Preliminary steps for evaluating the impact of AI and robotic technologies

Francesco Amigoni and Viola Schiaffonati
Artificial Intelligence and Robotics Laboratory
Politecnico di Milano

Nature 2016 October, 538: 311-313

Chicago police use algorithmic systems to predict which people are most likely to be involved in a shooting, but they have proved largely ineffective.

There is a blind spot in AI research

Fears about the future impacts of artificial intelligence are distracting researchers from the real risks of deployed systems, argue **Kate Crawford** and **Ryan Calo**.

A social-system analysis

- Relatively untested AI systems introduced without a rigorous analysis about their social, cultural, and political impact
- Social-system analysis to overcome the limitations of existing approaches
  - Compliance, values in design, thought experiments
- Engaging with social impacts at every stage
  - Conception, design, deployment, regulation

Testing AI systems rigorously

- Necessity of an integrated analysis (epistemology + ethics) for a rigorous evaluation
  - Testing how a system works to evaluate its impact
  - Not only ethical consequences, but radical epistemological shifts impacting on these consequences
- Focus on autonomous robotics as a case study
  - Robot systems with the ability to operate without continuous human intervention in places hardly accessible by humans or in cooperation with humans in common environments

My plan for today

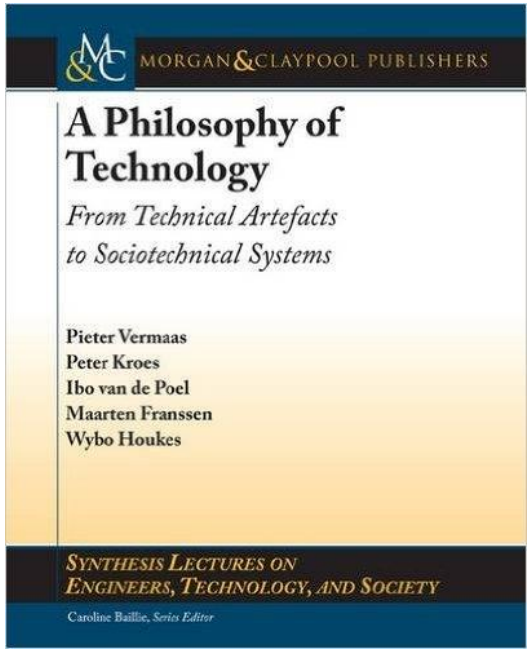
- Experiments in autonomous robotics
- Explorative experiments
- New technologies as social experiments
- Crisis of the traditional notion of direct control
  - Meaningful human control
  - Special testing zones
- Technical, scientific, ethical and societal challenges of responsible innovation

Experimental trends

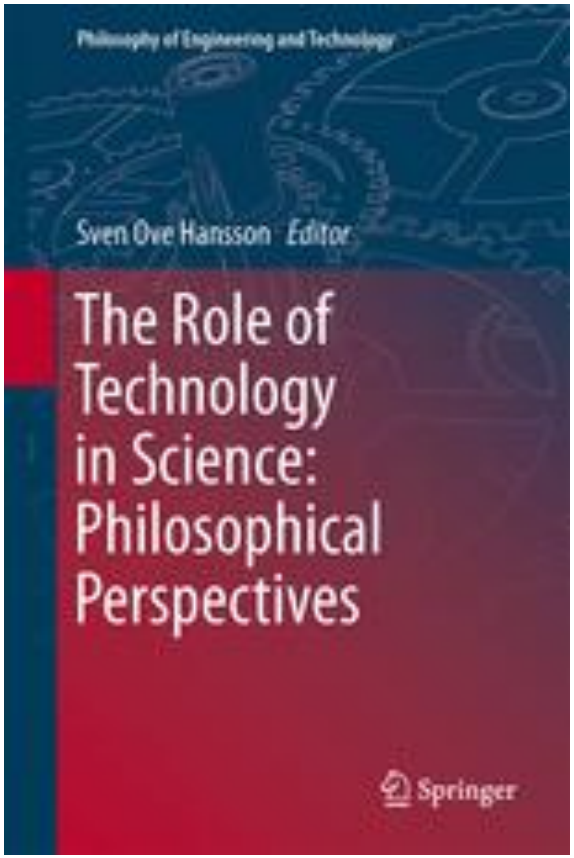
- Two different tendencies in autonomous robotics (Amigoni et al. 2014)
  - Principles of traditional experimental method (comparison, reproducibility, repeatability, generalization, justification, …) as inspiration
    - Development of comparable implementations using the same code (*comparison*)
    - Public distribution of code and/or data sets (*reproducibility*)
  - Rigorous approaches not yet fully part of the current research practice
    - Limited use of settings relative to different environments (*generalization*)
    - Rare reports of anomalies and negative results (*justification*)

Widening the framework

- Not simply adapting conceptual tools already adopted in the natural sciences (e.g., epistemic experiment)
- But proposing a novel notion of experiment fitting with the engineering sciences
  - Robotic systems as technical artefacts with a technical function and use plan designed and made by humans
  - Experiments carried out to check whether these artefacts meet the desired specifications via their technological production



A different type of experiment



"*An **experiment** is **directly action-guiding** if and only if it satisfies the following two criteria: (1) The **outcome** looked for should consist in the attainment of **some desired goal of human action**, (2) and the **interventions** studied should be potential **candidates** for being performed in a **non experimental setting** in order to achieve that goal. These criteria are satisfied for instance in a **clinical trial**.* […] *In contrast, an epistemic experiment aims at providing us with **information** about the **workings** of the **world we live in**.*"

(Hansson 2015)

Analgesics and autonomous robots

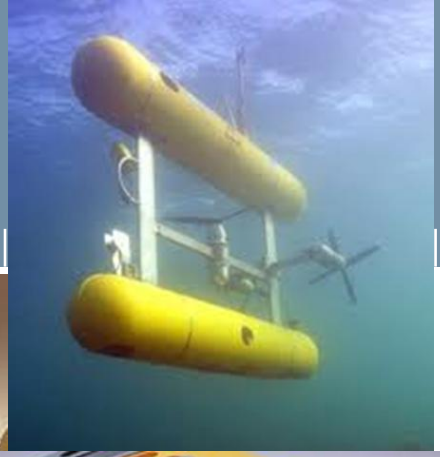
- Technological forms of experimentation driven by practical needs
- Clinical trial of an analgesic
  - Pain reduction (outcome looked for)
  - Treatment to be administered to patients (intervention)
- Systematic experimentation on an autonomous robot assisting an elderly person in her home
  - Proper interaction of the robot with the person (outcome looked for)
  - Careful tuning of the abilities of the robot to achieve the goal (intervention)

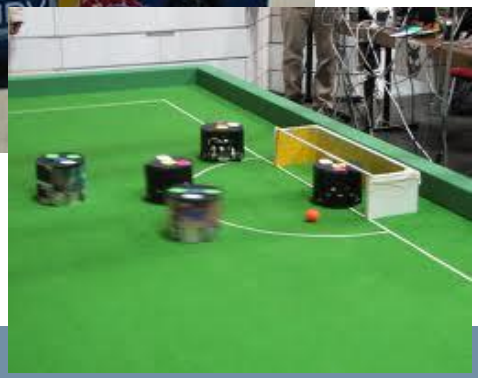
Explorative experiments

- Explorative experiments as forms of directly action-guiding experiments (Schiaffonati 2016)
  - Testing technical artefacts
  - Probing iteratively the possibilities and limits of the intervention (not testing a general theory)
  - Eliminating the distinction between designers and experimenters
  - Controlling the experimental factors not from the beginning, but after the insertion of artefacts into their environment

Forms of explorations

- Different forms of exploration in autonomous robotics (Amigoni and Schiaffonati 2016)
  - Investigating the relationship between values of parameters and behaviors of robot systems
  - Confirming expectations or hypotheses (in particular when inserting robots in their operating environments)
  - Getting insights on the behavior of the robot systems
  - Assessing the generality of robot systems





New technologies as social experiments

- New technologies having a serious impact on society
  - Impact largely unknown and very hard to predict
- New technologies introduced into society as a social experiment (NTaSE)

- Learning-by-experimentation

*"We might now position learning-by-experimentation between learning-by-doing and learning-by-anticipation. It is similar to **learning-by-doing** in that it takes place during the **actual introduction** of a **technology in society**. Still, it is **more anticipatory** than regular learning-by-doing because it takes place in a research setting with at least the **partial aim to learn** something. Ideally then, learning-by-experimentation allows for **learning things** that **cannot be learned by anticipation** and at the same time is **less costly than learning-by-doing**."*

(van de Poel forthcoming)

Explorative experiments and NTaSE

- Explorative experiments as social experiments
  - Necessity of introducing robotic systems into their environment to test them
  - Introduction of autonomous robotics technologies with large uncertainties, unknown and indeterminacies
  - Difficulties in modeling the interaction of the autonomous robotic system with the environment
  - Different notion of experimental control (a posteriori)

An issue of control

- Practitioners as experimenters in explorative experiments in autonomous robotics
- Creating and testing technical artefacts
- Loosing independence of the experimenter prescribed in the classical experimental protocol

*"In the traditional experimental protocol in natural sciences a researcher should be an **outsider** to the phenomenon to be explained — but it is uncertain how much a computer scientist can be an outsider to a **phenomenon** he or she has **created**"*
(Tedre 2011)

Crisis of the traditional control paradigm

Sci Eng Ethics (2016) 22:633–645
DOI 10.1007/s11948-015-9634-4

CrossMark

ORIGINAL PAPER

Experiments on Socio-Technical Systems: The Problem of Control

Peter Kroes

"*Because the **conditions** are **controlled**, **experiments** may be **replicated** in order to test the "internal" validity of the outcomes.* […] *The **experimenter** somehow is able **to intervene** in the system (s)he is experimenting on. The notion of intervention has a clear meaning: the experimentalist **is not part of the system** on which the experiment is conducted.* […] *In other words, the **experimentalist** operates from a **center of command** and **control outside** the experimental system. I will refer to these ideas as the **traditional control paradigm for experiments**. In my opinion, the notions of an intervention and of a center of command and control **become problematic** in the case of the **new technologies** that are **treated as social experiments** or involve complex socio-technical systems."*

(Kroes 2016)

Away from the ideal of direct control

- Meaningful human control (MHC)
  - Weapon systems

*"**Humans** not computers and their algorithms should remain **ultimately morally responsible** for potential **lethal operations** "*

(Horowitz and Scharre 2015)

  - Self-driving cars

*"Meaningful human control is required to make sure that every time that a **potentially wrong (criminal) action** is performed, for instance an injury or killing due to the reckless or negligent behavior of a driving system, some **human agent** is **morally** and **legally liable**"*

(Santoni de Sio 2016)

MHC (Santoni de Sio 2016)

ETHICS AND SELF-DRIVING CARS
A WHITE PAPER ON RESPONSIBLE INNOVATION IN AUTOMATED DRIVING SYSTEMS

Filippo Santoni de Sio*

- MHC different than 'being in the loop' and 'controlling'
- Meaningful not meaning direct
- In principle compatible with high automation
- MHC = system (robot + technical infrastructure + social/legal institutions) designed to respond to the relevant moral and legal reasons of the human designers and users

Special testing zones

- "Special zones" for testing robotic technologies created in some Japanese cities (Santoni de Sio 2016)
- Controlled space within real society with
  - Test robots already proven to be safe in laboratory
  - Special precautions (specific signs, specific insurance schemes) for those entering the zone
- Responsible innovation
  - Boosting highly autonomous robots while guaranteeing safety and human responsibility
  - Helping policy-makers to develop well-informed policies and legal regulations for introduction and use of robots

Exploration and responsibility

- Technological infrastructure and socio-political context too complex and risky for predicting behavior, side-effects, challenges of autonomous robots
  - Impossibility of learning-by-anticipation and difficulties of learning-by-doing
- Experimenting on autonomous robotic systems within the framework of NTaSE
  - Explorative experiments and learning-by-exploration (MHC, special testing zones)

A larger picture

- Not only technical challenges but also
  - Scientific: how to experiment rigorously but efficiently with robots and humans and robots (increasing importance of the human factor)
  - Ethical: how to promote responsible innovation and to anticipate and possibly solve values conflicts and tensions by design
  - Societal: how to design new licensing, training and liability schemes
- Engineers, designers, technologists, policy makers, philosophers working together from the beginning

Thank you for your attention

References

- Amigoni, F., Schiaffonati, V. (2016). "Explorative Experiments in Autonomous Robotics" in L. Magnani, C. Casadio (eds.), *Model-Based Reasoning in Science and Technology*, Springer, 585-599.
- Hansson, S.O. (2015). "Experiments before Science? – What Science Learned from Technological Experiments", in Sven Ove Hansson (ed.) *The Role of Technology in Science*, Springer.
- Horowitz, M., Scharre, P. (2015) "Meaningful Human Control in Weapon Systems: A Primer", *Center for a New American Security* http://www.cnas.org/human-control-in-weapon-systems.
- Kroes, P. (2016). "Experiments on Socio-Technical Systems: The Problem of Control", *Science and Engineering Ethics*, 22(3), 633–645.
- Santoni De Sio, F. (2016). "Ethics and Self-Driving Cars: A White Paper on responsible Innovation in Automated Driving Systems", *Dutch Ministry of Infrastructure and Environment*.
- Crawford, K., Calo, R. (2016) "There is a blind spot in AI research", Nature, 538, October 2016, 311-313.
- Schiaffonati, V. (2016). "Stretching the Traditional Notion of Experiment in Computing: Explorative Experiments", *Science and Engineering Ethics*, 22(3), 647-665.
- Tedre, M. (2011). "Computing as a Science: A Survey of Computing Viewpoints", *Minds and Machines*, 21, 361-387.
- Van de Poel, I. (forthcoming). "Society as a Laboratory to Experiment with New Technologies" in E. Stokes, D. Bowman and A. Rip (eds.) *Embedding and Governing New Technologies*. Singapore: Pan Stanford Publishing.