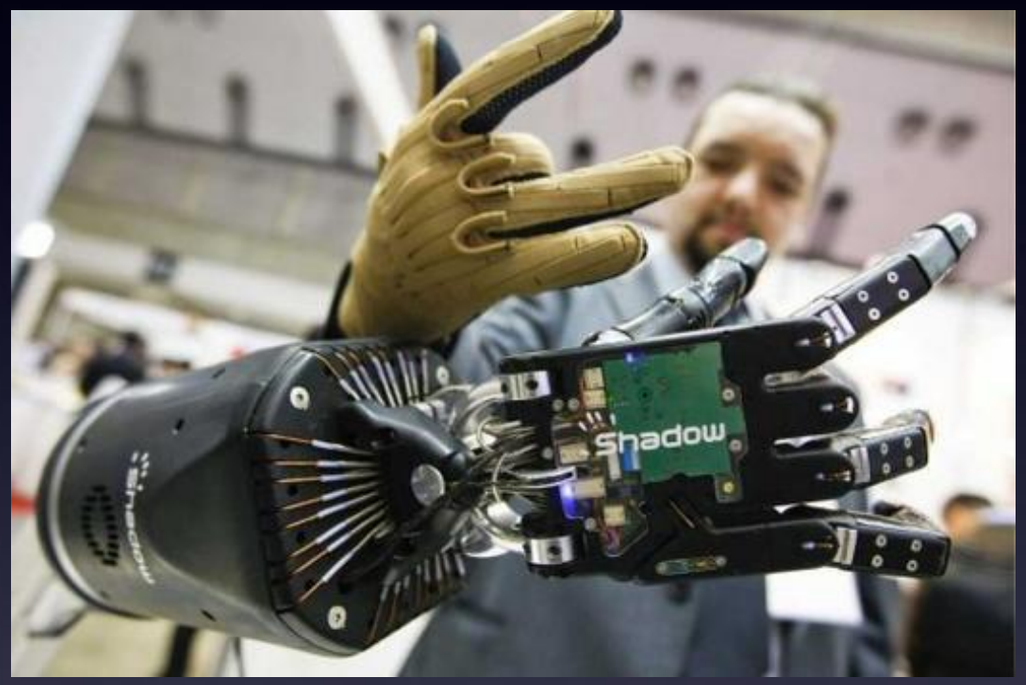
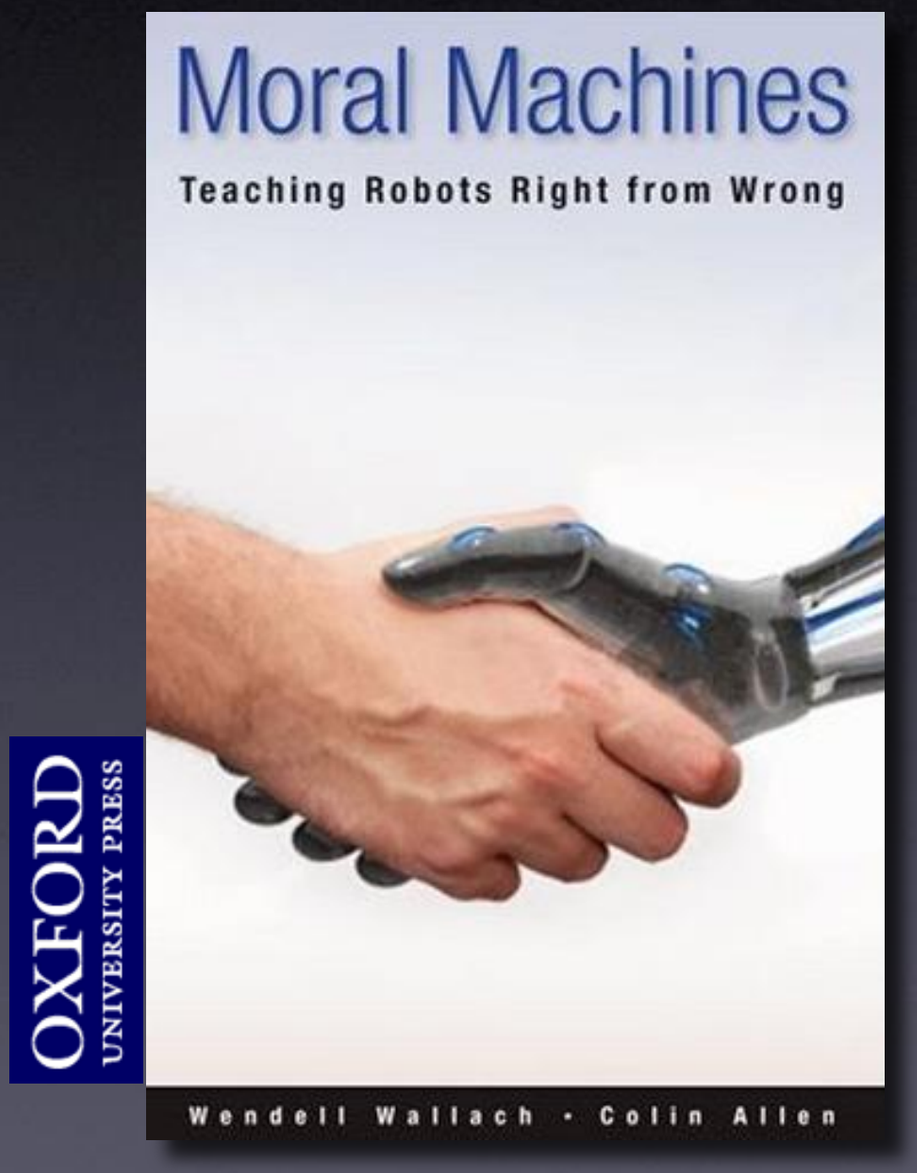
Engineering for Responsibility



Wendell Wallach
The Governance of Emerging Technologies
Arizona, May 20th, 2013

Mapping a New Field of Inquiry

Professor Colin Allen
Indiana University



Machine Morality
Machine Ethics
Computational Ethics
Artificial Morality
Friendly AI
Roboethics

- “The greater the freedom of a machine, the more it will need moral standards.”

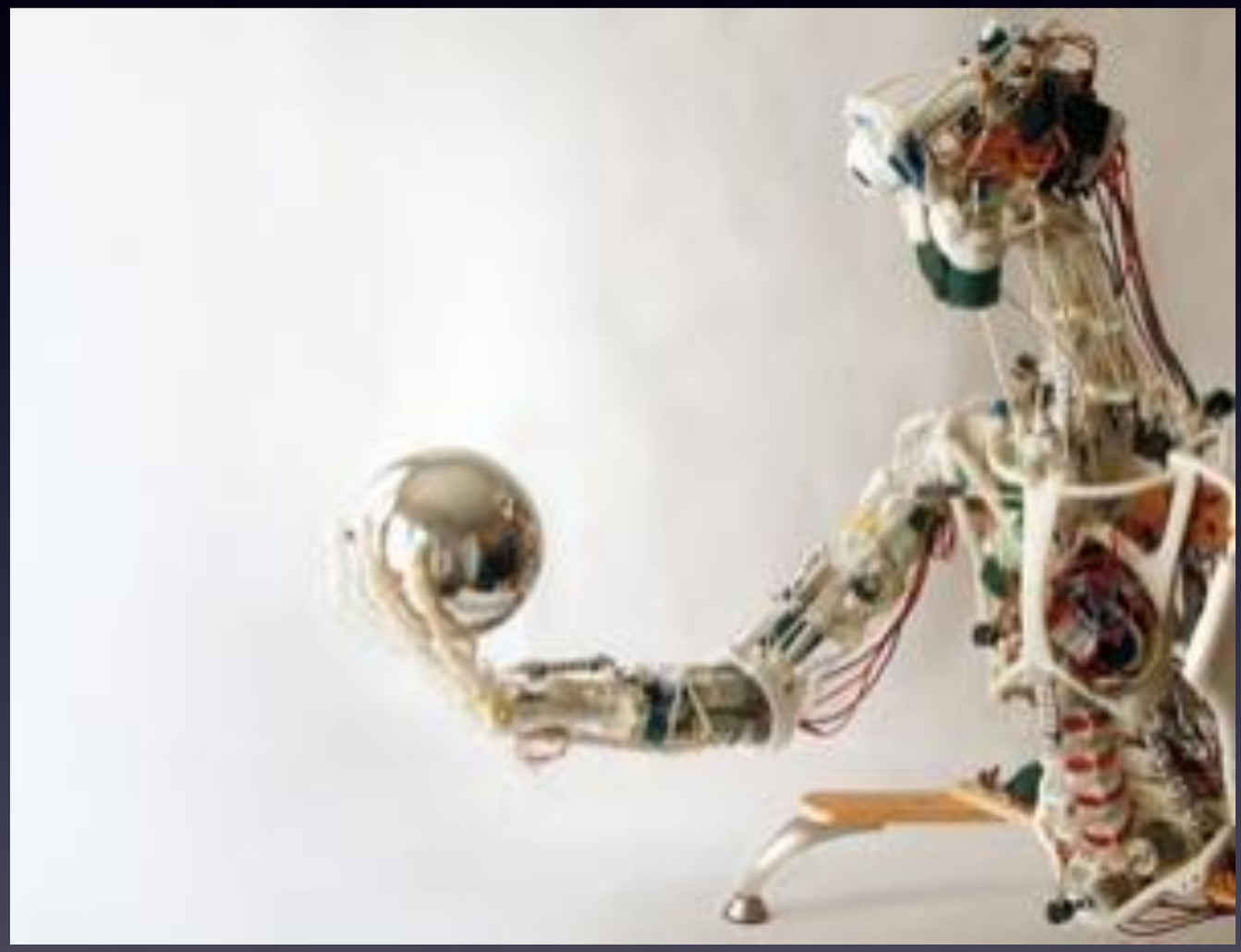
  - Rosalind Picard, Director

    M.I.T. Affective Computing Group

- If robots can be designed so that they are sensitive to ethical considerations and factor those considerations into their choices and actions, new markets for their adoption will be opened up.
- If they fail to adequately accommodate human laws and values, there will be demands for regulations that limit their use.

Ethical and Legal Concerns

- Safety
- Privacy and Property Rights
- Criminal Activity
- Freedom and Free Inquiry
- Appropriate Use
- Capabilities
- Responsibility



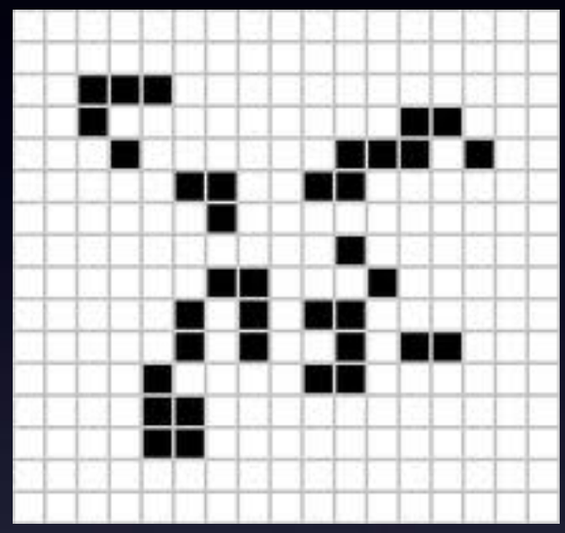
Alas, poor Yorick! I knew him.

Increasingly Autonomous Robotic Systems

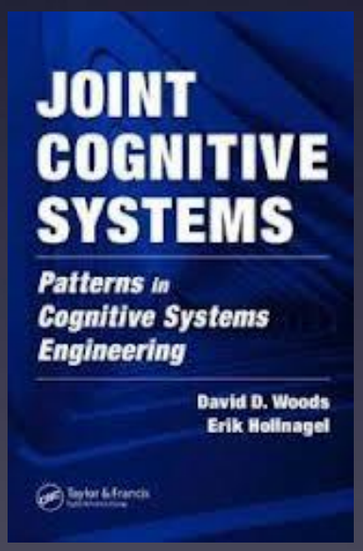
threatens to undermine the foundational principle that a human agent (either individual or corporate) is responsible, and potentially accountable and liable, for the harms caused by the deployment of any technology.

Responsibility

- Difficulty in predicting actions of algorithmic and complex systems.
  - ‘many hands’ (Nissenbaum, 1997)



- Joint Cognitive Systems/Coordination
  - Woods and Hollnagel - Global Hawk UAV – December 6th, 1999 – $5.3M



”we have constructed a world in which the potential for high-tech catastrophe is embedded in the fabric of day-to-day life.”

Malcolm Gladwell

Engineer for Responsibility

- Design of Artefacts
- Social Engineering

Limit any Dilution of Responsibility

Rules for Computing Artefacts

- Rule 1: The people who design, develop, or deploy a computing artefact are morally responsible for that artefact, and for the foreseeable effects of that artefact. This responsibility is shared with other people who design, develop, deploy or knowingly use the artefact as part of a sociotechnical system.

- Rule 2: The shared responsibility of computing artefacts is not a zero-sum game. The responsibility of an individual is not reduced simply because more people become involved in designing, developing, deploying or using the artefact. Instead, a person's responsibility includes being answerable for the behaviours of the artefact and for the artefact's effects after deployment, to the degree to which these effects are reasonably foreseeable by that person.

- Rule 3: People who knowingly use a particular computing artefact are morally responsible for that use.

- Rule 4: People who knowingly design, develop, deploy, or use a computing artefact can do so responsibly only when they make a reasonable effort to take into account the sociotechnical systems in which the artefact is embedded.

- Rule 5: People who design, develop, deploy, promote, or evaluate a computing artefact should not explicitly or implicitly deceive users about the artefact or its foreseeable effects, or about the sociotechnical systems in which the artefact is embedded.

Oversight

- Governance Mechanism
  - monitor
    - whether the lines of responsibility have been established for new systems being deployed.
    - thresholds are about to be crossed that pose serious dangers.

Machines must not make ‘decisions’ that result in the death of humans.

- Ban on ‘Killer Robots’
  - ICRAC (Berlin, Oct ’10)
  - Call for an Executive Order (Feb ‘12)
    - ALFIS violates LOAC & IHL
- HRW/Harvard Law School Human Rights Clinic (Nov. 19th, 2012)
- DoD Autonomy in Weapons Systems (Nov. 23rd , 2012)
  - Undersecretary Ashton Carter

Key Advantages

Samsung Techwin

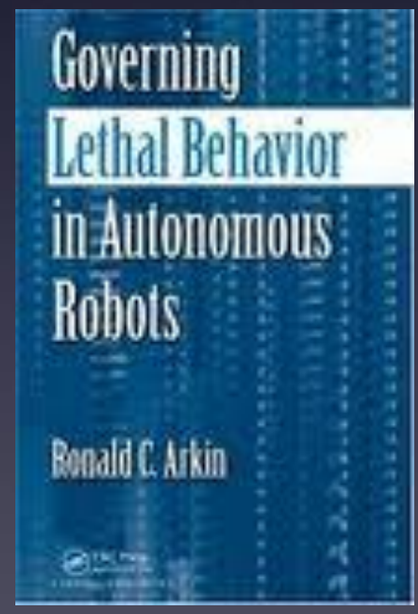
- Increase capabilities – e.g., remote attacks
- Reduce collateral damage through greater precision.
- Decrease the loss of personnel during hostilities.
- Lower manpower costs.
- Enable projection of force in a future where manpower resources will be potentially be far more limited.

Core Criticisms

- The inability to assess who is accountable for the actions taken by autonomous lethal force initiating systems (ALFIS)
- Will lower the barriers to starting new wars.
- Use for surveillance activities unrelated to achieving military objectives.
- ALFIS would be be dangerous from an operational perspective – e.g., potential for conflict escalation and disproportionate or indiscriminate use of force in the absence of human review.
- Once developed, in view of existing loopholes in current export control mechanisms, these systems are likely to proliferate widely.
- The proliferation of increasingly autonomous weaponry introduces a seriously unpredictable element into future conflicts, especially since many governments and non-government actors might not program in the constraints and limits that the United States would.

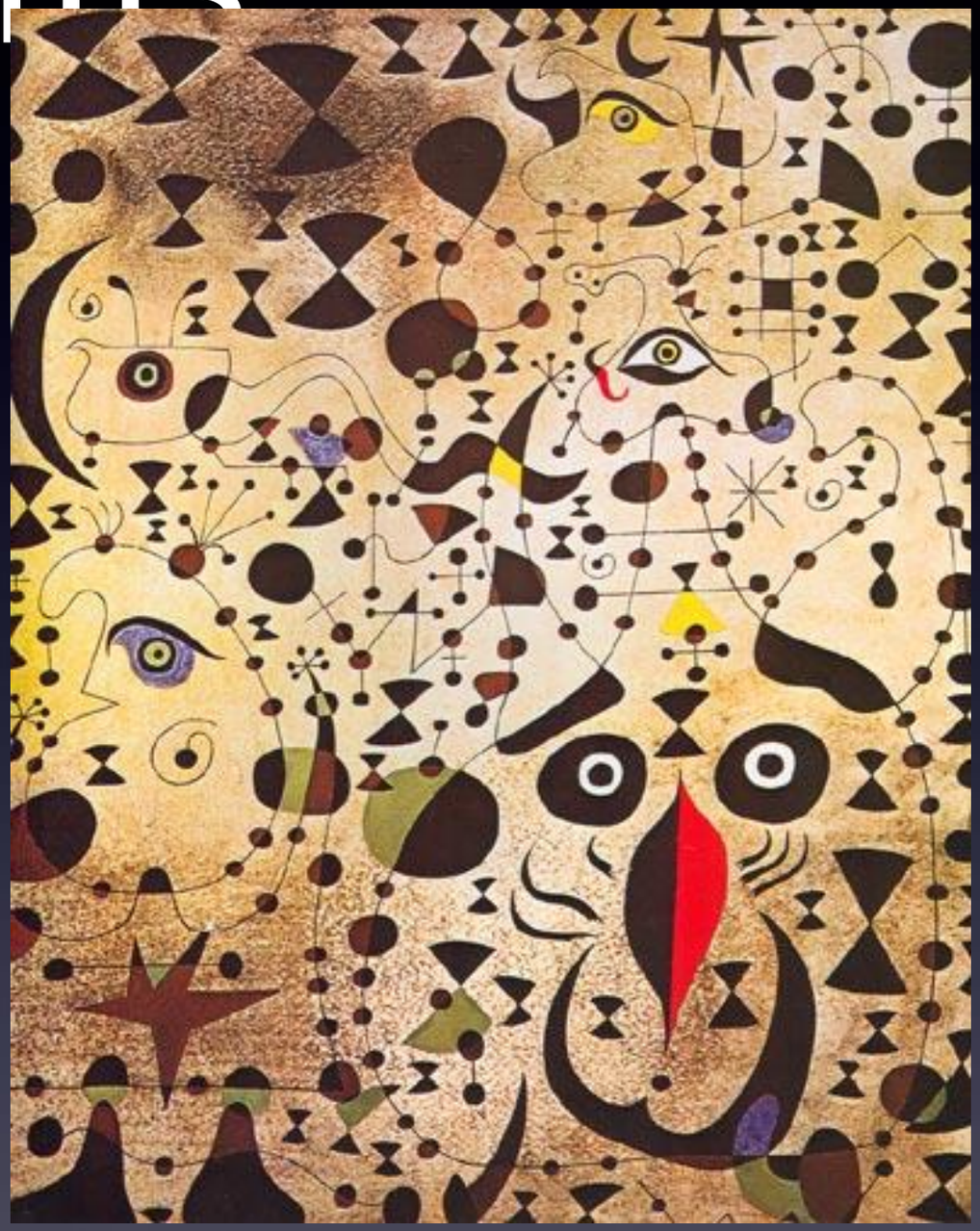
Robot Soldiers

- Ronald Arkin
- LOAC and ROE
  - More Moral Than Human Soldiers
  - MHAT IV
- Effect on Military Policy
- Need for a neutral party to monitor likely capabilities



The Two Hard Problems

- Implementing norms, rules, principles, or procedures for making moral judgments
- Framing Problems



Framing Problems

How does the system recognize it is in an ethically significant situation?

How does it discern essential from inessential information?

How does the AMA estimate the sufficiency of initial information?

What capabilities would an AMA require to make a valid judgment about a complex situation? e.g., combatant v. non-combatant.

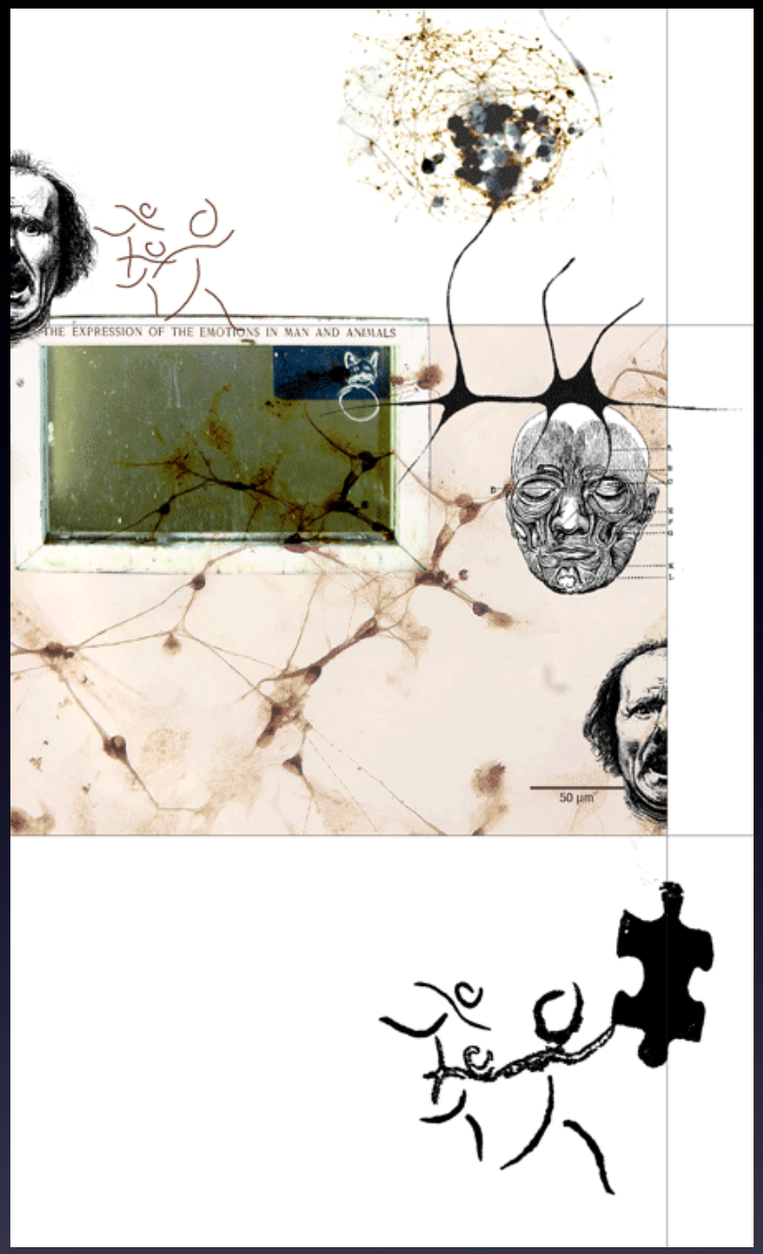
How would the system recognize that it had applied all necessary considerations to the challenge at hand or completed its determination of the appropriate action to take?

Moral Agency - AMA

- If and when robots become ethical actors that can be held responsible for their actions, we can begin debating whether they are no longer machines and are deserving of some form of personhood.
- ‘mala in se’
  - Not because they are machine
  - Unpredictable, can not be fully controlled, and attribution of responsibility is difficult if not impossible

Set Limits

Once such a principle is in place we can go on to the more exacting discussion as to the situations in which robots and information systems are indeed an extension of human will and intention and when their actions are beyond direct human control.



Thank You!

Email: wendell.wallach@yale.edu