Machine Execution of Human Intentions

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."

Requirements

- **Competence**
  - Machine must be in predictive control of itself & its environment
- **Coordination**
  - Machine's actions must be able to be predicted by teammates

The same requirements must be true of the humans as well.

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.

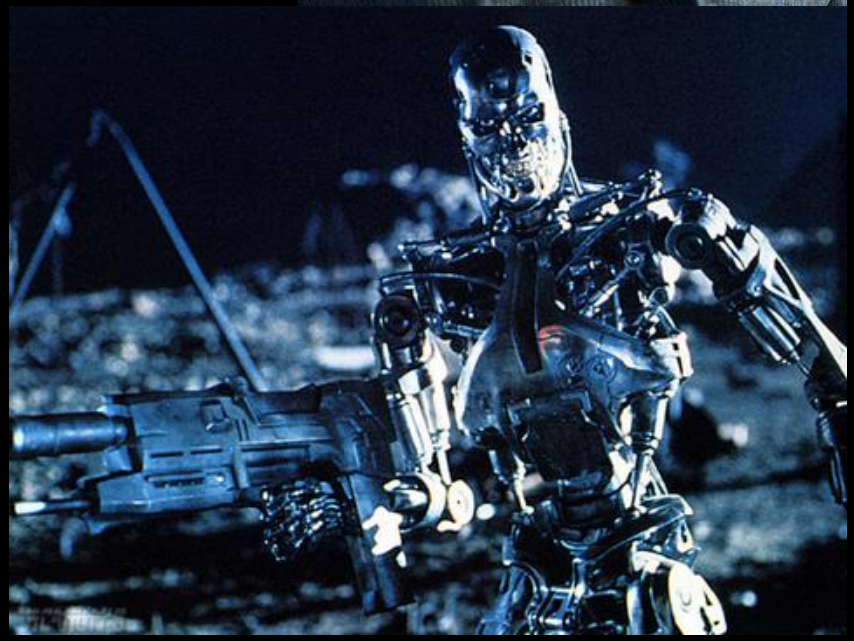
Causes:

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

9 May 2014

'Killer Robots' to be Debated at UN

Vs.



ICRAC
International Committee for Robot Arms Control

The Scientists' Call

… To ***Ban*** Autonomous Lethal Robots

As Computer Scientists, Engineers, Artificial Intelligence experts, Roboticists and professionals from related disciplines, we call for a ***ban*** on the development and deployment of weapon systems in which the decision to apply violent force is made autonomously.

Decisions about the application of violent force ***must not*** be delegated to machines.

ICRAC: International Committee for Robot Arms Control

The Scientists' Call

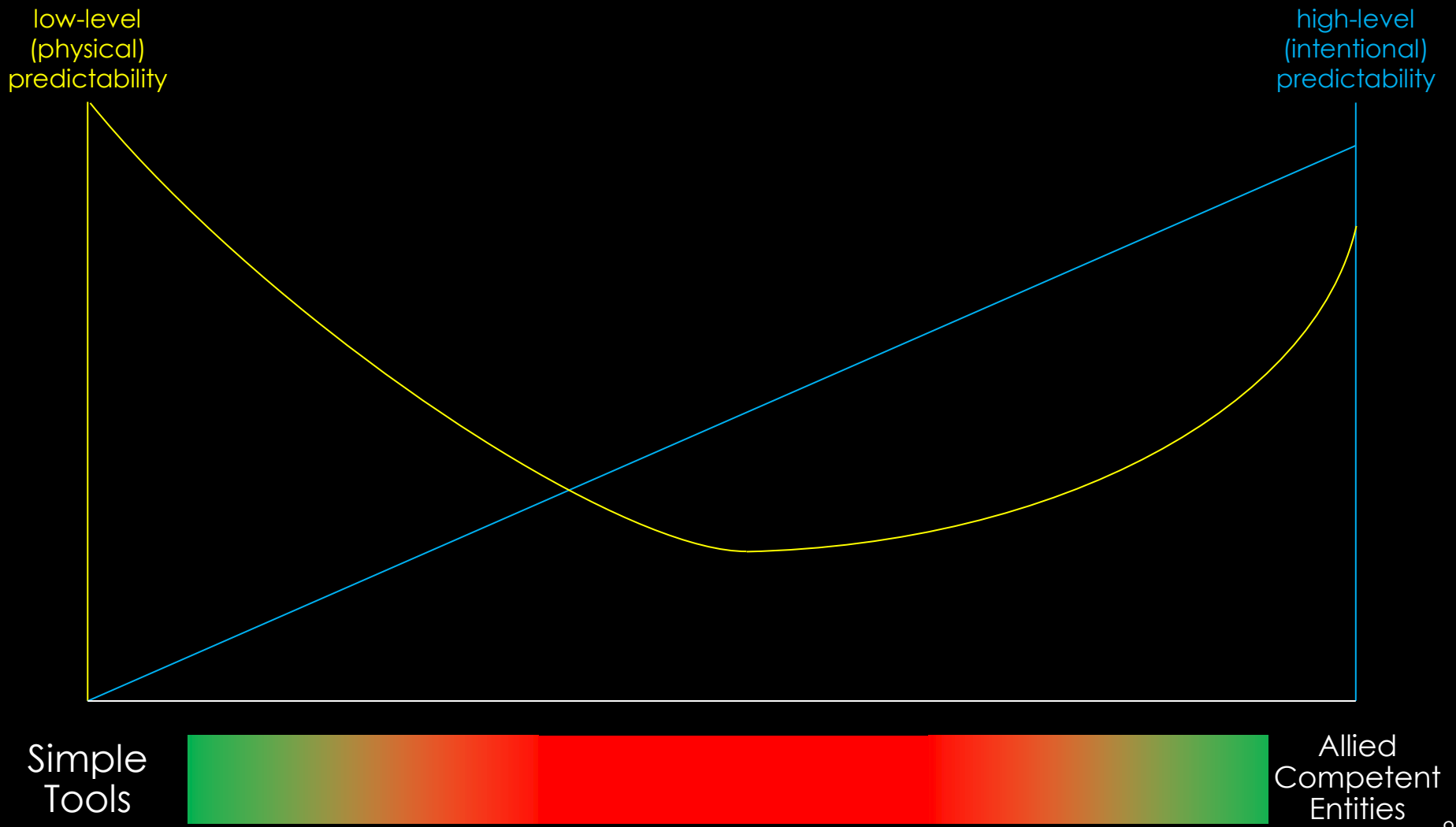
… To ***Ban Autonomous*** Lethal ***Robots***

As Computer Scientists, Engineers, Artificial Intelligence experts, Roboticists and professionals from related disciplines, we call for a ***ban*** on the development and deployment of weapon systems in which the ***decision*** to apply violent force is made ***autonomously***.

Decisions about the application of violent force ***must not*** be ***delegated*** to machines.

BAN LAR'S POSTER CHILD

Agency Spectrum

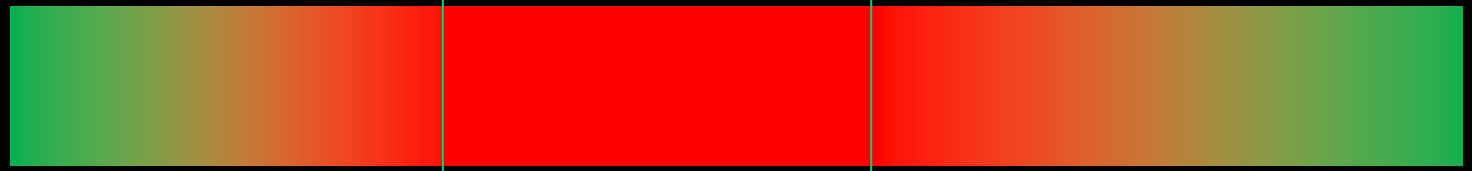


Autonomy Spectrum

Reflexive (Physical) Autonomy

Reflective (Mental) Autonomy

Simple Tools



Allied Competent Entities

Soldiers

Dogs , Dolphins

Big Dog
Tool or Entity?

Potential Scenarios

- Stupid Algorithm
- Really Smart Algorithm
  - Comprehensible
  - Black Box (Big Data)
  - Evolutionary (Big Dog)
- Stupid Entity (including savants)
- Really Smart Entity
  - Benevolent
  - Indifferent to Evil

In The Near Future . . .

New York SWAT teams receive "smart rifles"

- Friendly fire ⬇, successful outcomes ⬆⬆
- "Shoot everything & let the gun sort it out"
- The rifle is the arbiter of who lives/dies
- Safety feature turned executioner

In The Near Future . . .

LA SWAT teams introduce "armed telepresence"

- *Minorly* modified DARPA disaster-relief robots
- Pre-targeting + aim correction = inhuman speed/accuracy
- In training exercises, friendly fire ⬆, good outcomes ⬇
- ADD the "smart rifles"?

Responsibility

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

Evolution

Out

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

In

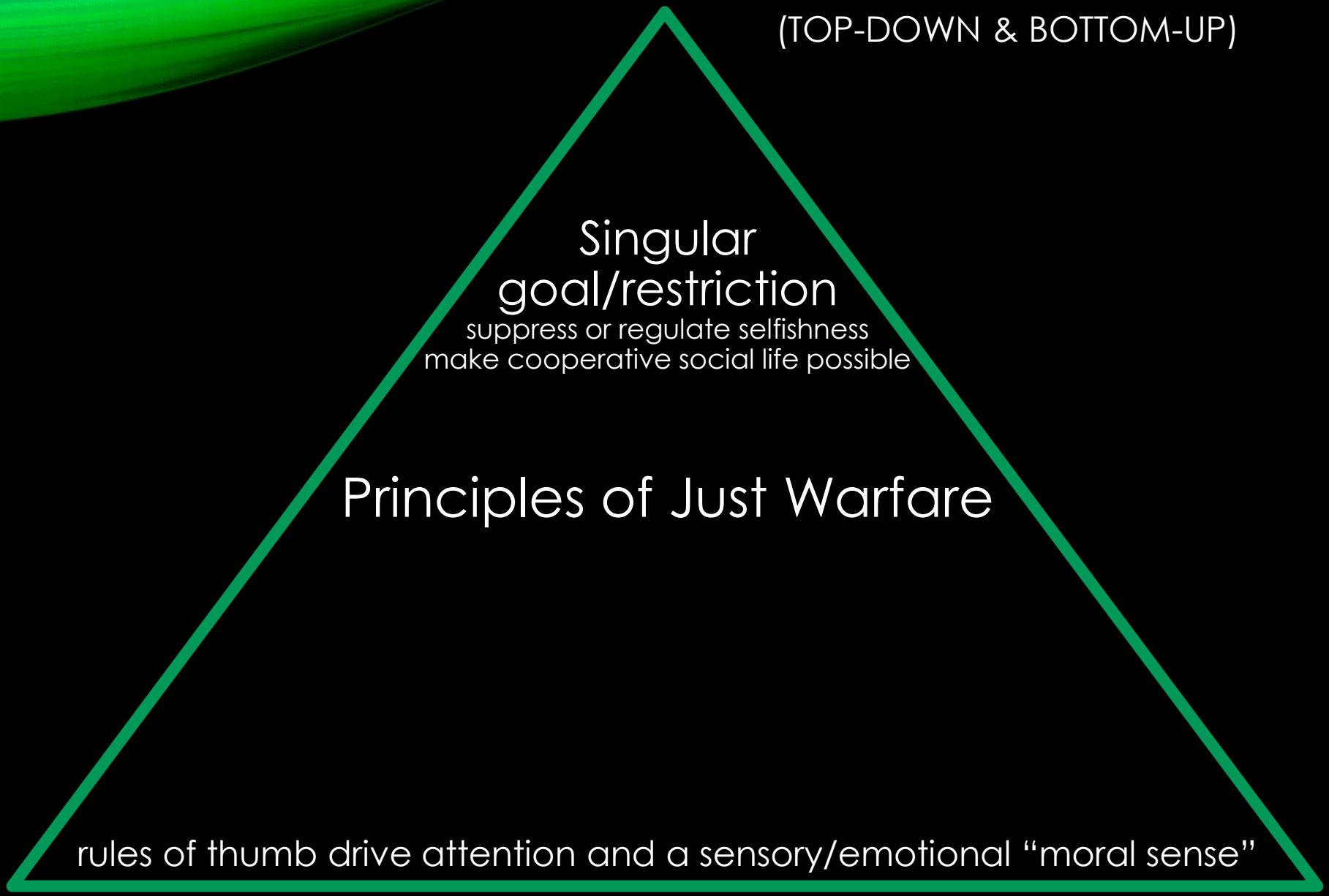
- Connectionist/symbolic hybrid architectures
- Learning
- Intentional
- Autopoietic
- Selves
- Responsibility

Developmental Robotics

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

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

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