Trust in Autonomy

(A Guide for the Perplexed)

Chief Perplexity Officer: Bill Casebeer, wcasebeer@riversideresearch.org
(Director, Artificial Intelligence & Machine Learning at Riverside Research)

1

Overview

- Definitions *cum* theories of trust—prefer the behavioral economics approach
- Trust—theoretically simply, cognitively complex
- Determinants of trust in autonomy: empirical example of workload and reliance (over and under)
- Broadening the aperture: heuristics and biases
- A rich research and development agenda, with a special plea for the development of an artificial conscience

2

Theories of Trust

- Definition choice is really a matter of theory choice
- Choose theories that are: as broad as possible, simple, coherent, operationalizable, explanatory, empirically justified
- A good example: a dispositional theory of trust…trust occurs when agent A is willing to expose itself to risk by relying on agent B
  - Other uses are metaphorical
- Some advantages: relatively straightforward, can be measured, can be a subject of engineering…calibrated trust
  - Straight lift from the behavioral economics literature
- Epicycles: traits of agents, effect of environment and context

3

Trust: Simple, and Yet Cognitively Complex

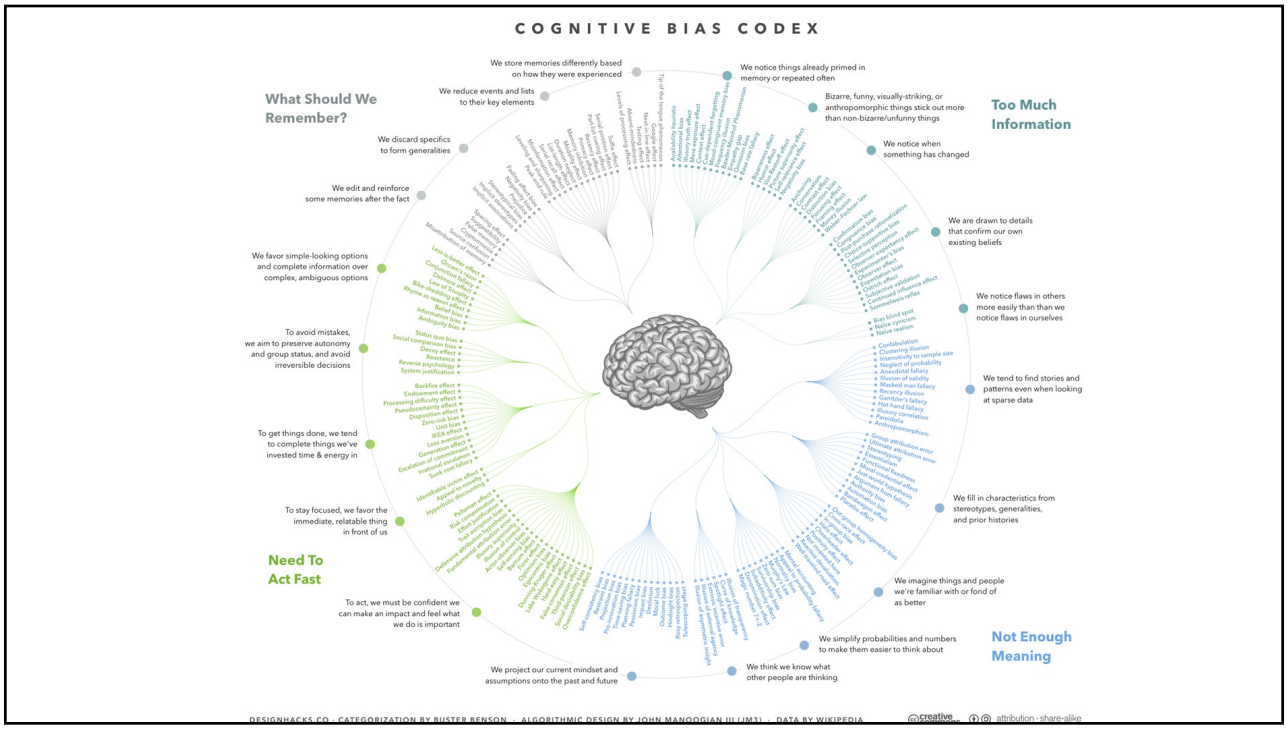
- This theory of trust is relatively straightforward (cf., Mayer et al)
- But: though our concepts of trust may be simple, their implementation in our cognitive architecture is very complex
  - Multiple intervening non-intuitive variables, multiple external influences
- Determinants of trust—some are intuitive, some are not
  - Intuitive: past experience
  - Not intuitive: workload
  - Entirely in the space of *causes* not reasons: brute neurobiology (e.g., temperature and inverse reasoning about parasympathetic nervous system activation, or oxytocin release and the stroking of fur)

4

An Example: Workload and Reliance on Autonomy

- Experimental setup: automated target recognition (ATR) algorithm delivers target class judgment along with confidence assessment, pilots are asked to give permission to unmanned aerial vehicle teammate to prosecute target (rules of engagement), vary the amount and complexity of other tasks pilot must also tackle
- Basic findings: in situations of high workload, pilots will over-rely on ATR algorithm even when confidence assessment is low; conversely, pilots are prone to second guess algorithm in low workload conditions even when confidence assessment is high (under-reliance)
- Theoretical explanation: Workload management as a heuristic to help us deal with what is important when we need to act quickly

5



6

An Example from the Moral Domain

- Calibrating trust across the moral domain will almost certainly require algorithms which can assess the environment of action and the cognitive and contextual influences on judgment, and provide a nudge or corrective to produce better judgment
- Example: prospect theory
- Call this what is is: an artificial conscience, or an artificial moral judgment/decision-making/action aide
- We should engineer these
- They have already saved lives: Auto-Ground Collision Avoidance System (GCAS) use

7

Review—Thank You!

- Definitions *cum* theories of trust—prefer the behavioral economics approach
- Trust—theoretically simply, cognitively complex
- Determinants of trust in autonomy: empirical example of workload and reliance (over and under)
- Broadening the aperture: heuristics and biases
- A rich research and development agenda, with a special plea for the development of an artificial conscience

wcasebeer@riversideresearch.org

8