6-3-2011

Evolving Machine Morality Strategies through Multiagent Simulations

David Burke
Galois, Inc.

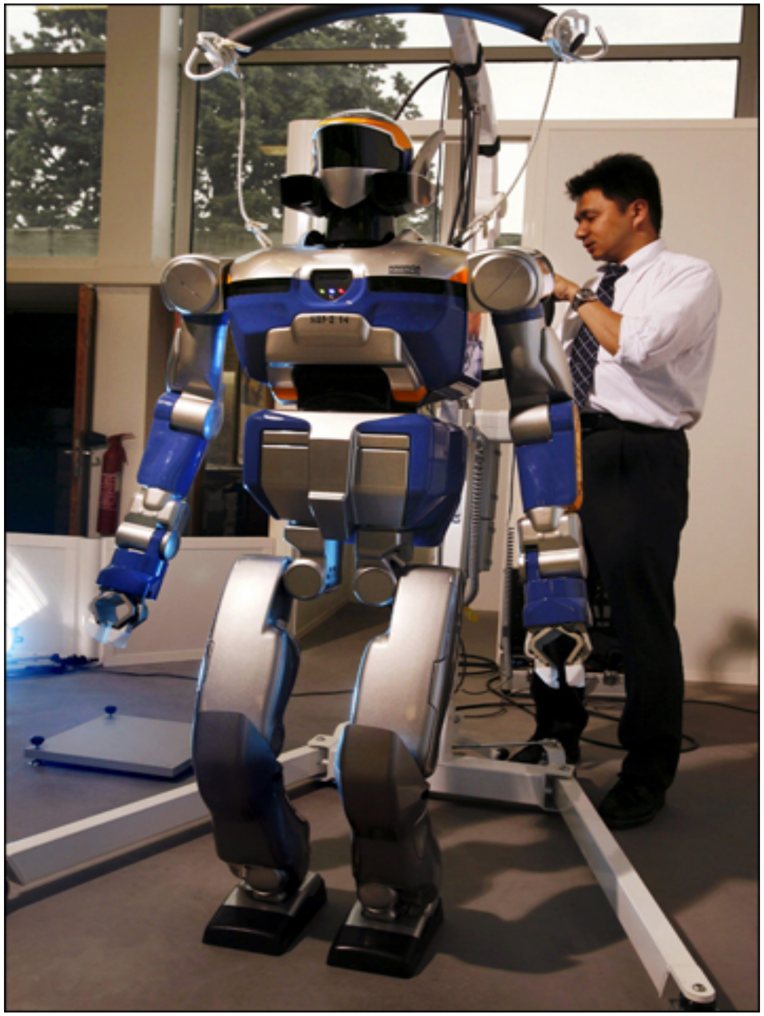
Recommended Citation

Burke, David, "Evolving Machine Morality Strategies through Multiagent Simulations" (2011). *Systems Science Friday Noon Seminar Series*. 65.
https://pdxscholar.library.pdx.edu/systems_science_seminar_series/65

Instilling Morality in Machines- Multiagent Experiments

David Burke | Systems Science Seminar | June 3, 2011

Robots are coming!



- In Japan, researchers anticipate that robot nurses will be the answer to demographic changes.
- iRobot builds various robots for bomb disposal, carrying payloads, gathering "situational awareness".
- Futurists like Ray Kurzweil predict "…we will have both the hardware and software to achieve human-level intelligence in a machine by 2029"

Huge Implications

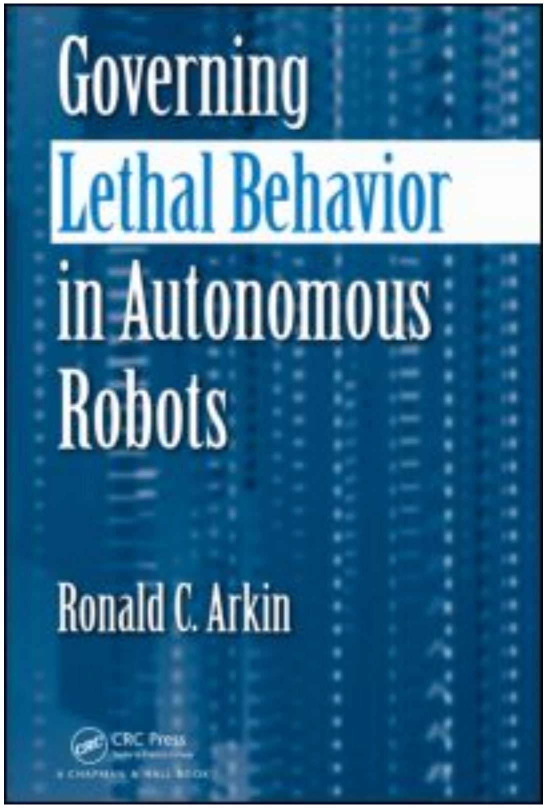
- Increasingly sophisticated information processing leads to more judgment and decision-making; hence, more autonomy.
- Human beings anthropomorphize at the drop of a hat -- yelling at cars & computers.
- Jesse Bering: "…we sometimes can't help but see intentions, desires, and beliefs in things that haven't even a smidgeon of a neural system."
- Result: we're dealing with them as moral agents -- they have beliefs, goals, responsibilities.
- *How do you instill morality in a machine?*

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

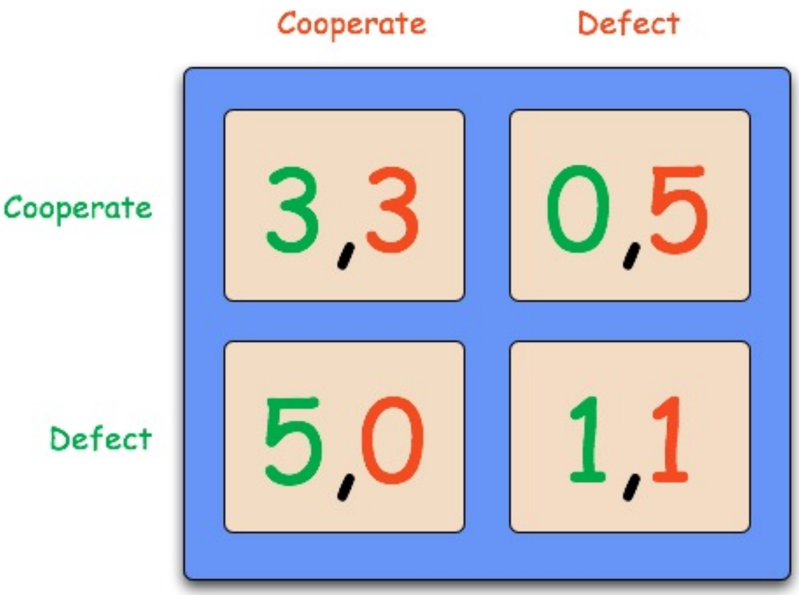
- "A robot can flawlessly obey a 'moral' code of conduct and still be thoroughly, stupidly, catastrophically immoral."
- "…control robot behavior by fundamental ethical principles encoded in deontic logic…"

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



	Cooperate	Defect
Cooperate	3,3	0,5
Defect	5,0	1,1

Moral Foundations Theory

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

Are any of these attributes more foundational than the others?

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.
- The basic version of the simulation is ~600 lines of Python.

Other Strategic Interaction Games

	Cooperate	Defect
Cooperate	3,3	0,1
Defect	1,0	1,1

"Stag Hunt"

	Cooperate	Defect
Cooperate	3,3	0,3
Defect	3,0	0,0

"Benevolence"

Attributes

- each agent assigned to a ‘tribe’
- ‘decStyle’- first attribute vs. weighted (two weighting schemes)
- each attribute votes ‘C’ or ‘D’ (>= or < 0)
- each attribute has a weight (0 to 1)
- ‘recip’ - default, and choices for last round being ‘CC’, ‘CD’,’DC’,’DD’
- ‘harm’ - delta between agent scores
- ‘auth’ - compare agent scores
- ‘loyal’ - compare agent tribes
- ‘disgust’ - agent1 checks to see if agent2’s tribe is a member of agent1’s disgust list.
- The 5 attributes are combined for a total (unless the decision style is ‘first’)

(very) Preliminary results

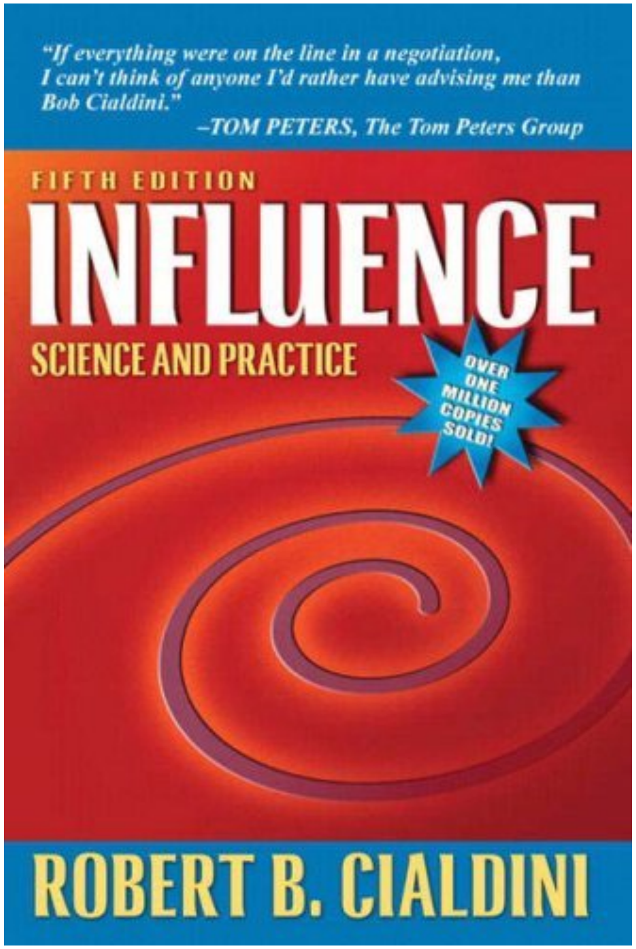
- Initial experiments featured five tribes, a population of 1000 agents, evolving over 250 generations, and runs for each of the three games.
- I had guessed that the “meaner” the game, the more we’d see traits like loyalty and authority dominate the population. (>80% of the population)
- Actual results: reciprocity and loyalty generally dominated the runs, but the “meaner” the game, the more likely that reciprocity came out ahead.
- More often than not, “first” decision-making outweighed “weighted” decision styles.
- A higher percent culled speeds up convergence, but doesn’t appear to affect the shape of the final landscape.

Playing with the model

- Number of tribes; number of agents; number of generations
- Topology of contacts
  - random
  - local
  - movement allowed each generation
- Percentage culled with each generation
- *What about cultural transmission?* Accounting for cultural influence during a lifetime - right now, the agents don't learn from experience.
- How can we make the model more endogenous?

Making the model endogenous: Social Influence

- Six keys to influence:
  - Reciprocity
  - Commitment & Consistency
  - Social Proof
  - Authority
  - Liking
  - Scarcity
- Add costs to these efforts


"If everything were on the line in a negotiation, I can't think of anyone I'd rather have advising me than Bob Cialdini."
–TOM PETERS, The Tom Peters Group
FIFTH EDITION
INFLUENCE
SCIENCE AND PRACTICE
OVER ONE MILLION COPIES SOLD!
ROBERT B. CIALDINI

Empathy

- 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 -- true empathy vs. “as if” 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)