Morality**

**Input** → ? → MG

**Input** → ? → PERMISSIBLE

**Input** → ? → IMPERMISSIBLE

**Input** → ? → MORTAL GRAMMATICAL

How much diversity?
Trolley Problems

Mikhail, Sorrentino & Spelke 1998; Mikhail 2000, 2002
“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

Side Effects

- Footbridge
  - Preventing train from killing men at $t^{(+n+o)}$
  - 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

- Battery at $t^{(+n+o)}$
- Preventing train from killing men at $t^{(+n+o)}$
- Turning train at $t^{(+n)}$
  - Throwing switch at $t^{(0)}$

End

Unconscious representations
Common legal analysis

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