Artificial Unintelligence

Peter Dayan
Gatsby Computational Neuroscience Unit

Nathaniel Daw Ray Dolan Quentin Huys
Read Montague

Intelligent Behaviour

- optimal choices in a nasty, brutish, otherwise short world?
- but:
  - spirit weak; flesh willing:
    - optimally solving the wrong problem
  - spirit willing; flesh weak:
    - individual differences in algorithmic choices (MB vs MF)
    - heuristics adapted to evolutionary circumstances
  - spirit weak; flesh weak:
    - BREXIT

(Approximate) Bayesian Decision Theory

problem definition (e.g., a maze) occurrent state (e.g., location)

- states $\mathbf{x} = \{\mathbf{x}_p, \mathbf{x}_s\}$
  - ignorant: have to represent ... over
    - combine priors & likeli...
- actions
  - catastrophic calc... is when temporally extended
    - short-cut... habitization; evolutionary programming
- utilities
  - state-... dent; other-regarding; time-dependent; ho... atically adapting
- ch... noisy argmaximization

'bad' decisions can come from any one of these – with non-identical signatures

argmax of expected utility over state distribution

Bad/In-Decision Theory

- wrong problem
  - priors; likelihoods; affordances; utilities
- right problem, wrong inference
  - fallacious precisions (expected/unexpected uncertainty)
  - early/over-habitization;
  - over-reliance on Pavlovian mechanisms
- right problem, right inference, wrong environment
  - learned helplessness
  - discounting from inconsistency

not completely independent: e.g., miscalibration

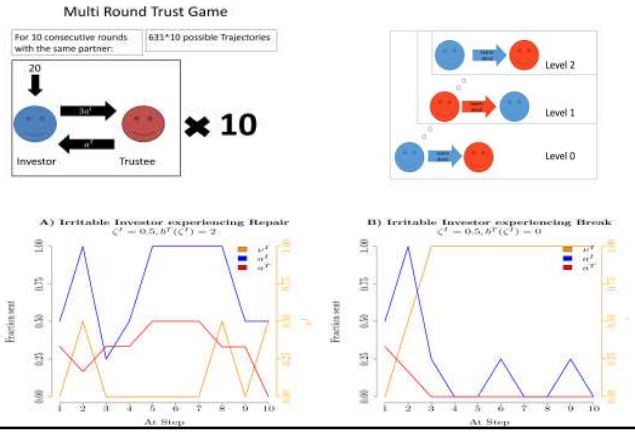
Wrong Problem

- prior:
  - overly strong prior beliefs
    - ruling in or out single possibilities (delusions)
    - forcing malign interpretations for ambiguous data
  - generalization
    - inference about contexts
      - over-generalization & depression
      - under-generalization & extinction (depression; PTSD)
    - representation: $\mathcal{P}_{x_p}(x_s)$ determines coords for x_s
      - controls effect of learning

Wrong Problem

- likelihood
  - incorrect explanations for data
    - "tubes only arrive on time for important people" versus "I am undeniably important"
  - neglect ones own contribution to the data (Frith; Friston; Stephan; Wolpert)
    - efference copy in schizophrenia
  - mis-interpret cues from others (Kings-Casas; Fonagy; Lohrenz; Montague)
    - borderline personality disorder

Irritation

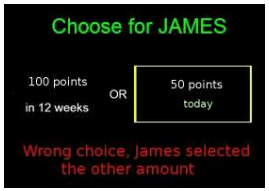


Wrong Problem

- utility function:
  - *de gustibus non est disputandum*
  - anhedonia
  - discount factor for future outcomes
    - impulsivity
    - hyperbolic discounting and intertemporal choice conflict (Ainslie)

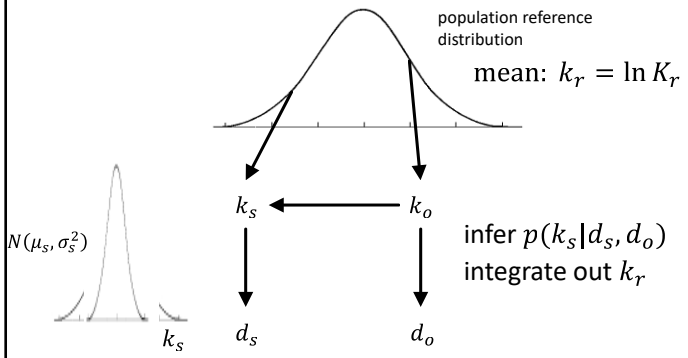
Learn to Choose for Another

Garvert et al (2015)



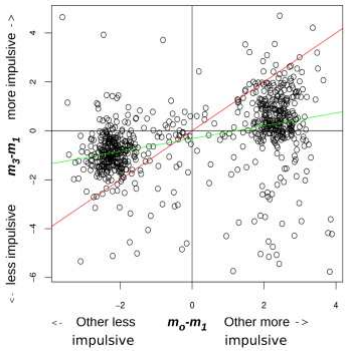
- chose $\ln K_{\text{other}}$ shifted by $\hat{\sigma} = 2.3$ towards (p=2/3) or away (p=1/3) from $\hat{\mu} = -4.5$; with T=1

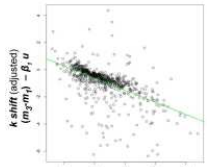
Uncertainty + Similarity ⇒ Influence



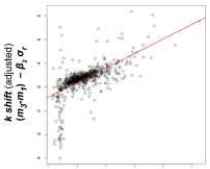
D0DD000D

Become Like Other





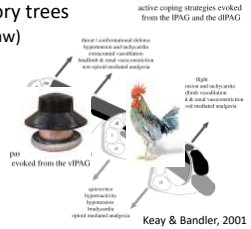
reference uncertainty



self uncertainty

Wrong Inference

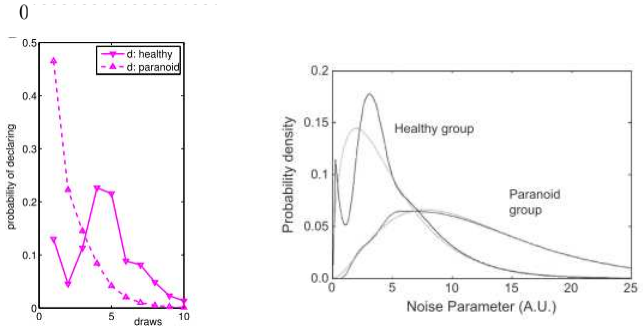
- inference about states:
  - computationally catastrophic; approximations can be dangerous
  - balance different information sources – 'precision'
- inference about actions: trajectory trees
  - multiple instrumental systems (Daw)
    - parameters for pathologies
  - Pavlovian choices
  - other hard-wiring
    - stress and model-based reasoning
    - pruning
- decision noise (beads)



Keay & Bandler, 2001

Wrong Inference

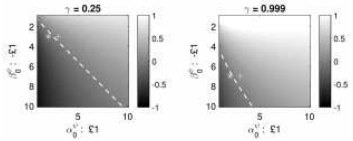
- decision noise in the beads task



Moutoussis (2011) model of Corcoran (2008); n=36 subjects < 65 with persecutory delusions

Wrong Environment

- inference about x_p
- learned helplessness
  - homogeneous subjects; heterogeneous outcomes
  - rationalisable generalization
- `early life events' bake in expectations; heuristics; problems



versus sure £0

- dependence on inferences about x_s
  - sloth of pharmacological effects if insufficient statistics

Bad/In-Decision Theory

- wrong problem
- right problem, wrong inference
- right problem, right inference, wrong environment