

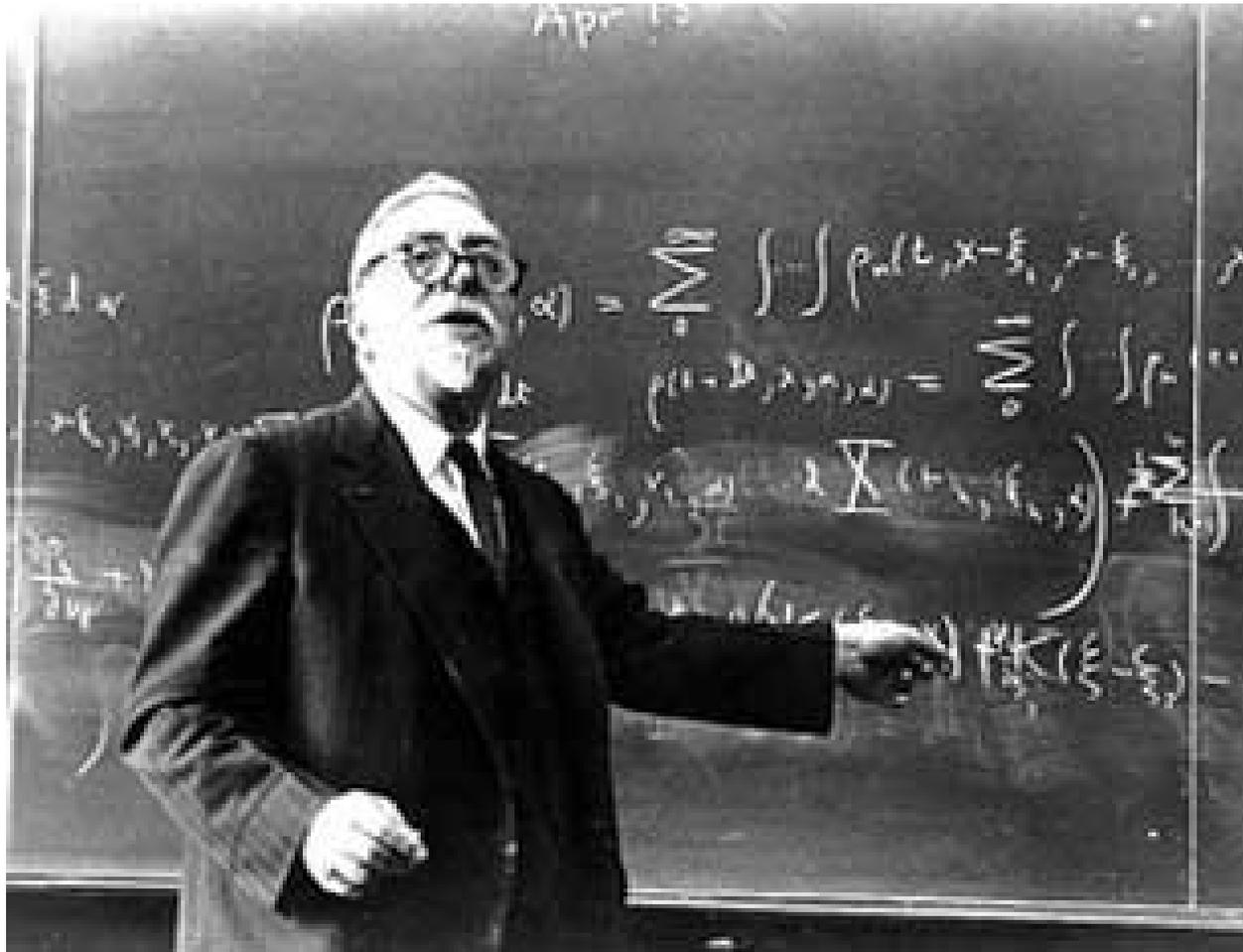
War Games & AI-Complete Problems

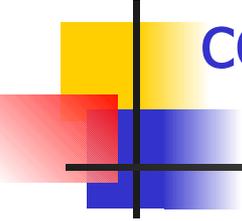
Guglielmo Tamburrini
Università di Napoli Federico II



Euron Atelier on RoboEthics
Genova, 2 march 2006

Norbert Wiener's concerns

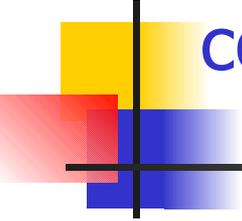




N. Wiener, "Some moral and technical consequences of automation", *Science*, May 1960

It is quite in the cards that **learning machines** will be used to program the pushing of the button **in a new push-button war...**

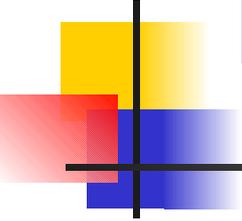
...the programming of such a learning machine would have to be based on some sort of war game...



N. Wiener, "Some moral and technical consequences of automation", *Science*, May 1960

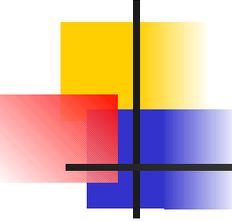
Here, however, if the rules for victory in a war game do not correspond to what we actually wish...

such a machine may produce a policy which would win a nominal victory on points **at the cost of every interest we have at heart...**



Main Source for Wiener's concern

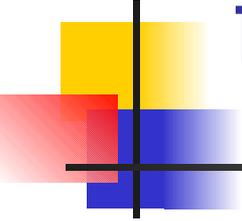
- **Unpredictability of machine behaviour from machine learning**
 - **Is Wiener's concern properly addressed today?**



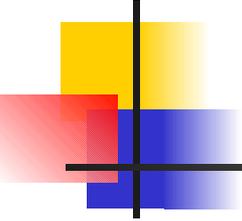
Ethical concepts involved

- Respect for human life, dignity, integrity
- Ascription of moral and legal responsibility
 - EU Charter of the Fundamental Rights of the European Union
 - Art. 1 (human dignity): human dignity is inviolable. It must be respected and protected.
 - Art. 3: right to the (physical and mental) integrity of the person.
 - Art. 6: right to liberty and security.

Additional (and belittled) Motives for Wiener's Concern Today



- **Unpredictability of machine behaviours**
 - **Developments of Machine Learning**
 - **Theoretical Computer Science**
 - **General Philosophy of Science**
 - **AI and Robotics state-of-art surveys**

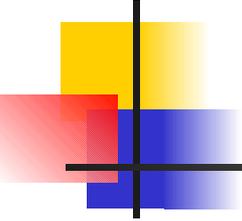


Roboethics and Machine Learning

- **Learning from examples**

- **Inductive hypothesis (Hume's problem):** Any learning hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over unobserved examples.

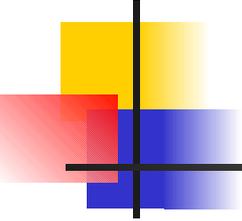
Is this inductive hypothesis justified in the case of machines?



Roboethics and Machine Learning

No “Superhuman” learning

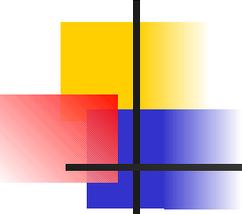
- Fallible inductive bias (theoretical expectations about the environment)
- Fallible choice of training sets (driven by hypotheses about representativeness of training sets)



Roboethics and Machine Learning

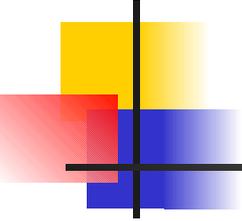
What can be reliably and efficiently learnt?

- Limited classes of concepts can be learnt with arbitrarily high probability from randomly drawn training examples using a reasonable amount of computational resources (PAC-learning)
- Computational complexity impediments to (neural) learning



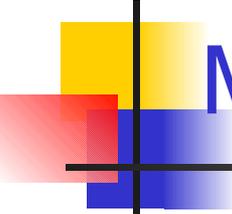
Roboethics and Philosophy of Science

- Regularities/Laws hold “ceteris paribus”, that is, when no disturbing factors arise
- Can one exhaustively list disturbing factors?
- ...and especially those arising “in the wild”?
 - AIBO’s discrimination between fires and red patches
 - Aegis radar system on USS Vincennes



Roboethics and Philosophy of Science

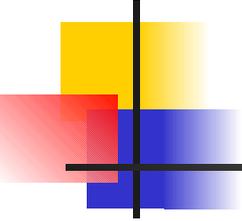
- Mechanisms are supposed to work in *normal* situations, and are usually tested in selective experimental settings or by computer simulations based on theoretical models of task environments



Misconceptions at war

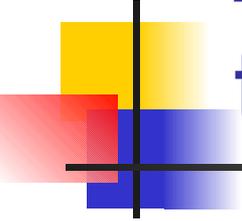
The American military is working on a new generation of soldiers, far different from the army it has. "They don't get hungry," said Gordon Johnson of the Joint Forces Command at the Pentagon. "They're not afraid. They don't forget their orders. They don't care if the guy next to them has just been shot. **Will they do a better job than humans? Yes.**" The robot soldier is coming.

Front-page article, NYT 16 feb. 2005 T. Weiner



Roboethics and state-of-the-art robotics research

- Open context interpretation
 - Recognizing surrender gestures
 - Telling bystanders from foes

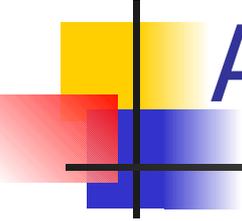


Are these problems properly taken into account?

- Saving life of soldiers
 - Acceptance of side effects, collateral damages?

VS

- Violating respect for human life
- Skirting moral or legal responsibility



A multiple-actor enterprise

Robo-ethics challenges: exchange of specialized knowledge, multidisciplinary dialogues & reflections

- Ordinary citizens
- Legal experts
- Computer scientists
- Sociologists
- Roboticists
- Philosophers
- Theologians
- Journalists
-