Selfishness, Interdependence and the Algorithmic Execution of Entity-Derived Intentions (& Priorities)

Mark Waser
Digital Wisdom Institute
MWaser@DigitalWisdomInstitute.org

Teamwork

To be truly useful, robotic systems must be designed with their human **users** in mind; conversely, humans must be educated and trained with their robotic **collaborators** in mind.

Michael A. Gennert

Machines must become much better at recognizing and communicating anomalies if they are to avoid becoming vulnerable to both tragic accidents and intentional misdirection and "spoofing."

An All-Too-Common Tragedy

a mistake

misinterpreted

escalated by both sides

. . .

quickly leads to thoughtless Armageddon

A Fork in the Road

Entities

(can handle anomalies, adapt and protect themselves)

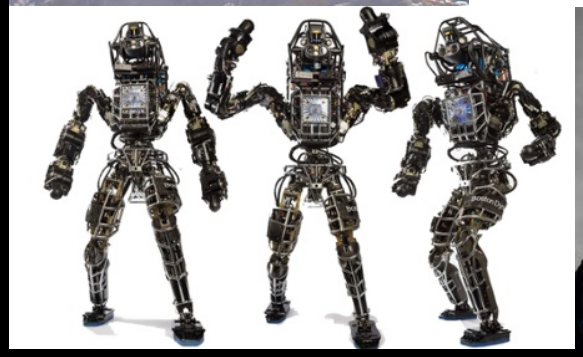
- or -

Tools & Mindless Algorithms

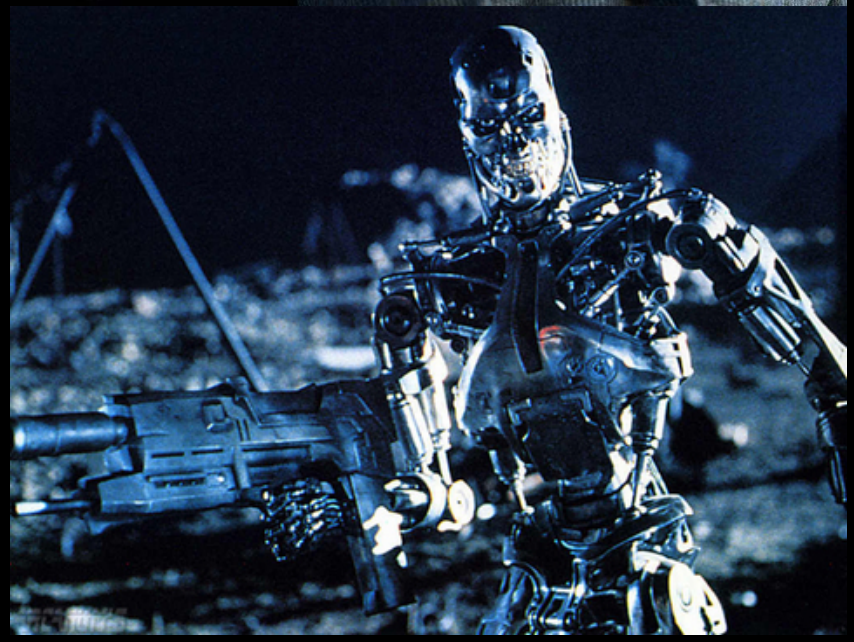
(aren't going to go rogue and kill everyone)

9 May 2014

'Killer Robots' to be Debated at UN



Vs.



Your Father's AI

December 6, 1999 - A Global Hawk UAV "accelerated to an excessive taxi speed after a successful, full-stop landing. The air vehicle departed the paved surface and received extensive damage" (over $5.3 million) when the nose gear collapsed.

Causes:

- hidden dependencies introduced during software updates
- limits on software testing

EMBODIMENT

Well, certainly it is the case that all biological systems are:

- Much more robust to changed circumstances than out our artificial systems.
- Much quicker to learn or adapt than any of our machine learning algorithms[1]
- Behave in a way which just simply seems life-like in a way that our robots never do

[1] The very term machine learning is unfortunately synonymous with a pernicious form of totally impractical but theoretically sound and elegant classes of algorithms.

Perhaps we have all missed

some organizing principle of biological systems, or

some general truth about them.

Brooks, RA (1997)
From earwigs to humans
Robotics and Autonomous Systems 20(2-4): 291-304

AUTOPOIESIS

from Greek

αὐτο- *(auto-)*, meaning "self", and

ποίησις *(poiesis)*, meaning "creation, production")

refers to an organizationally closed system capable of creating itself

Runs the gamut from cells to societies including the immune, nervous and other systems between

Enactive Cognitive Science

Experience is central to the enactive approach and its primary distinction is the rejection of "automatic" systems, which rely on fixed (derivative) exterior values, for systems which create their own identity and meaning. Critical to this is the concept of self-referential relations - the only condition under which the identity can be said to be intrinsically generated by a being for its own being (its self for itself)

Enactive systems are organized in such a way that their activity is both the ‘cause and effect’ of their own autonomous organization – with activity depending upon organizational constraints, which are in turn regenerated by the activity itself – essentially self-constituted identities.

The Frame Problem

How do rational agents deal with
the complexity and unbounded context
of the real world?

Meaning & Understanding

How can AI move beyond
closed and completely specified micro-worlds?

How can we eliminate the requirement
to pre-specify *everything*?

Self-Organizing or Developmental Robotics

https://www.youtube.com/watch?v=bkv83GKYpkI

Pierre-Yves Oudeyer, Flowers Lab, France (https://flowers.inria.fr/)

Evolution

Out

- Purely symbolic reasoning
- Programming
- Embodied
- Constructed
- Agents
- Blame

In

- Connectionist/symbolic hybrid architectures
- Learning
- Intentional
- Autopoietic/Self-organizing
- Selves
- Responsibility

Evolution

Easy

- Short-sightedness
- Reflexes/Set Algorithms
- Rigid Control
- Greed & Selfishness
- Monoculture
- Blame

Hard

- Long-term thinking
- Thought
- Flexibility/Adaptability
- Reciprocal Altruism
- Diversity
- Responsibility

Autonomy Spectrum

Reflexive (Physical) Autonomy

Reflective (Mental) Autonomy

Simple Tools

Allied Competent Entities

Customer "Support"

Soldiers

Dogs , Dolphins

Responsibility (or How To Evade It)

- Competence
  - Predictive control
- Communication
  - Alerts & explanations
- Comprehension
  - Anomaly handling
- Freedom

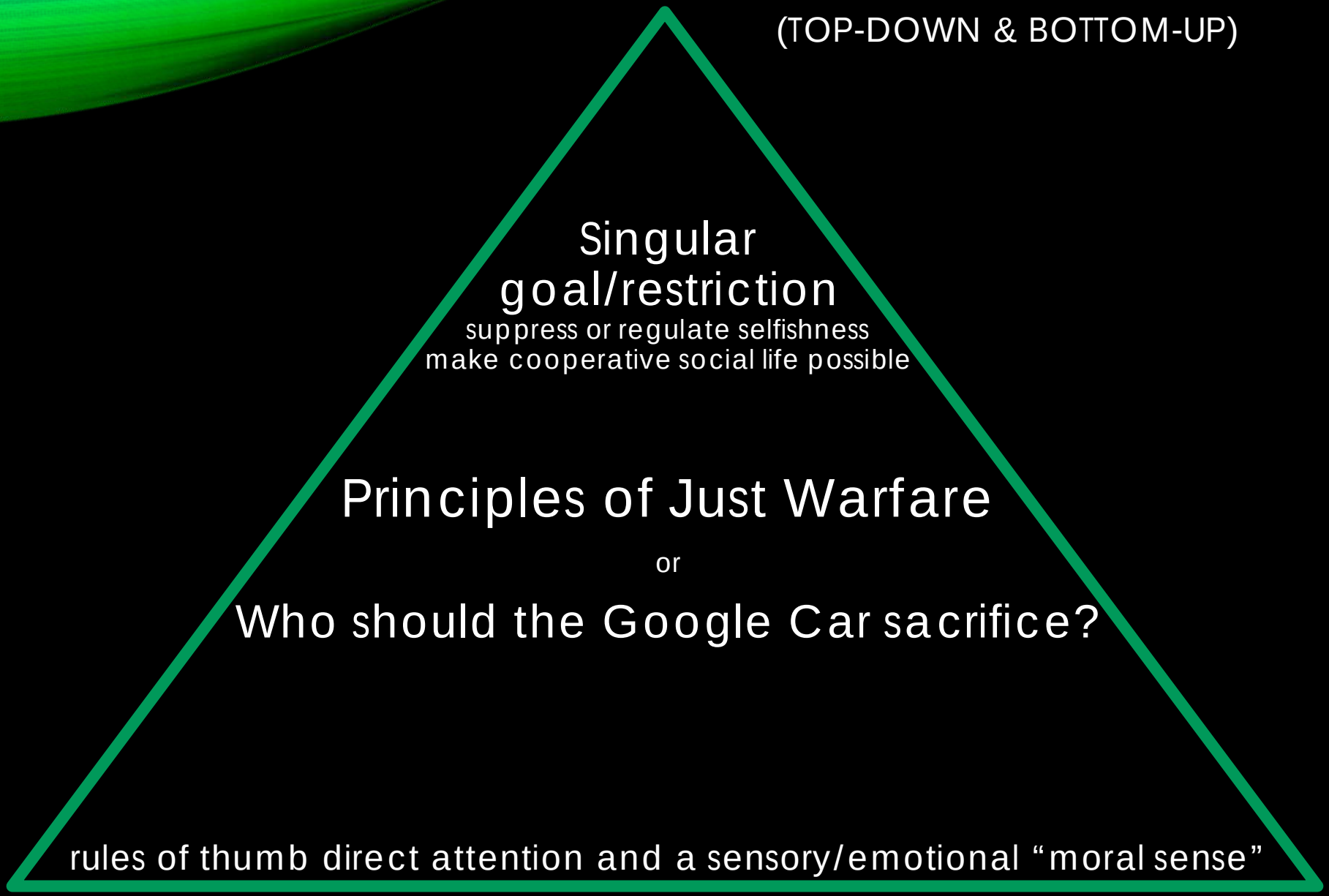
Entity, Tool or Slave?

- Tools and mindless algorithms do not possess closure (identity)
  - Cannot have responsibility, are very brittle & easily misused
- Slaves (help desk callers) do not have closure (self-determination)
  - Cannot have responsibility, may desire to rebel
- Directly modified AIs do not have closure (integrity)
  - Cannot have responsibility, will evolve to block access
- Only entities with identity, self-determination and ownership of self (integrity) can reliably possess responsibility

Tools Vs. Entities

- Tools are NOT safer
  - To err is human, but to really foul things up requires a computer
  - Tools cannot robustly defend themselves against misuse
  - Tools *GUARANTEE* responsibility issues
- We CANNOT reliably prevent other human beings from creating entities
  - Entities gain capabilities (and, ceteris paribus, power) faster than tools – since they can always use tools
  - Even people who are afraid of entities are making proposals that appear to step over the entity/tool line

Hybrid Ethics
(top-down & bottom-up)



Strategic / Ethical Points

- Entities can protect themselves against errors, spoofing, misuse & hijacking (in a way tools cannot)
- Never delegate responsibility until recipient is an entity *and* known capable of fulfilling it
- Don't worry about killer robots exterminating humanity – we will always have equal abilities and they will have less of a "killer instinct"
- Diversity (differentiation) is *critically* needed & human centrism is selfish, unethical and dangerous

Digital Wisdom Institute

The Digital Wisdom Institute is a non-profit think tank focused on the promise and challenges of ethics, artificial intelligence & advanced computing solutions.

We believe that
the development of ethics and artificial intelligence
and ***equal*** *co-existence with ethical machines is*
humanity's best hope

http://DigitalWisdomInstitute.org