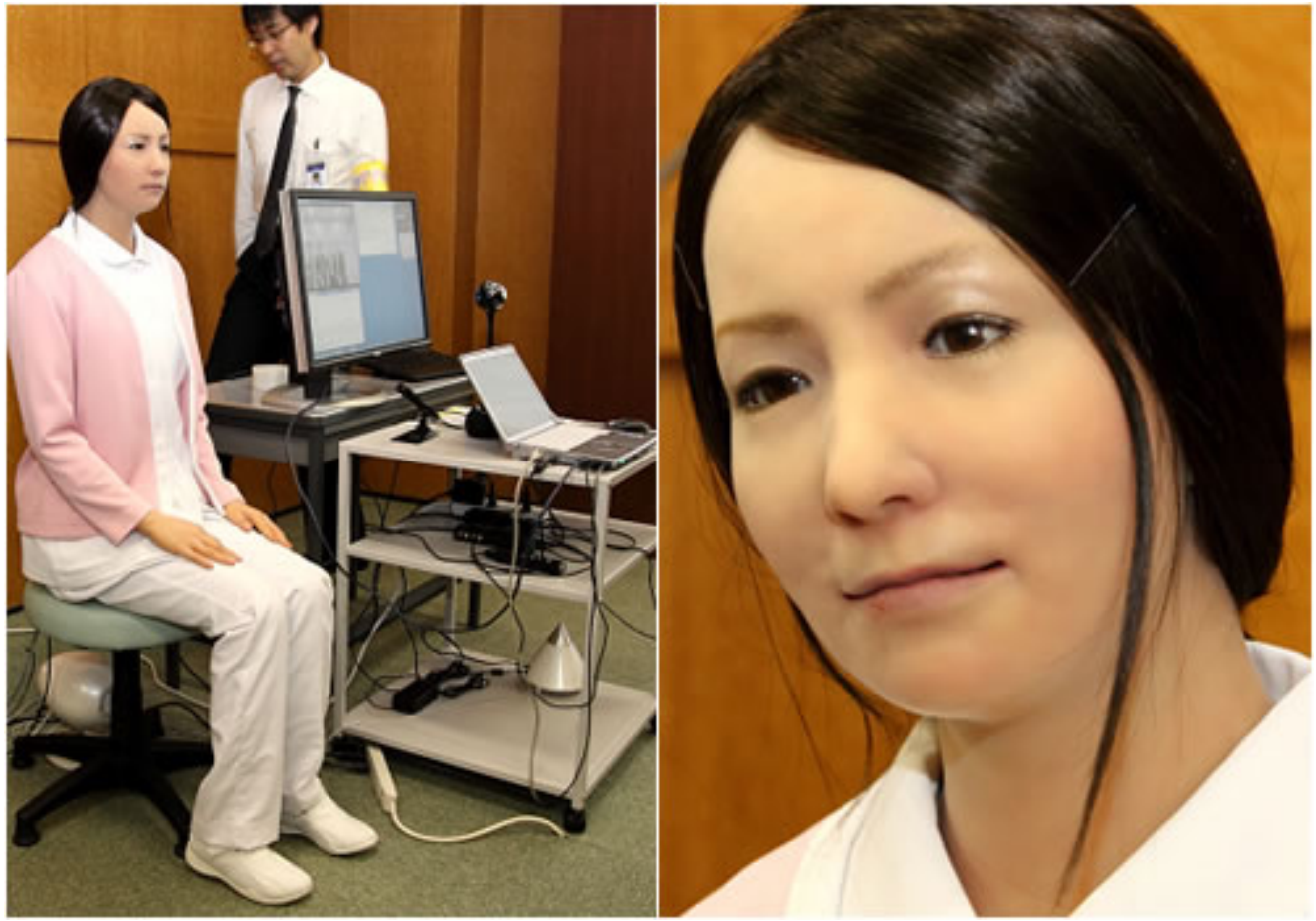
Instilling Morality in Machines

David Burke | Nuts & Bold Ideas Seminar | February 9, 2011

Robots are coming!



Actroid F - Robotic Nurse

Tsukuba Center, National Institute of Advanced Industrial Science and Technology, Japan

Huge Implications

- Increasingly sophisticated information processing leads to more judgment and decision-making; hence, more autonomy.
- (Aside: Human beings anthropomorphize at the drop of a hat.)
- Result: we're dealing with them as moral agents -- they have beliefs, goals, responsibilities.
- *How do you instill morality in a machine?*

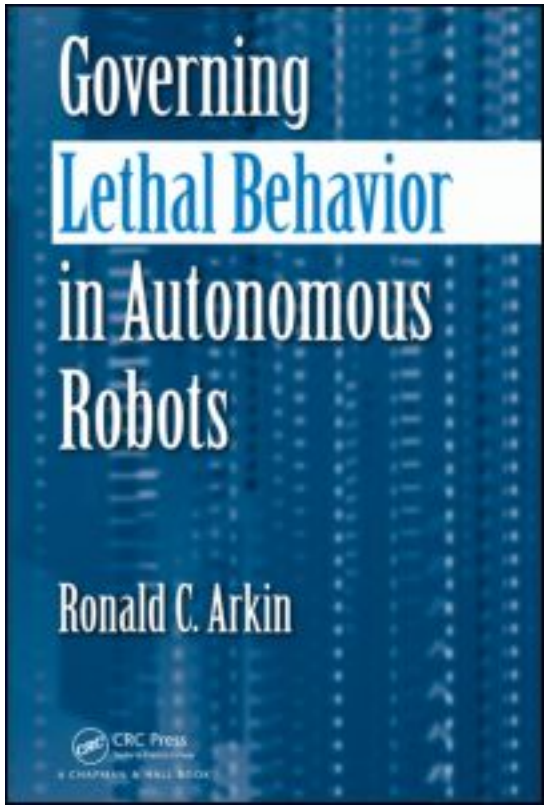
| galois |

Didn't Isaac Asimov Solve This Problem Already?

- Asimov's Laws of Robotics:
  - 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
  - 2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.
  - 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
  - 0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Ronald Arkin's Work

- "*Humane-oids* - robots that can potentially perform more ethically in the battlefield than humans are capable of doing."
- Approach: codification of the Laws of War (LOW) and Rules of Engagement (ROE).



Logic-based approaches

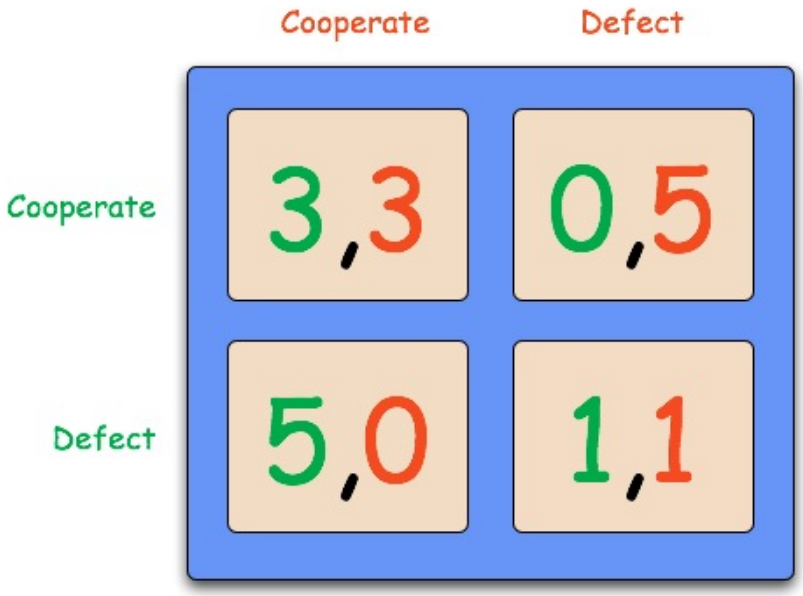
- “A robot can flawlessly obey a ‘moral’ code of conduct and still be thoroughly, stupidly, catastrophically immoral.”
- “…control robot behavior by fundamental ethical principles encoded in deontic logic…”

Moral Monocultures

- Fascinating Tradeoff:
  - perfect copying - one of the defining characteristics of software
  - diversity - ubiquitous strategy in biology
- Imagine the eventual large-scale successors to today's swarm robotics experiments -- do we want a 'moral monoculture'?
- My proposal: some kind of moral pluralism for autonomous systems.

Strategic interactions

- "The prisoner's dilemma is to game theorists what the fruit fly is to biologists"
- Many multiagent simulations & tournaments are based on this simple game.
- Idea: play the prisoner's dilemma (as well as other games) with a diverse population w.r.t. moral decision-making



| galois |

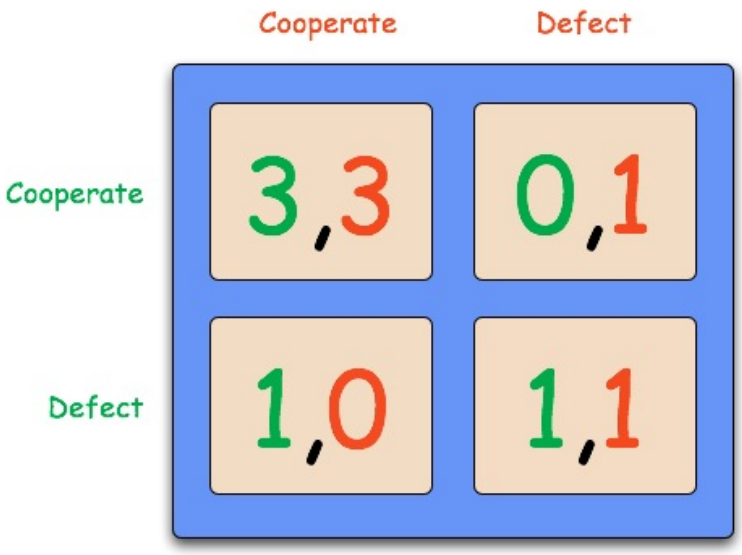
Moral Foundations Theory

1. Reciprocity/Fairness
2. Harm/Care
3. Ingroup/Loyalty
4. Authority/Respect
5. Purity/Disgust

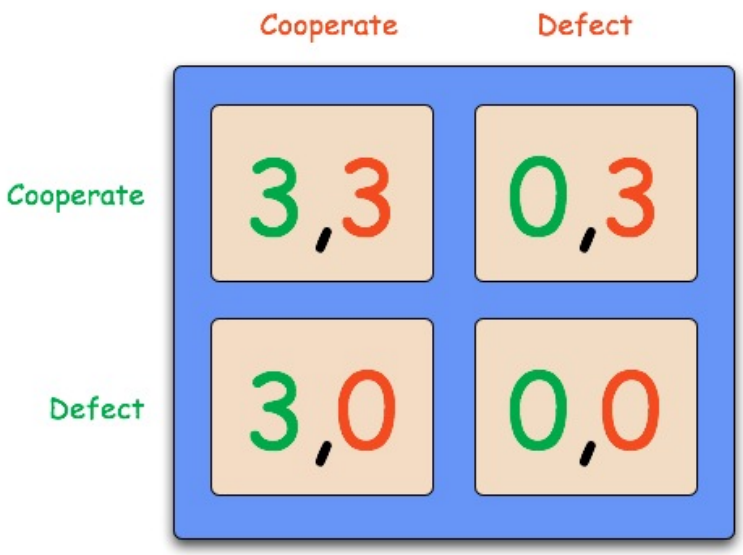
Multiagent Simulation

- Implement a genetic algorithm:
  - Instantiate a starting set of agents with various strengths for the five moral attributes
    - For each attribute, we have a value, and a weighting.
    - Each agent also has an attribute ordering, and a decision style.
  - Let the agents interact; the successful ones breed
  - Watch the population evolve through the generations.

Other Strategic Interaction Games



"Stag Hunt"



"Dependence"

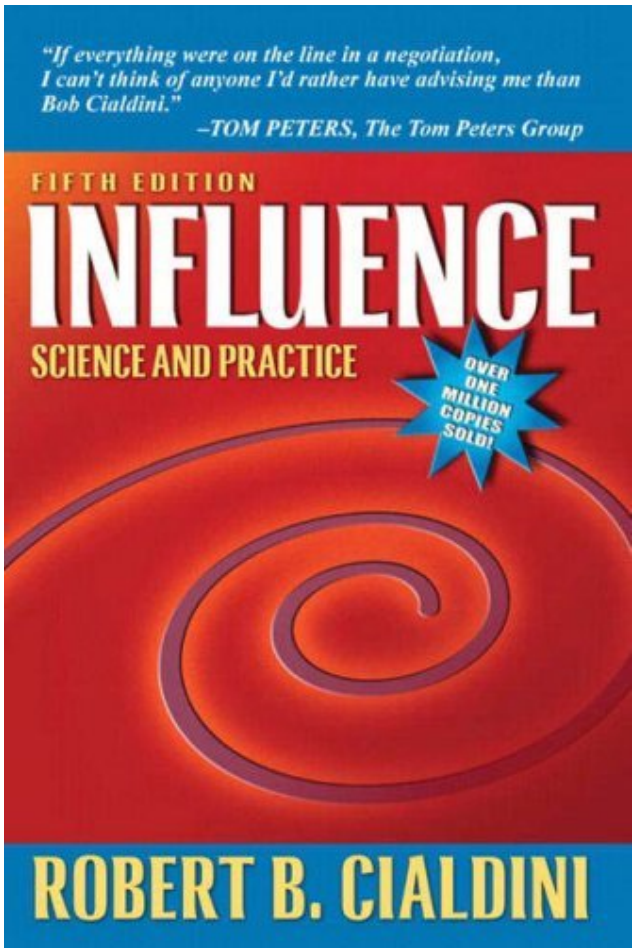
|galois|

Playing with the model

- Some preliminary results…
- Topology of contacts
  - random vs. locality, ability to move, etc.
- Percentage culled with each generation
- *What about cultural transmission?* Accounting for cultural influence during a lifetime.
- (Also, we'd like the model to be as endogenous as possible.)

Social Influence

- Six keys to influence:
  - Reciprocity
  - Commitment & Consistency
  - Social Proof
  - Authority
  - Liking
  - Scarcity



|galois|

Empathy - the ‘Holy Grail’?

- Prosociality of human beings
- Some versions of empathy:
  - Knowing somebody’s else’s thoughts or feelings
  - Coming to feel as another person feels
  - Imagining how another person is thinking and feeling
  - Feeling distress at somebody else’s suffering
- Computational Empathy?

Selected Links

- Ronald Arkin
  - Home page: http://www.cc.gatech.edu/aimosaic/faculty/arkin/
- Selmer Bringsjord (RAIR lab)
  - Home page: http://www.rpi.edu/~brings/
  - A video of his talk on this subject: http://www.vimeo.com/4032291
- Jonathan Haidt
  - Home page: http://people.virginia.edu/~jdh6n/
  - Moral foundations page: http://faculty.virginia.edu/haidtlab/mft/index.php

| galois |

Contact Info

David Burke
davidb@galois.com
(503) 808-7175 (office)
(503) 330-9512 (cell)