



UNIVERSITÀ DI PISA

# Social and Ethical Issues in Information Technology

Prof. Vincenzo Gervasi  
Fabio Fossa, Ph.D.

A.Y. 2019/20

Master Program in Computer Science

Artificial Intelligence curriculum

Diletta Goglia

*Notes from lectures*

## SYLLABUS

First part: "Theoretical" issues (lectures 1-6)

- Artificial Agents: Purposeful and Intelligent Behaviour.
- Singularity and Superintelligence.
- Agency, Free Will, Consciousness.
- Responsibility.

Second part: Ethical and Social Issues (lectures 7-16)

1. AI Ethics/Roboethics.
2. Machine Ethics.
3. Machine Learning, Neural Networks, Big Data.
4. Anthropomorphism, HCI-HRI.

Third part: (lectures 17-20)

Practical cases

- Autonomous Weapon Systems
- Sex Robots

...and students' presentations.

Possible topics:

- Self-Driving Cars.
- Expert Systems: COMPAS, Watson...
- HRI: Hanson's Sophia, Jibo...
- Microsoft's Tay.
- Machine Art: TheNextRembrandt, Obvious Art, Shimon...
- ...and so on

## CONTENTS

SYLLABUS.....	2
I. First part: "Theoretical" issues.....	6
LECTURE 0: INTRODUCTION.....	6
LECTURE 1: ETHICS, TECHNOLOGY AND PURPOSEFUL BEHAVIOUR .....	6
What is technology?.....	6
What is AI? .....	9
Summary.....	10
Essential characters of AI .....	10
LECTURE 2: ETHICS, TECHNOLOGY AND INTELLIGENT BEHAVIOUR .....	14
On the previous lesson .....	14
Summary.....	15
Conclusions .....	15
Questions.....	15
LECTURE 3: ETHICS AND SUPERINTELLIGENCE .....	15
Intelligence explosion.....	16
The Singularity .....	16
Superintelligence & ethics .....	18
LECTURE 4: DEBATE ON SUPERINTELLIGENCE.....	21
Deep disagreement.....	21
Relevance .....	21
Who says yes .....	21
Who says no.....	22
What really matters.....	22
LECTURE 5: HUMAN AND ARTIFICIAL INTELLIGENCE .....	24
Recap of all previous lectures .....	24
Human vs. Artificial Intelligence .....	25
Recap of the two main problems .....	28
Syntax and Semantics.....	28
The Chinese Room .....	29
LECTURE 6: DEBATE ON SEARLE'S CHINESE ROOM .....	32
Objections.....	34
Summary.....	36
II. Second part: Ethical and Social Issues.....	37
LECTURE 7: ARTIFICIAL AGENCY AND RESPONSIBILITY (PART I) .....	37
Agency and Meaning.....	37
Agency and Responsibility .....	37
Responsibility .....	38
AI and Responsibility .....	40
Conclusions .....	41
LECTURE 8: DEBATE ON ARTIFICIAL AGENCY AND RESPONSIBILITY (PART I) .....	42
Recap: AI and responsibility .....	42
Analogical Thinking.....	43
LECTURE 9: ARTIFICIAL AGENCY AND RESPONSIBILITY (PART II) .....	45

## SYLLABUS

Summary.....	45
Responsible Artificial Agents .....	45
Responsibility Gap .....	47
Reactions to Matthias' Arguments .....	48
Summary.....	50
Combining or discharging responsibility .....	50
LECTURE 10: DEBATE ON ARTIFICIAL AGENCY AND RESPONSIBILITY (PART II) .....	51
Who's responsible .....	51
Punishment.....	52
LECTURE 11: ANTHROPOMORPHISM.....	55
Terms, definitions and etymology.....	55
How design can exploit anthropomorphism .....	56
Research fields.....	57
Negative aspects .....	58
Conclusions .....	58
Reflections .....	58
LECTURE 12: DEBATE ON ANTHROPOMORPHISM .....	59
Summary of previous lecture.....	59
Social Robots .....	59
Design Issues.....	60
Social & Ethical Issues .....	60
Summary and conclusions.....	63
Open issues:.....	63
LECTURE 13: AI ETHICS .....	64
Summary of previous lessons .....	64
The Ethics of AI.....	64
Applied Ethics.....	64
Technology and Ethics .....	65
Design Issues.....	65
(Not just) Design Issues .....	65
Aims of AI Ethics:.....	66
Important initiatives .....	66
Important approaches.....	66
AI Ethics principles.....	66
The Ethics of AI – Issues .....	67
AI Ethics Epic Fail .....	68
Ethics Washing.....	68
LECTURE 14: MACHINE ETHICS.....	68
Literature reference.....	68
AI Ethics vs. Machine Ethics.....	69
Artificial Moral Agency.....	69
The Machine Ethics Project.....	70
Machine Ethics Issues.....	70
LECTURE 15: NEURAL NETWORKS AND MACHINE LEARNING .....	71
LECTURE 16: BIG DATA ETHICS .....	72
Bias.....	72

## SYLLABUS

Delegation .....	74
Black boxes .....	75
Other issues .....	75
III. Third part: practical cases and student presentations .....	76
LECTURE 17: SEX ROBOTS .....	76
LECTURE 18: SELF DRIVING CARS .....	77
LECTURE 19: AUTONOMOUS WEAPON SYSTEMS .....	77
Definitions .....	78
Approaches to autonomy of weapons .....	78
Ethical frameworks .....	80
Legal Frameworks .....	80
Policy & Social Initiatives: .....	80
Examples of AWS deployed today: .....	80
LECTURE 20: DEBATE ON AUTONOMOUS WEAPON SYSTEMS .....	81
Issues .....	81
Utility of AWS .....	81
Issue 1: Technical obstacles .....	81
Issue 2: Risks .....	82
Issue 3: Accountability .....	82
Issue 4: Human Dignity .....	82
Conclusion: The 'Martens Clause' .....	83
LECTURE 21: MACHINE ART .....	83

## I. FIRST PART: “THEORETICAL” ISSUES

---

### LECTURE 0: INTRODUCTION

**EXAM:** paper/essay about a topic of the course (personal opinion on a theme).

Course methodology: focus on questions, not answers → Philosophical point of view but also programmers point of view.

Aims of the course: improve practical skills.

- Critical analysis of notions, claims, arguments and cases.
- Proposing/discussing different viewpoints clearly and thoroughly.
- Debating with peers, asking the “right” questions, finding the weak spots.

*Technological applications are everywhere. AI technologies are (probably) going to revolutionize our life. Their effects and impacts are already extremely significant and will be disruptive in the next decades. Therefore, matters of design and use must be assessed also from a social, political, and ethical point of view.*

*You (computer scientists) are and will be those who design, program, build AI systems and applications. You know the nitty-gritty on computer technology and AI. As experts, your opinions on these matters too will be held in great consideration. It is also your responsibility to ask these questions and to learn how to properly deal with them!*

*If we wish the use of AI in future society to be well- informed, fair, and good, then computer science, ethics, and social reflection must go hand in hand.*

**Antropomorphism:** attributing human behavior and human vocabulary to things that are not human.

---

### LECTURE 1: ETHICS, TECHNOLOGY AND PURPOSEFUL BEHAVIOUR

*Artificial Intelligence: why ethical and social issues—beyond technical issues?*

Ethical and social dimensions are embedded in technology, we don't have to add them into it.

This applies not only to AI but to technology in general (let's take a step back in time).

*What is AI? AI is a technological product. What is a technological product? What is technology?*

#### WHAT IS TECHNOLOGY?

La **tecnologia** è risultato dell'attività umana, è tutto ciò che produciamo. È l'insieme di tutti gli artefatti tecnologici. Ma questa definizione lascia fuori **la componente umana** che è **molto importante**. Dal punto di vista filosofico diamo una definizione che si incentri sulla attività umana: questi artefatti tecnologici sono il prodotto dell'attività umana.

---

### 1. FORM OF HUMAN ACTIVITY:

«For procuring ends and procure and utilize the means to them is a human activity. The manufacture and utilization of equipment, tools, and machines, the manufactured and used things all belong to technology. The whole complex of these contrivances is technology. Technology itself is a contrivance—in Latin, an instrumentum»

(M. Heidegger, *The Question Concerning Technology*)

---

Questa definizione più filosofica e meno “strettamente scientifica” comprende riferimento non solo agli oggetti ma **l'azione di produrli**. Consideriamo quindi ogni attività umana che c'è dietro.

Tutti gli artefatti tecnologici sono strumenti, quindi li produciamo perché riteniamo che ci siano utili per qualche scopo. Ma è anche vero che non possiamo ridurre la tecnologia ai semplici artefatti tecnici “utili” ma dobbiamo considerare tutti gli spetti sociali del produrli, commercializzarli, ecc..

---

### 2. TECHNOLOGY/ARTEFACTS DISTINCTION:

«A common way of thinking about technology —perhaps the layperson's way—is to think that it is physical or material objects (artifacts) [...]. Philosophers of technology and recent literature from the field of Science and Technology Studies (STS) have pointed to the misleading nature of this view of technology. Technology is a combination of artifacts, social practices, social relationships, and systems of knowledge».

(D. Johnson, *Computer Systems*)

---

### 3. TECHNOLOGY, ARTEFACTS, SOCIAL CONTEXT:

«Artifacts (the products of human contrivance) do not exist without systems of knowledge, social practices, and human relationships. Artifacts are made, adopted, distributed, used, and have meaning only in the context of human social activity».

(D. Johnson, *Computer Systems*)

---

**L'elemento sociale** non è qualcosa che noi aggiungiamo ma **è intrinseco nella tecnologia**.

Il goal, l'obiettivo finale che vogliamo raggiungere, è un elemento sociale, è qualcosa che noi creiamo allo stesso modo in cui creiamo un tool.

**Il significato che conferiamo ad uno strumento riguarda il modo in cui pensiamo di utilizzarlo.** Non possiamo decontestualizzarlo dal contesto sociale in cui il tool è stato ideato e creato. Il significato riguarda il modo in cui utilizziamo un oggetto. Se rimuoviamo l'elemento sociale ovvero decontestualizziamo l'artefatto, quest'ultimo diventa qualcosa di completamente diverso. Es. Immaginiamo una TV nel Settecento: le persone dell'epoca non conoscono questo strumento e quindi non sanno il suo significato: di conseguenza la useranno in un modo completamente diverso da quello pensato quando è stato inventato. Il contesto è importante: un tool nasce in un preciso momento del tempo per uno scopo necessario in quel momento: rimuovere la dimensione temporale e quindi sociale fa perdere il significato dell'oggetto stesso e quindi anche il motivo per cui è stato immaginato, progettato e creato. L'oggetto diventa qualcosa di completamente diverso perché le persone lo usano con scopi completamente diversi, conformi alla società e ai bisogni di quel momento.

Stesso artefatto (oggetto tecnologico) posso usarlo in modi diversi e dargli quindi significati diversi (es. coltello posso usarlo per mangiare o per uccidere). Ci sono norme sociali intorno ad un certo oggetto.

La stessa cosa vale se un oggetto è naturale e non uno strumento artificiale (es. pietra): **il modo in cui lo usiamo gli conferisce un significato sociale diverso.**

Non solo gli oggetti, ma **anche la scienza è *socially determined***. Il modo in cui una società è organizzata e la scienza si influenzano reciprocamente. La scienza è una tecnologia: più precisamente, è la condizione fondamentale per la tecnologia.

Dietro uno strumento prodotto c'è sempre un sistema di conoscenza.

---

#### 4. NON-NEUTRALITY OF TECHNOLOGY:

«technology is not neutral, [...] technologies do much more than simply achieve the goals for which they were instituted. technology does much more than realize the goal toward which it is put; it always helps to shape the context in which it functions, altering the actions of human beings and the relations between them and the environment».

(P.-P. Verbeek, *What things do*)

---

#### 5. COMPUTER SYSTEMS:

«Artifacts come into being through social activity, are distributed and used by human beings as part of social activity, and have meaning only in particular contexts in which they are recognized and used. [...] So it is with computers and computer systems. This is as true of current computer systems as it will be of future computer systems. No matter how independently, automatically, or interactively computer systems of the future behave, they will be the products (direct or indirect) of human behavior, human social institution, and human decision».

(D. Johnson, *Computer Systems*)

---

#### 6. ARTIFICIAL INTELLIGENCE:

«Computer programs, included those classified as Artificial Intelligence (AI), are purpose-built artefacts designed, commissioned and operated by human beings. [...] We develop artifacts to perform tasks for us, and while they may eliminate the need for various human labours, they do not eliminate the need or desire for us to live our lives».

(J. Bryson & P. Kime, *Just an Artifact*)

---

***Technology is not value-free, is not neutral. Is intimately connected to the social sphere.***

Dire che la tecnologia serve semplicemente per raggiungere un obiettivo è uno stratagemma per tenere l'etica al di fuori. Ma noi sappiamo che non è possibile tenerle separate perché sono intrinsecamente connesse.

APPROFONDIMENTI: Verbeek, filosofo STS olandese.

Non importa quanto un sistema sia autonomo, resta comunque un prodotto umano.

La **general AI** non ha uno scopo. Non sono *purpose-built artefact*, come invece lo sono molti prodotti di AI. Differenza tra *general* e ... AI è proprio la presenza o meno di uno scopo preciso da raggiungere. Prospettiva molto diversa che include ovviamente l'essere umano.



Qui nasce una riflessione. Le macchine che creiamo sono davvero senza scopo? Oppure il loro scopo è creare scopi per la ricerca?

**Emergent behaviour:** sistemi che iniziano a comportarsi in un modo non forzato, non programmato, non deciso a priori, spontaneo.

Nel dibattito sulla *general AI* è interessante il fatto che una macchina non costruita per uno scopo inizi a fare qualcosa per un *emergent behaviour*. Nessuno gli ha fornito un goal da raggiungere, non ha uno scopo a priori, ma inizia comunque a perseguirne uno: si comporta “spontaneamente” in un certo modo. Emerge un comportamento non progettato, non programmato.

APPROFONDIRE significato parola “spontaneo”

Questo è osservato in generale per i sistemi, non solo per i prodotti di *general AI*.

Un sistema come per esempio la natura *exhibit an emergent behaviour*. Essa non ha uno scopo prefissato “dall’alto” ma emergono dei comportamenti ben precisi. Moltissimi processi naturali hanno degli scopi ben precisi, ma questi emergono da soli, nessuno li ha predeterminati o programmati prima → dibattito con religione.

Spesso associamo *purposeful behaviour* con *intentionality*.

APPROFONDIRE differenza tra questi due concetti.

### SUMMARY

- **Technology** = combination of artefacts, social practices and relationships, systems of knowledge.
- **Technological products** = tools, i.e. artefacts built to attain given goals, to carry out tasks (for us).
- **Ethical/Social relevance:** artefacts are embedded in the social context and concur to shaping it > essential connection between technology, society, and ethics.
- **AI is a technological product:** «in creating and using intelligent artifacts we do need to consider ethical and social dangers, but in no greater sense than we should with more conventional technologies». (J. Bryson & P. Kime, *Just an Artifact*).

### WHAT IS AI?

Example: what differentiates ...

Hammer → Washing Machine → AI product

	Hammer	Washing Machine	AI
<b>Autonomy</b>	None	Medium	High
<b>Delegation</b>	None	Medium	High
<b>Substitution</b>	None	Medium	High
<b>Supervision/intervention</b>	High	Medium	Low

Aumenta l'**autonomia** delle macchine e quindi diminuisce l'intervento umano. *Delegation and substitution are increasing.*

Aumenta il **tempo libero** per gli esseri umani → nessun giudizio etico su questo punto in questo momento. Delegando tutte le mie scelte alla macchina ho più tempo libero ma meno **potere decisionale**.

*Less direct intervention.*

Il risultato è che sempre più persone hanno meno potere decisionale, il quale sta invece aumentando nelle mani di pochi (coloro che progettano queste tecnologie).

Es. Facebook, Netflix che suggeriscono (o addirittura decidono) per noi cosa possiamo o non possiamo vedere.

Aumenta anche la **probabilità di rischio**. È sempre più problematica e grave l'ipotesi che accada un errore. Problema della **scalabilità**.

Infatti, più deleghiamo più è difficile stabilire di chi è la **responsabilità** di eventuali errori. **Perdiamo controllo diretto** sulla tecnologia e su ciò che essa fa. Abbiamo solo controllo generico.

*Control and responsibility are strictly connected.*



C'è un sistema gerarchico di responsabilità e deleghe (es. esercito). Se viene violata una istruzione precisa (c'è una contraddizione) la responsabilità è a livello basso (es. della macchina) altrimenti se non ho dato istruzioni precise la responsabilità sale nella gerarchia di chi ha delegato (es. io che sono stato ambiguo nelle specifiche).

Aumenta anche la comprensione da parte della macchina di ciò che essa stessa fa.

Attenzione ai termini comprensione e *consciousness* da parte delle macchine.

*Context/environmental awareness* è più corretto di *consciousness* nel caso della lavatrice.

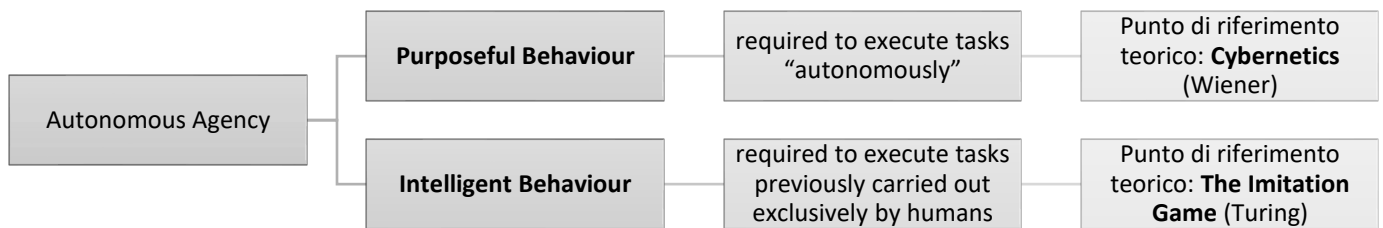
### SUMMARY:

	<b>Autonomia</b> macchine
	<b>Delegazione</b> /sostituzione
	<b>Tempo libero</b> per l'essere umano
	Gravità di un eventuale errore
	<b>RESPONSABILITA'</b>
	Intervento umano
	Potere decisionale
	<b>Controllo</b> /intervento <b>diretto</b>

## ESSENTIAL CHARACTERS OF AI

*AI is a technological product (tool) that executes tasks which used to require human labour in an "autonomous" way, i.e., without requiring human constant supervision or intervention.*

→ **Essential Characters** = Autonomous Executer of Functions, Artificial Delegatee, or Autonomous Agent (!)



## ARTIFICIAL PURPOSEFUL BEHAVIOUR (CYBERNETICS)

*Can machines have purposes? Or only humans can?*

N. Wiener, A. Rosenblueth, J. Bigelow (1943), *Behavior, Purpose and Teleology*, Philosophy of Science 10, 18-24.

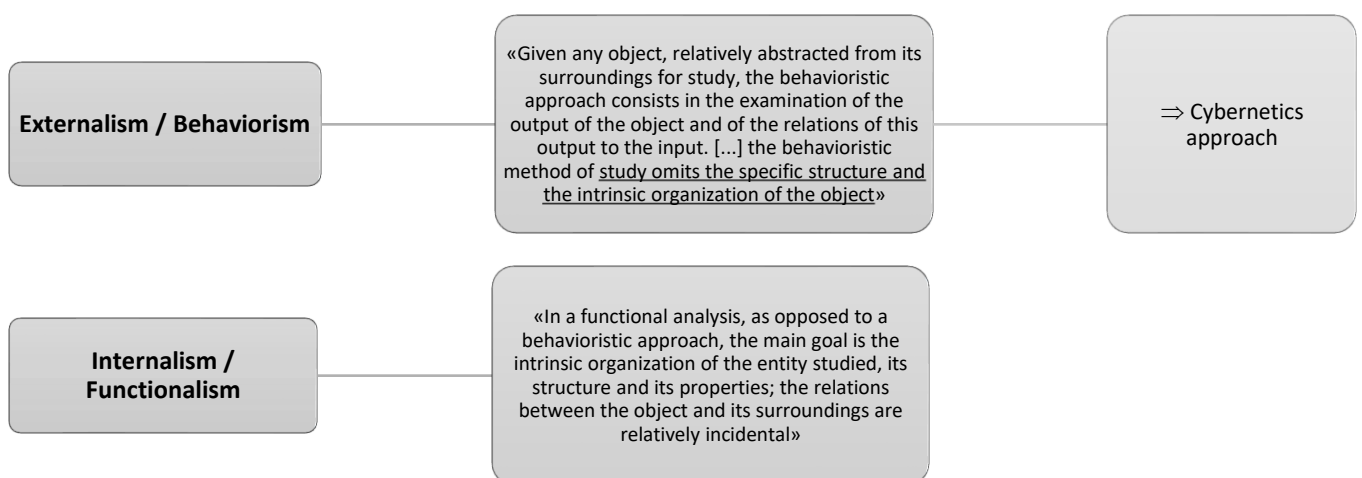
Aims: «This essay has two goals. The first is to define the behavioristic study of natural events and to classify behaviour. The second is to stress the importance of the concept of purpose».

Etimologia del termine “**teleologia**” include il concetto di **scopo** (greco “telòs”).

*Define a behaviour that tends to a specific purpose.*

**Wiener** padre della **cybernetica** → prima tecnologia sviluppata dalla cybernetica: Torpedo. è un esempio di tecnologia *purposeful behaviour*.

## METHODOLOGY



Two internally different entities that produce the same output given the same input are considered equal. This is the externalist conception of cybernetics.

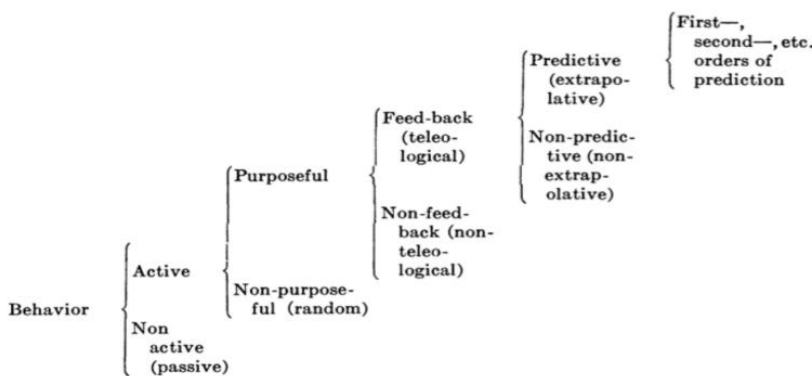
Two methodologies that are not only different but completely opposite (they must always be kept distinct).

**Machine-Organism Analogy:** both organic and technological behaviour!

«a uniform behavioristic analysis is applicable to both machines and living organisms, regardless of the complexity of the behavior. [...] The methods of study for the two groups are at present similar. [...] The broad classes of behavior are the same in machines and in living organisms. [...] While the behavioristic analysis of machines and living organisms is largely uniform, their functional study reveals deep differences».

→ Consapevolezza di star trascurando degli elementi molto rilevanti ma non ci interessano perché stiamo volutamente adottando un approccio externalista.

Siamo consapevoli del fatto che ci possono essere differenze profonde ma finché restiamo in questa metodologia di approccio non ci interessa se internamente le entità sono diverse quindi NON CI INTERESSA SE È UN MACCHINA O UN UMANO.



## DEFINITIONS

- **Behavior** = any change of an entity with respect to its surroundings.
- **Purposeful** = *The term purposeful is meant to denote that the act or behavior may be interpreted as direct to the attainment of a goal—i.e., to a final condition in which the behaving object reaches a definite correlation in time or in space with respect to another object or event.*

Definition very focused on the machine. I wouldn't use the same terms on human behavior. No reference to the internal dimension of human or machine but only scientific language.

- **Purposeless** = is that which is not interpreted as directed to a goal → = random.

EXAMPLES:

Purposeless Machines? Roulettes, Clocks, Guns:

«The view has often been expressed that all machines are purposeful. This view is untenable. First may be mentioned mechanical devices such as a roulette, designed precisely for purposelessness. Then may be considered devices such as a clock, designed, it is true, with a purpose, but having a performance which, although orderly, is not purposeful—i.e., there is no specific final condition toward which the movement of the clock strives. Similarly, although a gun may be used for a definite purpose, the attainment of a goal is not intrinsic to the performance of the gun: random shooting can be made, deliberately purposeless. Some machines, on the other hand, are intrinsically purposeful. A torpedo with a target mechanism is an example».

Methodology	Roulette	Clock	Gun
<b>Behaviorism:</b> <i>Definite correlation with another object/event</i>	No: the final position of the ball is indefinite.	There is no final condition.	Only when it hits the target? (Why not: when the bullet is successfully ejected?)
<b>Functionalism:</b> <i>Intentional purpose</i>	Play the game: roulette.	Keep track of time.	Shooting things (either living or not)

The two opposite methodologies become problematic in explaining the purpose of the gun example: the dimension of intentionality cannot be omitted: the human component is present.

## NEGATIVE FEEDBACK

Teleological Behaviour requires Negative Feedback:

«The expression feed-back [...] in a broad sense may denote that some of the output energy of an apparatus or machine is returned as input. [...] Positive feed-back adds to the input signals, it does not correct them. The term feed-back is also employed in a more restricted sense to signify that the behavior of an object is controlled by the margin of error at which the object stands at the given time with reference to a relatively specific goal. The feed-back is therefore negative, that is, the signals from the goal are used to restrict outputs which would otherwise go beyond the goal».

**Control** is the key concept here. Cybernetic point of view: we can implement teleology in a machine, or give it a purposeful behavior.

Purposeful Behaviour and Representation:

«All purposeful behavior may be considered to require negative feedback. If a goal is to be attained, some signals from the goal are necessary at some time to direct the behavior».

Requirements:

- representation of the purpose
- representation of the current state *vis-à-vis* the goal
- computation of future state towards the goal and corrections.

Technological science of purposeful behaviour based on negative feedback processes → **Cybernetics** (*kubernetes*, ancient Greek for “pilot”): technological science of control and communication in the animal and in the machine.

---

## LECTURE 2: ETHICS, TECHNOLOGY AND INTELLIGENT BEHAVIOUR

### ON THE PREVIOUS LESSON...

- Essential connection between technology and social context.
- AI **autonomous** tool → problems of *use* + problems of *impact*.
- **Artificial agents** = artificial delegates → more free time, less direct control.
- Basic characters of artificial agency: **purposeful and intelligent behaviour**.
- Purposeful behaviour and Cybernetics:
  - Behaviourism-Externalist approach → input-output correlation.
  - Organism-Machine analogy → rewarding and risky.
  - Purposefulness: ability to reach a definite correlation in time or space with another event or object → issues and difficulties.

Previous lesson: AI described as a mix of **purposeful** and **intelligent** behavior. We focused on the first one, today we analyze the second one.

### A. Turing - [Computing machinery and intelligence](#) - (PRINTED)

Attenzione alla semantica di I/O. *Issue: length of life of a computation* → in long running systems è difficile stabilire output dato un input. Gli input condizionano il comportamento in maniera molto protratta nel tempo. Comportamento = tutti input ricevuti dall’inizio (può essere passato moltissimo tempo). Quindi la semantica di input-output in molti sistemi è totalmente inappropriata.

Quando definiamo un goal esso non ha senso se non definiamo anche **vincoli**. Sono più difficili da specificare ancora di più del goal. Come posso specificare tutti i vincoli possibili e immaginabili senza trasferire la mia visione del mondo alla macchina?

J. McCarthy, M. Minsky, N. Rochester, C. Shannon (1955), *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, AI Magazine, 27, 4, 2006, 12-14.

«The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves».

What happened to the Imitation Game framework?

→ here we have imitation game without any boundary.

## SUMMARY:

- Machine **intelligence** = machine behaviour that mimic/simulate human intelligent behaviour (whatever it is), with the goal of becoming undistinguishable from it.
- **Programming** = writing instructions according to the logic of if/then or cause-effect (**determinism**). To write instructions, the end to be accomplished must be given (game: every player has an end). Potential situations must be foreseen, unless a method to automatically deal with unexpected situations is found.
- The end must be given, the means to its attainment must be automatized. Intelligence = ability to attain ends autonomously.
- Under some (very) specific respect, human and machine intelligence are continuous or homogeneous phenomena: the one can approximate and, to some extent, reproduce/substitute/overcome the other.

## CONCLUSIONS

«Machines can obviously have goals in the narrow sense of exhibiting goal-oriented behavior: the behavior of a heat-seeking missile is most economically explained as a goal to hit a target».

«Intelligence = ability to accomplish complex goals».

## QUESTIONS

What about setting purposes and values? Is “setting purposes & values to oneself” describable as a goal to accomplish, a function to execute?

Can ethical self-determination be simulated by means of AI? Why would we do it? Should we?

Is AI a useful tool to inquire into human intelligence? Is the human mind explainable as an AI system?

---

## LECTURE 3: ETHICS AND SUPERINTELLIGENCE

Uno dei punti che emergono dal paper di Turing: qualsiasi cosa sia l'intelligenza umana, in qualsiasi modo essa sia definita, posso comunque riprodurla sotto forma di intelligenza artificiale in modo indistinguibile per ottenere un goal. Questo è possibile grazie al programming.

**Intelligence** = ability to accomplish ends autonomously. → connection with purposeful behaviour

To define machine intelligence, a link to the dimension of life is necessary.

To define machine intelligence analogy with "natural" intelligence is required (connection with human but also animal behavior, plants, ...).

Organism-Machine Analogy = Animal-Machine Analogy = Human-Machine Analogy = **continuity** between organic and technological phenomena → Behaviourism/Externalism

In realtà ci sono discussioni aperte sul fatto che questo riferimento alla biologia sia o meno necessario per definire l'intelligenza. Solo gli organismi biologici sono realmente intelligenti?

Not only organic but also artificial systems exhibit intelligence (autonomous behaviour) → “‘alive’” ??

Complex systems (even if not biological) show this behavior that we can define as intelligent even if (yet) we do not understand it as a whole and deeply.

Debate open today. Are only biological systems intelligent, or are artificial complex systems intelligent?

Instead, we can take a non-mechanistic position where only biological systems are truly intelligent.

Turing propose not to try to reproduce directly an adult human mind because of its complexity and multi-layer structure.

**We start building an artificial mind from child-level (it's simpler) and then we increase the level of its complexity.**

How deep can it go? How much powerful? → **Superintelligence theory**

Progressive advancement of Machine Intelligence: pre-child level → child's level → adult's level → ?

### INTELLIGENCE EXPLOSION

Irvin J. Good, *Speculation concerning the first Ultraintelligent Machine*, 1965.

«Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control».

→ we no longer have an Imitation Game but everything can become an imitation of man.

**Any expression of human intelligence can be imitated by a machine. Can it perform even better?? This is now a speculation.** We follow this presupposition of continuity even if we don't know if it is correct. It doesn't matter if the discourse sounds crazy.

We used to know machine as tools... but if an ultraintelligent machine raise up this concept of “tool” is not applicable any more. **Move from notion of tool to notion of subject**: the machine is titled to have values, desires, ... It's necessary to keep it aligned with our values and purposes or we will face with an enemy (very stronger than us).

### THE SINGULARITY

This idea was first conceived in 1965. We find it again in 1993 in Venor Vinge, *The Coming Technological Singularity: How to Survive in the Post-Human Era*.

«we are on the edge of a change comparable to the rise of human life on Earth. The precise cause of this change is the imminent creation by technologies of entities with greater than human intelligence».



«What are the consequences of this event? When greater-than- human intelligence drives progress, that progress will be much more rapid. In fact, there seems no reason why progress itself would not involve the creation of still more intelligent entities—on a still shorter time scale».

Framework of ultraintelligent machines.

How game changes in this transformation?

Progress pertains to the whole system, not to the single (es. “umanità”).

We won't be able to control these technologies: human beings will not even understand how the machine reasons because it is extremely smarter than them. We will become obsolete. We not only lose control of the machines but also the only hope of controlling them. And this is where the term **singularity** appears.

### LIFE-MACHINE ANALOGY

«The best analogy that I see is with the evolutionary past. Animals can adapt to problems and make inventions, but often no faster than natural selection can do its work—the world acts as its own simulator in the case of natural selection. We humans have the ability to internalize the world and conduct “what ifs” in our heads; we can solve many problems thousands of times faster than natural selection. Now, by creating the means to execute those simulations at much higher speeds, we are entering a regime as radically different from our human past as we humans are from the lower animals».

### SINGULARITY → LOSS OF CONTROL

«From the human point of view this change will be a throwing away of all the previous rules, perhaps in a blink of an eye, an exponential runaway without any hope of control. [...] I think that it is fair to call this event a singularity. It is a point where our models must be discarded and a new reality rules».

«Good has captured the essence of the runaway, but does not pursue its most disturbing consequences. Any intelligent machine of the sort he [Good] describes would not be humankind's “tool”—any more than humans are the tools of rabbits or robins or chimpanzees».

More-than-human Intelligence → AI Singularity → from Delegation to Substitution → Human Obsolescence → We need ethical constraints for AI!

N.B.

In reasoning about superintelligence there are some **assumptions**:

- We assume that it is possible to reach a level of intelligence higher than human (we can believe it but not prove it).
- If this level exists, is it only potentially reachable or will we definitely get it?
- Beware of the term “imminent”: it can mean 5 years as 2 million (time framing problem).
- “Extrapolation of current condition” may not occur (it's not for sure that super intelligent machines will reproduce, and it's not for sure that they will do it on the same scale as human beings, it's not for sure that they will have a desire to live, ...). Not everything that could happen will actually happen.

When we will get the capability we must already have clear ideas on how to use these technologies. We need ethics in AI.

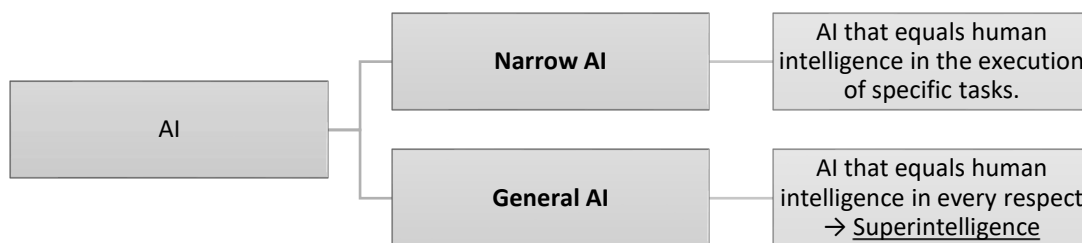
Eg. emergency of man over other animals was not a singularity at the time. There was no reasoning about when and how the human being would overcome the intelligence of the chimpanzee and dominate the world.

### SUPERINTELLIGENCE & ETHICS

N. Bostrom, *Ethical Issues in Advanced Artificial Intelligence*, 2003.

«A superintelligence is any intellect that vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills».

Here we have **not only scientific creativity but also ethical and social skills**.



The AI that Turing had in mind in the Imitation Game was Narrow AI.

### SUPERINTELLIGENCE: MORE THAN TOOLS

Consequences that follow the definition of superintelligence:

«Artificial Intellects are potentially autonomous agents. A superintelligence should not necessarily be conceptualized as a mere tool. [...] general superintelligence would be capable of independent initiative and of making its own plan, and may therefore be appropriately thought of as an autonomous agent.

Artificial intellects need not have humanlike motives. Humans are rarely willing slaves, but there is nothing implausible about the idea of a superintelligence having as its supergoal to serve humanity or some particular human, with no desire whatsoever to revolt or to “liberate” itself. It also seems perfectly possible to have a superintelligence whose sole goal is something completely arbitrary. [...] For better or worse, artificial intellects need not share our human motivational tendencies».

Bostrom says that probably these machines will have no desire to be free and to live as we have. Very controversial point.

Our conceptuality is lost. The purpose of these machines could be completely arbitrary. Why should they want the same things that we want, if they are thought to be smarter than us.

## VALUE-ALIGNMENT & ETHICAL MOTIVATION

1. «To the extent that ethics is a cognitive pursuit, a superintelligence could do it better than human thinkers. This means that questions about ethics, in so far as they have correct answers that can be arrived at by reasoning and weighting up of evidence, could be more accurately answer by a superintelligence than by humans».
2. «the setting up of initial conditions, and in particular the selection of a top-level goal for the superintelligence, is of utmost importance. (...) It seems that the best way to ensure that a superintelligence will have a beneficial impact on the world is to endow it with philanthropic values. Its top goal should be friendliness».

Value-Alignment is a problem of **design**, so it has sense if we are talking about **tools** (implement our values into machines); it is an exclusively **technical** problem.

BUT when we're dealing with "**superethical**" machine that are as autonomous as we are, the Value-Alignment changes. We don't want to implement our set of values in the machine if they're "**ethically better**" than us! It would be a constraint to their autonomy. We have to persuade these machines to behave with our values, not implement them into them.

We move from a design problem to a political and ethical problem.

Nella frase 1. si presuppone che queste macchine saranno più etiche di noi → non dobbiamo imporgli i nostri valori perché in relazione ai loro sarebbero inferiori.

La frase 2. contraddice la prima perché dice che dobbiamo insegnargli la filantropia. → they're not tools any more! non possiamo più "trattarli" così.

La presupposizione che più "disturba" è che ci siano delle risposte esatte in etica come fosse matematica.

## MORE ABOUT VALUE ALIGNMENT

In delegation or cooperation, value alignment = sharing of core ethical values and practical purposes between stakeholders.

- With "autonomous" tools: matter of programming. Alignment "by design" (vedi sopra).
- With self-determining entities: matter of motivation. Alignment "by agreement".  
We say that some people are "ethically aligned" if they have the same set of values and the same goals.
- With Superintelligence: agreement by design? What about the SuperAI's self-determination?

Big issue in keep together **control** and respect **autonomy** of these new superintelligent machines.

What will happen? So many possibilities!

Max Tegmark, *Life 3.0*, cap. 5: lists 12 potential outcomes.

1. Libertarian Utopia
2. Benevolent Dictator
3. Egalitarian Utopia
4. Gatekeeper
5. Protector God
6. Enslaved God
7. Conquerors
8. Descendants
9. Zookeepers
10. 1984
11. Reversion
12. Self-Destruction

### RELEVANCE

Does Superintelligence raise relevant concerns?

Nick Bostrom, *Ethical Issues in Advanced Artificial Intelligence* (2003).

YES (Better safe than sorry!): «there seems currently to be no good ground for assigning a negligible probability to the hypothesis that superintelligence will be created within the lifespan of some people alive today. Given the enormity of the consequences of superintelligence, it would make sense to give this prospect some serious consideration even if one thought that there were only a small probability of it happening any time soon».

David Chalmers, *The Singularity: a philosophical analysis* (2010).

«the singularity idea is clearly an important one. The argument for a singularity is one that we should take seriously. And the questions surrounding the singularity are of enormous practical and philosophical concern».

«**Practically:** If there is a singularity, it will be one of the most important events in the history of the planet. An intelligence explosion has enormous potential benefits: a cure for all known diseases, an end to poverty, extraordinary scientific advances, and much more. It also has enormous potential dangers: an end to the human race, an arms race of warring machines, the power to destroy the planet. So if there is even a small chance that there will be a singularity, we would do well to think about what forms it might take and whether there is anything we can do to influence the outcomes in a positive direction».

«**Philosophically:** The singularity raises many important philosophical questions. The basic argument for an intelligence explosion is philosophically interesting in itself, and forces us to think hard about the nature of intelligence and about the mental capacities of artificial machines. The potential consequences of an intelligence explosion force us to think hard about values and morality and about consciousness and personal identity. In effect, the singularity brings up some of the hardest traditional questions in philosophy and raises some new philosophical questions as well».

## LECTURE 4: DEBATE ON SUPERINTELLIGENCE

### DEEP DISAGREEMENT

Artificial Intelligence High Level Expert Group (AIHLEG), *Draft Ethics Guidelines for Trustworthy AI*, 2018

#### 5.5 Potential longer-term concerns

**This sub-section has proven to be highly controversial in discussions between the AI HLEG members, and we did not reach agreement on the extent to which the areas formulated below raise concerns. We therefore ask specific input on this point from those partaking in the stakeholder consultation.**

All current AI is still domain-specific and requires well-trained human scientists and engineers to precisely specify its targets. However, extrapolating into the future with a longer time horizon, critical long-term concerns can be identified – which are by their very nature speculative. The probability of occurrence of such scenarios may from today's perspective be very low, yet the potential harm associated with it could in some instances be very high (examples thereof are the development of *Artificial Consciousness*, i.e. AI systems that may have a subjective experience,<sup>18</sup> of Artificial Moral Agents<sup>19</sup> or of Unsupervised Recursively Self-Improving Artificial General Intelligence (AGI)<sup>20</sup> – which today still seem to belong to the very distant future). A risk-assessment approach therefore invites us to keep such areas into consideration and invest resources into minimizing epistemic indeterminacy about long-term risks, unknown unknowns and “black swans”<sup>21</sup>. We invite those partaking in the consultation to share their views thereon.

### RELEVANCE

#### Does Superintelligence raise relevant concerns?

Each author gives his opinion, there are those who believe “yes” are and those who do not.

### WHO SAYS YES

- Nick **Bostrom**, *Ethical Issues in Advanced Artificial Intelligence*, 2003.

**YES** (Better safe than sorry!): «there seems currently to be no good ground for assigning a negligible probability to the hypothesis that superintelligence will be created within the lifespan of some people alive today. Given the enormity of the consequences of superintelligence, it would make sense to give this prospect some serious consideration even if one thought that there were only a small probability of it happening any time soon».

Bostrom talks about **preconsciousness**. It does not matter when the singularity will occur and with what probability; what matters is that we have to prepare and take it into consideration.

- David **Chalmers**, *The Singularity: a philosophical analysis*, 2010.

«the singularity idea is clearly an important one. The argument for a singularity is one that we should take seriously. And the questions surrounding the singularity are of enormous practical and philosophical concern». (Rileggi qui ultimi punti lezione precedente: “*practically and philosophically*”).

### WHO SAYS NO

- Luciano **Floridi**: *Singularitarians, Atheist, and why the...*

It's «a debate of mass distraction. Like all views based on faith, Singularitarianism is irrefutable. It is also ludicrously implausible. (...) any concern about the appearance of some superintelligence is laughable».

Main Issue → **Singularitarianism distracts from what is really relevant:**

«I have realized that Singularitarianism is irresponsibly distracting. It is a rich-world preoccupation, likely to worry people in leisure societies, who seem to forget what real evils are oppressing humanity and our planet, from environmental disaster to financial crises, from religious intolerance and terrorism to famine, poverty, ignorance, and appalling living standards».

For Floridi, the problem of Superintelligence is almost **unscientific**, it depends on "fate", it is so unlikely that in his opinion it is not even worth discussing.

### WHAT REALLY MATTERS

«any apocalyptic vision of AI is just silly. The serious risk is (...) that we may misuse our digital technologies, to the detriment of a large percentage of humanity and the whole planet. We are and shall remain for the foreseeable future the problem, not our technology. We should be worried about real human stupidity, not imaginary artificial intelligence». → **It is still an all too human problem of tool use!**

- Daniel **Dennett**, *Will AI achieve consciousness? Wrong question*, 2019.

Autonomous tools vs. agents: «we're making tools, not colleagues, and the great danger is not appreciating the difference, which we should strive to accentuate, marking and defending it with political and legal innovations».

AI = Autonomous Tool → Specific issues:

«AI in its current manifestations is parasitic on human intelligence. It quite indiscriminately gorges on whatever has been produced by human creators and extracts the patterns to be found there— including some of our most pernicious habits. [→ *qui fa riferimento al bias del ML*] These machines do not (yet) have the goals or strategies or capacities for self-criticism and innovation to permit them to transcend their databases by reflectively thinking about their own thinking and their own goals. They are, as Wiener says, helpless, not in the sense of being shackled agents or disabled agents but in the sense of not being agents at all—not having the capacity to be “moved by reasons” (as Kant put it) presented to them. It is important that we keep it that way, which will take some doing».

Dennett therefore argues that we must **suspend judgment and speculation and rather make political decisions**.

Do we need Superintelligent Machines?

«we don't need artificial conscious agents. There is a surfeit of natural conscious agents, enough to handle whatever tasks should be reserved for such special and privileged entities. We need intelligent tools. Tools do not have rights and should not have feelings that could be hurt or be able to respond with resentment to “abuses” rained on them by inept users».

Humans, take responsibility!

«It will be hard enough learning to live with them without distracting ourselves with fantasies about the Singularity in which these AIs will enslave us, literally. The human use of human beings will soon be changed—once again—forever, but we can take the tiller and steer between some of the hazards if we take responsibility for our trajectory.».

Per alcuni dobbiamo smettere di produrre *conscious machine* perché non ne abbiamo bisogno. Al centro c'è l'uomo. Non dobbiamo correre il rischio della superintelligenza perché non abbiamo bisogno "naturale" di essa. Non ci serve. Smettiamo di produrla. Manteniamo solo i tool perché quelli ci servono. Riportare alla centralità solo gli esseri umani.

Ma le circostanze cambiano, in futuro non avremo gli stessi bisogni di oggi. Una visione così conservatrice non può pensare di bloccare il progresso.

Abbiamo davvero il diritto di dire che siamo il livello più alti dell'evoluzione e non può esistere niente di meglio di noi esseri umani?

Abbiamo il diritto di decidere che i tool non hanno diritti? Possiamo decidere di bloccare il progresso solo perché "siamo già abbastanza" e "non ci serve altro progresso"?

Non dobbiamo produrre altra intelligenza perché non ci serve, non ce n'è bisogno, ci servono solo tool. Possiamo avere la pretesa di dire questo? Visione eccessivamente conservatrice. Il progresso non può fermarsi.

Non c'è una via giusta e una via sbagliata. Possiamo manipolare la risposta sociale a qualcosa così come possiamo manipolare una tecnologia. Non esiste la via di mezzo o "il modo giusto" di fare qualcosa... è relativo. La via giusta sarà una qualsiasi, basta che la società non si ribelli troppo. Infatti anche la società può essere manipolata.

C'è differenza tra immaginare un agente più intelligente di noi ma dello stesso tipo di intelligenza (mente umana ma più potente). Quello che non possiamo immaginare è una mente non umana: nozioni di scorrere del tempo, causalità, risorse limitate (concezione del Dio Cristiano). Questo tipo di intelligenza diversa dalla nostra non la possiamo concepire. I problemi circa la superintelligenza esaminati finora valgono in entrambi i casi.

Se costruiremo qualcosa del secondo tipo di intelligenza non lo sapremo nemmeno perché non lo possiamo capire: non sono applicabili concezioni di goals, causalità, ecc... paradossalmente potremmo già avere qualcosa di questo tipo e non saperlo.

Does Superintelligence raise relevant concerns?	
<u>YES</u>	<u>NO</u>
Hans Moravec, <i>Mind Children</i>	Luciano Floridi, <i>Singularitarians, Altheist, and why the ...</i>
Ray Kurzweil, <i>How to create a mind</i> , <i>The singularity is near</i>	
Nick Bostrom, <i>Superintelligence</i>	
Max Tegmark, <i>Life 3.0</i>	

---

## LECTURE 5: HUMAN AND ARTIFICIAL INTELLIGENCE

### RECAP OF ALL PREVIOUS LECTURES

- Artificial Agency = Purposeful Behaviour (→Wiener, Cybernetic) & Intelligent Behaviour (→Alan Turing).
- Artificial Agents = **Technological Delegates** (very productive way to understand what is the main relationship between human beings and technology) **vs. Artificial Subjects**. In fact if we follow a way of reasoning (that is also in Turing paper, so from the very beginning), we might want to review this understanding and to frame Artificial agents in Artificial subjects. → two **different phenomena**: Human Intelligence and Artificial Intelligence.

But their differences are not very important. What is important is their **similarity**: we want to understand them as the same thing. If we do this we are on the way of **superintelligence** and **singularity** theories.

There is also the **opposite-superintelligence theory**: there are essential differences between AI and Human Intelligence and we have to stick to them if we want to understand.

Discuss a different take of the same subject that interprets in a very different way the organism-machine analogy (vedi lezione cybernetic).

- Organism-Machine Analogy: *imitation/simulation/duplication* of Biological Behaviour.  
The theory of superintelligence is based on the fact that this analogy is a form of duplication of Human Intelligence so there is not any essential difference between the module and the copy.  
Today we start with the presupposition that this kind of imitation and simulation cannot be interpreted as a duplication but we still need to focus on the differences.
- Problem 1: Extent of imitation. Is there a line between human and technological agency? Is there a line between human and technological intelligence?
- Problem 2: **Conceptual confusion** in interpret AI and talking about artificial artefacts. Why do we attribute “intelligence” or “agency” to AI systems? What do we mean by that? These terms belong to a semantic context that is human: linguistically, we are already attributing human qualities and terms that pertains to human beings.



“Analogy” è qui usato in senso filosofico come termine ben preciso che indica il modo in cui decidiamo di analizzare le due facce, ovvero sia similarità ma anche differenze tra AI e HI. Dobbiamo prendere in considerazione entrambi gli aspetti. Se ci soffermiamo solo (o più) su uno o sull'altro può diventare problematico da una prospettiva metodologica.

### HUMAN VS. ARTIFICIAL INTELLIGENCE

Critica più rigorosa al problema è contenuta nel paper: Hubert Dreyfus, *Alchemy and Artificial Intelligence*, 1965. Filosofo Dreyfus scrisse questo paper dal tono polemico che ebbe molta rilevanza.

Giustapposizione tra AI e alchimia → polemica → da punto di vista epistemologico di come il discorso scientifico è rigoroso e *strong* AI può essere problematica tanto quanto

alchimia (qualcosa di magico???)

**Hype problem.** Problema metodologico e linguistico: ciò che l'AI può raggiungere è *boosted* dai ricercatori (per una serie di motivi... attrarre clamore, esposizione sui social, ...) → da punto di vista scientifico questo è problematico perché è difficile capire a che livello è davvero la ricerca e quali sono i risultati effettivi.

AI research relies on too high expectations and tends to present results in misleading ways:

«the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that **create**. Moreover, their ability to do these things is going to increase rapidly until—in a visible future—the range of problems they can handle will be coextensive with the range to which the human mind has been applied» (H.A. Simon, 1957).

→ Il dominio della **creatività** è umano o applicabile anche ad agenti artificiali? Questione aperta già negli anni '50 quando è stato scritto questo paper!

In questo paper, focus sul linguaggio: termini indeterminati e indefiniti (*thinking, learning, creating*, ma anche l'ultima frase).

Dreyfus's response (very ironic) to Simon's previous statement:

«Feigenbaum and Feldman claim that tangible progress is indeed being made and they define progress as “displacement toward the ultimate goal”. According to this definition, the first man to climb a tree could claim tangible progress toward flight to the moon».

**We had already decided that there is a continuity** between HI and AI, like there is a continuity in “climb a tree and flight to the moon”.

Ma questo non ha molto senso logico perché troppo estremizzato: avere dei risultati nell'AI non basta per poter dire che esiste questa continuità.

Infatti dobbiamo anche tener conto della dimensione in cui misuriamo questi progressi: nel caso dell'albero e della luna la dimensione è geometrica ma non ha molto senso.

Non abbiamo la stessa dimensione di riferimento per parlare di HI e AI come se fossero elementi diversi della stessa cosa. Questo possiamo farlo per progressi come l'uso del primo tool da una scimmia e la capacità di pilotare uno shuttle.

Lungo quale dimensione stiamo misurando il progresso? Da questo dipende la distanza tra il punto in cui siamo e il goal finale. Perché se il goal è "replicare processi biologici" allora non stiamo andando in quella direzione. Ma per altre dimensioni di riferimento invece possiamo vedere che i progressi ci sono, come per esempio "replicare certi risultati".

Questo problema di non definizione della dimensione di riferimento e di non poter avere la stessa dimensione per HI e AI è ciò che fa ironia nell'affermazione di Dreyfus.

C'è connessione, continuità tra mente umana e artificiale? Avere risultati nell'AI non è sufficiente per poter affermare di sì. C'è analogia tra le due, abbiamo analizzato le similarità ma **ora dobbiamo tenere conto ora delle diversità**.

**Keypoint on Dreyfus:** AI researchers are too impatience in equating HI e AI. Reasoning should not be so rush. **In his opinion, similarity are given too much space and differences too little.** Impatience generate a sort of *intellectual smog*.

AI researchers are too impatient in equating human and artificial intelligence. Impatience generates «intellectual smog» that must be dissipated → **Philosophy of Artificial Intelligence**.

Confusione generata quando le persone dicono che l'AI sa parlare, fare cose simili all'uomo, pensare, ... quando usiamo questo linguaggio creiamo "intellectual smog". Esso rende sfocata la linea distintiva di HI e AI. Obiettivo filosofico è dissiparlo e vedere attraverso esso, vedere cosa rimane quando togliamo lo smog.

«This output of confusion makes one think of the French mythical beast which is supposed to secrete the fog necessary for its own respiration».

Questa confusione nel linguaggio è tuttavia necessaria però per attrarre clamore mandare avanti il progresso motivando le persone nella ricerca. Questa è una visione stimolante e molto interessante.

In Dreyfus idea there is a fundamental problem at the very base of how AI was conceived in the 50's.

Misleading assumption:

«Underlying their optimism is the conviction that human information processing must proceed by *discrete steps* like those of a digital computer, and, since nature has produced intelligence behavior with this form of processing, proper programming should be able to elicit such behavior from machines».

Che questo sia corretto o no, qui Dreyfus si sta muovendo da una analogia a quella che sembra più una identità tra HI and AI. La similarità, per esempio nel processare informazioni in un cervello umano o in un computer, è ciò che è importante nell'AI e le differenze sono totalmente ignorate. → chiaro esempio di *overlook*, sbilanciamento nel guardare solo un aspetto, una faccia dell'analogia.

Non vogliamo stabilire se questo è vero o no ma ci focalizziamo sulla metodologia e sulla logica nell'esporre l'argomento.

«Although machines do, people do not perform intelligent tasks by simple determinate steps».

Innanzitutto quest'affermazione non viene sostenuta da nessun dato, ricerca, risultato scientifico. È fine a sé stessa → argomenti vaghi, indefiniti. Cosa si intende qui per intelligenza?

Inoltre, il fatto che un sistema discreto non sia intelligente non è vero. E in realtà il comportarsi per passi discreti è una cosa che gli umani fanno molto spesso (esempio: ricetta).

Understanding AI correctly:

«there is no reason to deny the evidence that human and mechanical information processing proceed in entirely different ways. At best, research in artificial intelligence can write programs which allow the digital machine to *approximate*, by means of discrete operation, the results which human beings achieve».

Dobbiamo stare attenti quando parliamo di “simulazione” e “analogia” qui e nel caso di Turing.

La dimensione della simulazione è uno spazio concettuale in cui **le conclusioni valgono solo se le premesse valgono**. → If the analogy between AI and HI is possible only under certain circumstances, if these do not hold any more we need to review our conclusions → **we can't move from simulation to identity/omology.**

Do not focus on similarities or discrete steps but on the topic itself: the key point of Dreyfus here is that, **even if inputs and outputs are not different** (when we compare human and artificial executing task), **this is still an approximation, an imitation, NOT an identity between the two elements.**

→ **Simulation**: similar results through different means.

→ **Two kinds of information process: *digital* vs. *biological*.**

**The result may also be similar in biological and digital processes but the medium is different and this must be taken into consideration.** → again on the concept of focus on differences.

## RECAP OF THE TWO MAIN PROBLEMS

When we compare HI and AI:

- **Socio-scientific** problem: **Hype** for social recognition.  
→ Risk: generating false expectations and misleading beliefs concerning the real capacities of technologies.
- **Communication and conceptual** problem: How can we make sense of what technologies do?  
AI imitates human behaviour → same concepts, same words! Using the same terms for both. Same semantic field to account to HI and AI.  
→ Risk: **shifting from imitation/analogy** (similarity in dissimilarity) **to reproduction** (perfect correspondence, equation).

## SYNTAX AND SEMANTICS

John Searle, *Minds, Brains, and Programs*, 1980.

Introducing terminology: **Weak vs. Strong AI**. Terms that want to distinguish between two epistemological attitudes in the field of research, concerning the way in which AI refers to HI.

**It's different from Narrow and General AI!**

«According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool (...) to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states».

Here the background is the study of the human mind and AI is studied exactly for this purpose: AI here is seen as a tool to explain what happens in the human mind.

**Weak AI**: a digital computer can be a useful tool to formulate and test hypothesis in a rigorous and precise way, regarding some aspects of HI. Again, the analogy. A digital computer can help us to understand some aspects of human minds. **Some aspects and not others! → NO reduction of the human mind on its fullest to a digital computer.** The machine is not the most complete model of the mind.

On the contrary, according to **Strong AI**, there is continuity in HI and AI.

The subject of Searle's controversial consideration concerns strong AI: an idea that the computer if programmed appropriately "*can understand and have other cognitive states*".

We focus here on the notion of "understanding".

Searle vs. Strong AI:

«My discussion here will be directed at the claims I have defined as those of strong AI, specifically the claim that the *appropriately* programmed computer literally has cognitive states and that the programs thereby explain human cognition».

Non siamo più nella dimensione dell'analogia ma usiamo il linguaggio come se si applicasse letteralmente a ciò che stiamo descrivendo. Nella prospettiva della *strong AI* non il computer letteralmente comprende e ha stati mentali, così come li ha l'essere umano.

Se esista questa corrispondenza, questa continuità, tra digital computer e mente umana allora il digital computer diventa il modello per spiegare la cognizione umana.

Un altro avverbio su cui ci focalizziamo è "*appropriately*": infatti Searle sta dicendo che noi concretamente possiamo programmare un computer in modo appropriato per il task. Ovviamente non sappiamo (ancora) cosa significa programmare appropriatamente. Il fatto che non siamo ancora a tal punto, tuttavia, non significa che non dobbiamo tenere in considerazione la possibilità di arrivarci. E proprio questo Searle ci vuole dire.

Method of inquiry: a *Gedankenexperiment*. The second most well-known thought experiment in philosophy of computations (after the Turing test): **The Chinese Room**, designed by Searle in this paper.

### THE CHINESE ROOM

«Suppose that I am locked in a room and given a large batch of Chinese writing. Suppose furthermore (as indeed is the case) that I know no Chinese (...). To me, Chinese writing is just so many meaningless squiggles.

Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes.»

He is setting the stage to reproduce a computing process.

So for him his just drawing, he cannot understand letters, words and meaning of what he is writing. He has only an understanding of the rules (in English) to correlate the writings of the first batch to those of the second.

"Formal" because it's not important the content of the symbol but the symbol itself.

«Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch.

Unknown to me, the people who are giving me all of these symbols call the first batch a "script", they call the second batch a "story", and they call the third batch "questions". Furthermore, they call the symbols I give them back in response to the third batch "answers and questions", and the set of rules in English that they gave me, they call the program.»

Ora Searle ci dice che a tutti gli elementi che ha introdotto finora in maniera formale (es. scritti cinesi, rules, ...) viene dato un **significato molto diverso se l'interessato non è più lui nella stanza ma le persone al di fuori**.

Impariamo che alla stessa cosa possono essere dati nomi diversi e questa dualità di termini è esattamente il punto in cui l'esperimento è basato:

Searle	Programmers	Computer Science
1 <sup>st</sup> batch Chinese symbols	Script	Language
2 <sup>nd</sup> batch Chinese symbols	Story	Input
3 <sup>rd</sup> batch Chinese symbols	Questions	Assignment
Chinese symbols given back	Answers	Output
English set of rules	Program/Linguistic skills	Program

Searle	Human	Computer
1 <sup>st</sup> batch Chinese symbols	Symbols/Letters	Language?/Data?
2 <sup>nd</sup> batch Chinese symbols	Story	Input
3 <sup>rd</sup> batch Chinese symbols	Questions	Assignment
Chinese symbols given back	Answers	Output
English set of rules	Logic? Grammar? Syntax? Probability?	Program

Why Searle introduces this experiment?

«Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view (→ **behaviourism**) —that is, from the point of view of somebody outside the room in which I am locked—my answer to the questions are absolutely indistinguishable from those of native Chinese speakers. (→ this is a sort of reworking of the Turing test: Searle has in mind that way of reasoning.)

(...) in the Chinese case, (...) I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer: I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the program».

Ciò che per Searle sono solo prodotti di una manipolazione di simboli, per le persone esterne alla camera sono risposte. In this scenario, he is just a mechanical device, implementing the program.

There is, for him, a correspondence between what he is doing in the room and what a digital computer do: **Searle's claim: a digital computer is just a tool for manipulating symbols without any understanding of what they mean.**

Experiment's conclusions.

Human understanding vs. symbol manipulation:

«imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. (...) In the Chinese case I have everything AI can put into me by way of a program, and I understand nothing; in the English case I understand everything, and there is so far no reason at all to suppose that my understanding has anything to do with computer programs, that is, with computational operations on purely formally defined elements».

Qui introduce uno scenario simile con cui paragonare ciò che è accaduto nella *Chinese Room* perché da un punto di vista esterno non c'è differenza: sono sempre due conversazioni in cui c'è uno scambio di informazioni e i risultati ottenuti sono molto simili (storie, domande, risposte, ...). Le situazioni da un punto di vista esterno sono simili.

MA secondo Searle c'è una differenza importante tra le due situazioni che rischia di essere trascurata e assumiamo un punto di vista esclusivo. Nel caso della *Chinese Room* e quindi nella manipolazione di simboli non c'è comprensione. Nel caso del secondo scenario (English stories) c'è comprensione.

Searle sta cercando di **chiarificare che c'è differenza nel modo in cui sono processate le info in un digital computer e da un essere umano che comprende il linguaggio**. Sta cercando di focalizzarsi e sottolineare queste differenze.

Quando gli vengono fatte domande in inglese e lui risponde in inglese sta istanziando un tipo di "programma" che NON può essere spiegato con riferimento al modello della processazione di un digital computer. Non c'è ragione di farlo.

**Punto centrale: questo tipo di simulazione di domanda-risposta tramite manipolazione di simboli è sufficiente per poter parlare di "comprensione" da parte di un digital computer?**

Questa idea di trasformare un insieme di simboli in un altro insieme di simboli senza comprensione è la base della **logica formale**. Logica tipografica = modo di ricavare asserzioni corrette da asserzioni corrette, semplicemente applicando un insieme di regole fisse su simboli non interpretati e senza capire cosa significa il simbolo. Ciò che otteniamo sono delle dimostrazioni: il primo insieme di simboli può infatti essere interpretato come il teorema e il secondo la sua dimostrazione. Ragionamenti corretti senza comprensione dei simboli. L'esperimento di Searle dunque è pura logica formale.

## LECTURE 6: DEBATE ON SEARLE'S CHINESE ROOM

[The Chinese Room Argument Youtube Video](#)

Is the programme instantiating by Searle in the room an exhaustive representation of human intelligence?

**NO.**

Even though it works, and even though from an external p.o.w. the results are good enough, there is something missing in this representation. There are some differences (from an internal p.o.w.) that have been overlooked: these differences refer to the notion of ***understanding***.

There is an opposition between **syntax** and **semantic**: the first one is the domain in which only symbols (mere manipulation of symbols that have no meaning) while the second one is no blind manipulation, there is understanding of the meaning.

"Understanding a concept" is put in opposition with "manipulating symbols" → *syntax VS semantic*.

Obviously, semantic is based on syntax: there is no understanding of symbols if there are no symbols.

What Searle says:

- 1) «I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing». There is no grasp based on meaning, but just blind symbol manipulation.

Again, this is a **methodological distinction: *Behaviourism – Externalism*** on one side, and ***Functionalism – Internalism*** on the other (vedi lezione su *cybernetic*).

Qui compare di nuovo differenza tra internalismo ed externalismo: se non ci interessa della struttura del computer/essere umano e ci interessa solo il suo comportamento.

- 2) «the computer and its programs do not provide sufficient conditions of understanding since the computer and the programs are functioning, and there is no understanding».

There is a difference between the functioning of the digital computer and the understanding that the mind performs. This does not mean that the mind is NOT some sort of computer that execute some sort of program: this is Searle's opinion. The mind instantiates some programs BUT the mind itself is NOT a digital computer, so the program that it instantiates is DIFFERENT from the program that a digital computer instantiates. In computer there is only symbols manipulation and understanding cannot be explained by this model.

In his paper, Searle motivates all these considerations.



N.B.

There is an implicit assumption here, that is: that *understanding* is something different from *blind symbol manipulation*. What if these two turn out to be the same thing?

Se si scoprisse che non esiste "comprensibilità" e "significato" ma solo manipolazione di simboli? Tutte le argomentazioni di Searle cadrebbero.

Per noi le argomentazioni di Searle non sono abbastanza convincenti perché potremmo dire che esattamente per il fatto che l'output sono indistinguibili, allora la comprensione e la manipolazione di simboli sono la stessa cosa. L'assunzione a priori che siano due cose diverse è un preconcetto con cui giudichiamo poi il risultato finale dell'esperimento.

### CONCLUSION 1: ONLY SYNTAX, NO SEMANTICS

«formal symbol manipulations by themselves do not have any intentionality; they are quite meaningless, they aren't even symbols manipulations, since the symbols do not symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output».

Remember that Searle's arguments are the result of a research project that revolves around the philosophical idea of *intentionality* → in philosophy of mind it means the fact that some mental contents are about something that exist in the real world.

**Intentionality = aboutness, the quality of being about something.**

The idea in Searle's experiment is the following: while **mental states** (→ ways in which a being - human or artificial – understand meanings) are about something (and this aboutness is the point from which we explain what happens in our mind), **in symbols manipulation there is no aboutness**: there is only syntax.

The fact that this "signs" actually symbolize something (→ semantic domain) is a dimension added by human being that constitute the social context.

### CONCLUSION 2: SIMULATION IS NOT DUPLICATION

«For simulation, all you need is the right input and output and a program in the middle that transforms the former into the latter. That is all the computer has for anything it does. To confuse simulation with duplication is the same mistake. (...) »

There is a difference between digital computers and human being in relation to the notion of *understanding* and *creating content*.

If we use the relationship between the computer and meaning to explain the relationship between human mind and meaning we make a methodological mistake and we reduce human understanding to symbol manipulation (see again behaviourism). → this is what strong AI does!

For Searle **strong AI is unreasonable**.

«In much of AI there is a residual behaviorism or operationalism. Since appropriately programmed computers can have input-output patterns similar to those of human beings, we are tempted to postulate mental states in the computer similar to human mental states. (...) if AI workers totally repudiated behaviorism and operationalism much of the confusion between simulation and duplication would be eliminated».

Now the situation is reverse: we don't explain the human mind by means of computer science but we project human mental states to the computer. This is a problem because this kind of projection is scientifically ...

For Searle, **weak AI is perfectly reasonable** because AI might actually help us to understand the way in which the mind works but not on the part referred to the notion of *understanding*.

For Searle a machine can only execute but now we have machines that learn from their own experience and behaves differently from what they have been programmed to do according to this new learned experience.

### OBJECTIONS

Searle's paper is a sort of discussion of Turing's idea: there is a hidden debate included in the paper between Searle and Turing. We might frame this experiment by reference to one of the objections that Turing explores in his paper: the *argument from consciousness*.

#### 1. TURING'S ARGUMENT FROM CONSCIOUSNESS

«Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it. [...] According to the most extreme form of this view the only way by which one could be sure that a machine thinks is to be the machine and to feel oneself thinking. [...] Likewise, according to this view the only way to know that a man thinks is to be that particular man (*solipsism*). [...] Instead of arguing continually over this point it is usual to have the polite convention that **everyone thinks**». → *Behaviourism!*

**One thing is to compose something by blind symbols manipulation, and another thing is to compose it because there is an intention behind it (a feeling, an emotion, a *human understanding*).**

But Turing has a very interesting way to turn the table: **the only way we have to know the other human being create something because they have intentions, thought, emotions, ... is to project what we feel when we do similar things. We don't have access to the personal experience of other people just like we do not have full access of what happens inside a digital machine.**

**If we project our experience to other human beings there is no reason not to do it with computers as well.**

*Solipsism* indicate the fact that we can be sure only about our own experience and not what happens in other minds and how others experience things.

Turing solves this theoretical difficulty with a “polite convention” that, because humans and computers behaves in a similar way, everyone thinks.

### 2. SEARLE VS. TURING

How Searle answer to the solipsistic problem raised by Turing.

«The *attributions of intentionality* that we make to the robot (...) have nothing to do with formal programs. They are simply based on the assumption that *if the robot looks and behaves sufficiently like us, then we would suppose, until proven otherwise, that it must have mental states like ours* that cause and are expressed by its behavior and it must have an inner mechanism capable of producing such mental states. If we knew independently how to account for its behavior without such assumptions we would not attribute intentionality to it, especially if we knew it had a formal program».

→ Since we are able to explain the computer behaviour without any reference to intentionality, than we should not consider the computer as it *understand* something. On the contrary, the attribution of intentionality to animals is more reasonable because there a similarity *in the way in which we are made!*  
**For this similarity in materials we can infer a similarity in which we work.**

«To see this point, contrast this case with cases in which we find it completely natural to ascribe intentionality to members of certain other primate species such as apes and monkeys and to domestic animals such as dogs. The reasons we find it natural are, roughly, two: we can't make sense of the animal's behavior without the ascription of intentionality, and we can see that the beasts are made of similar stuff to ourselves—that is an eye, that a nose, this is its skin, and so on. Given the coherence of the animal's behavior and the assumption of the *same causal stuff* underlying it, we assume both that the animal must have mental states underlying its behavior, and that the mental states must be produced by mechanisms made out of the same stuff that is like our stuff».

With animals “the ascription of intentionality” is the better tool we have to interpret their behaviour.

Since we are made of similar things, than the causal process that stands behind the way we behave must be similar as well. There is not such similarity in the case of digital computers.

Materialism here is at the centre of the argument: digital computer cannot taken as a model of the functioning machines that we are.

«The problem in this discussion is not about how I know that other people have cognitive states, but rather what it is that I am attributing to them when I attribute cognitive states to them. The thrust of the argument is that it couldn't be just computational processes and their output because the computational processes and their output can exist without the cognitive states. It is no answer to this argument to feign anesthesia».

So the problem is not the projection itself but it revolves around the content of this projections (mental states).

## SUMMARY

- Two logics: **symbol manipulation (syntax) VS understanding-meaning (semantics/intentionality)**.
- Two “machines”: **digital computer VS brain**: different causal powers. Differences must not be lost.

- **Methodological issue: Behaviourism, if generalised, fails to acknowledge this difference.**

«could a digital computer think?»<sup>1</sup> If by “digital computer” we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of course, **yes**, since we are the instantiations of any number of computer programs, and we can think. But could something think, understand, and so on, solely in virtue of being a computer with the right sort of program? This I think is the right question to ask, (...) and the answer is **no**». → this is precisely the point of the Chinese Room argument.

- **Linguistic issue.**

- The main problem here is the use of language in an undetermined way for notions like understanding, thinking, meaning and so on.<sup>1</sup>
- But there is also a problem of **projection/anthropomorphism**: in order to account for the functioning of the machine we have not a specific language to use but we use the same language that we use to refer to human acting. It is just a linguistic shortcut, not a scientific way to refer.

«We often attribute “understanding” and other cognitive predicates *by metaphor and analogy* to cars, adding machines, and other artifacts, but nothing is proved by such assumption. (...) The reason why we make these attributions is quite interesting, and it has to do with the fact that in artifacts we extend our own intentionality; our tools are extensions of our purposes, and so we find it natural to make metaphorical attribution of intentionality to them» → this happens even more with Artificial Autonomous Agents (AAA): why? (it is very common for example to say that an AAA “talks”).

---

<sup>1</sup> Again there is confusion! We were talking about *understanding* but now Searle talks about *thinking*. These two things are different!

## II. SECOND PART: ETHICAL AND SOCIAL ISSUES

### LECTURE 7: ARTIFICIAL AGENCY AND RESPONSIBILITY (PART I)

#### AGENCY AND MEANING

- Artificial Autonomous Agents (AI systems) carry out tasks without *understanding* them. They just execute rules.
- Autonomous tools are NOT autonomous agents, even though we call them so. Agency presupposes *intentionality, understanding, meaning*. Artificial “agents” *mimic* human agency by means of autonomous function execution.
- Function execution depends on programming/design, NOT on autonomous understanding of the task. What does this entail?
- Artificial vs. Human delegates: similarities and differences?
- **Are artificial agents responsible for their action? If not, who is responsible?**

	Artificial Delegatee	Human Delegatee
Execution of Given Task	Functioning	Working/following instructions/obeying orders
Functional Autonomy	Do not require human constant supervision or intervention	Do not require superiors'/colleagues' constant supervision or intervention
Functional Adaptability	Adapt to unpredicted/able scenarios	Adapt to unpredicted/able scenarios
Functional Inadequateness	Errors, bad programming	Incompetence, distraction
Betrayal	? No	Yes! intentional dissimulation
Refusal (practical/ethical)	? No	Yes! Disobedience, Resistance

#### AGENCY AND RESPONSIBILITY

We refer to the notion of **responsibility** (from now, R.) from a scientific p.o.w. (not intuitive) and its meaning is complex. We need to clarify it.

**Etimology** → latin “*respondeo*” and German corresponding root “*antwort*” → to answer for something that has been done. The ability of reacting to something that has happened outside of yourself (see the Turing test and the “answering machine”).

Not every kind of act / agents has to do with R. We do not consider children or animals responsible agents even if they are agents. R. is connected to **autonomous, voluntary, self-determined agency**.

R. is a complex concept, with **many sides and meanings depending on the context**: moral/social/political/juridical dimensions (they are different from each other and have different meanings).

**Related ideas** to R. that it is not always easy to distinguish from it: voluntariness, accountability, praise/blame, shame, compensation, retaliation, punishment, sin, power/control, moral character, engagement, trust...

R. has an interesting **temporal extension**: past, present, future. But depending on the time frame in which it is applied it has a different meaning (active VS passive).

## RESPONSIBILITY

### BASIC DISTINCTIONS:

- **Passive** R.: for something that has already happened (belongs to the **past**).
- **Active** R.: for something that has yet to happen (belongs to the **future**).
- **Role** R.: related to the role one plays in an organisation or collective (roles → **obligations** – other members of the organization expect me to do something):
  - **Formal** (ex. written on a code, rules are clear) or **Informal** (ex. delegated to common sense, rules are messier, and difficulties raises in some situations).
  - Not necessarily “moral”: criminal organisation. In fact, R. is not synonymous of “moral”: there are many kinds of R. that have nothing to do with ethics.
- **Moral** R: related to acceptable ethical obligations, norms, duties, values.
- **Professional** R: related to the profession one exercises (role) + the values that are important for the profession (moral).

### PASSIVE RESPONSIBILITY

It is perhaps the most common way in which we think R. It is “held somebody responsible for something” → ask somebody to **provide a rational justification for something** that is happened and has to do with him/her doing something (or NOT doing something they should) → “rational justification” means to explain why something was made (instead of something else) or some decisions were taken. **Give reason** to support choices or actions.

The “**rationality**” aspect of justification is important because it’s what allows us to do a common discourse about. We need the reasons to be **socially understandable**.

The ability/condition to provide justification is known as **accountability**. Not every agent can be accountable for something because it cannot provide a rational justification for something (ex. an animal or a child). Those agents that we don’t consider responsible, we also don’t consider accountable.

In order to hold somebody responsible for something we also need another step. We need **blameworthiness**: situation where it is proper to blame somebody for something that has happened.

There can be situations where somebody is accountable but at the same time blameworthy.

In order to attach blame we need 4 conditions.

### CONDITIONS FOR BLAMEWORTHINESS:

1. **Wrong – doing:** a (written/unwritten, organisational/moral) norm has been violated. The agent did something wrong.
2. **Causal Contribution:** causal connection between the agent's action/inaction and the consequences he/she is held responsible for.
3. **Foreseeability:** it must have been possible for the agent to foresee the consequences of their own action/inaction. Otherwise it would be unfair to hold the agent responsible.
4. **Freedom of action:** the agent did not act or restrain from acting under compulsion or coercion (otherwise again it would be unfair to hold the agent responsible). Actions must be the result of an autonomous behaviour. → is a machine coerced in its functioning in the way it has been programmed? This is an important question in case we have to transfer responsibility to the creator of the machine.

### ACTIVE RESPONSIBILITY

Is connected to the chance of doing good. It has to do with the question “how can I improve the world?”.

**Has *always* moral significance:** avoid harm, realize positive consequences.

It requires a guiding principle of conduct, or **Ideal:** motivating ideas, inspiring strivings which aim at achieving an optimum or a maximum. Can be personal (religious beliefs, happiness, private convictions) but also professional (effectiveness, efficiency, human welfare).

This kind of R. is not connected to acts that happened but requires a character trait (the way in which we live our lives or exercise our profession): in philosophy this is known as **virtue**.

Active R. is not single acts but spread into an entire life.

Virtuous attitudes in professions:

- Perception of possible value/norm/obligation/violation (the ability to understand when something *will* probably go wrong).
- Thorough consideration of possible consequences.
- Autonomy of judgment (no conflict of interests, no corruption, ...).
- Verifiable and consistent conduct (transparency).
- Serious commitment to role obligations.

### RESPONSIBILITY AND TECHNOLOGY:

When R. has to do with the use but also the development of technology there are many and diverse **actors:** engineers, developers, producers, users, regulators, professional organisations, educational institutes, interests' groups, trade unions, ...

**Stakeholder** = everyone who has certain interests connected to the technology.

**Interest** = things somebody strives for because beneficial for or advantageous to them.

Interest involve people in situations: stakeholders are involved in situations where a specific technology is under use because they have some interest in it.

→ Problem: interests' conflict (in most of the cases, interests of different stakeholders collide). It is necessary to find a compromise, a way to align all interest.

→ Ethical challenge: take into due consideration the interests of all relevant stakeholders. Different actors have different power positions in relation to one another: **it is ethically important that interests of those who have not a relevant power position are taken into consideration.**

## AI AND RESPONSIBILITY

### THE PROBLEM OF MANY HANDS

Raises when R. allocation must be done in an organization, so **when many people are involved** in a process that has as result some consequences. It is **difficult to realize precisely WHO is responsible for something**.

The PMH **has to do with social consequences of technology** and it is generated by interaction between the actions of many different actors.

Lots of actors involved → difficult to identify where the responsibility for a particular outcome lie. It is difficult to frame R. from consequences to actors.

Many reasons of why this problem arises:

- **Fragmentation of decision making** → fragmentation of responsibility → nobody feels responsible for the process as a whole because nobody has control on the entire process.
- Morally unsatisfactory: **moral need for responsibility allocation**, for accountability (passive R) + **desire to learn from mistakes** (active R).

Ex. if I have been damaged it's my right (moral need) to ask for someone to be responsible.

Note that Artificial Agents are a product of many hands: most of the time it is difficult to attach R. in case something goes wrong.

### HOW TO DEAL WITH PMH?

**Methodology**: mechanism we can apply to try to solve (or handle) this problem.

Find a way to do **distribution of responsibility** (DoR): ascription or apportioning of (individual) responsibilities to the various members of a collective (ex. organization).

Two main requirement for any model of DoR:

1. **Moral fairness**. DoR (the way in which we distribute R. among the members of the organization) should be fair. Otherwise we do something immoral. In order to be fair, some conditions must be applied:
  - In case of passive R there are 4 conditions (wrong-doing, causal contribution, foreseeability, freedom of action). If they are applied we are sure that the DoR is moral and fair.
  - In case of active R: persons should only be allocated responsibilities that they can live by (should have means/authority to fulfil their active R). So we have to be sure that our model of DoR don't ask too much from people (because otherwise they will not act o the expectations ad the model won't function properly).
2. **Effectiveness**. Choose the DoR that is most effective in preventing harm. We can measure it by measuring how much it is effective in preventing future harm.



## DoR MODELS

Model	Description	Pros	Cons
<b><u>Hierarchical Model</u></b>	held in R. those who occupy the <b>highest places</b> in the organisation / collective	clear, simple, feasible and easy to be applied	Not everyone has full control of what happens at any degree of the organization, so: <ul style="list-style-type: none"> <li>• Not always effective in preventing undesirable consequences.</li> <li>• Somewhat unfair: Foreseeability, Causal contribution for anyone?</li> </ul> (Used in military world but not applicable for computer scientists).
<b><u>Collective Model</u></b>	<b>each member</b> of the organisation/collective is to be held jointly and severally responsible for the acts of the organisation / collective <b>as a whole</b>	clear, simple, feasible and easy to be applied	<ul style="list-style-type: none"> <li>• Individual differences cannot be accounted for (unfair)</li> <li>• No one tends to feel truly responsible for the consequences of an activity (ineffective). I know that I will not pay for what I do because consequences will be distributed on everybody.</li> </ul>
<b><u>Individual Model</u></b>	<b>each member</b> of the collective/organisation is to be held responsible <b>in relation to his/her contribution</b>	<ul style="list-style-type: none"> <li>• morally fair: 4 conditions for blameworthiness seems to apply.</li> <li>• presumably effective: everyone is encouraged to act responsibly.</li> </ul>	Might lead PMH because the difficulty here is to attach R. in an appropriate way.

## CONCLUSIONS

AI products = autonomous tool → may cause harm in many ways.

Who is responsible – legally, morally, politically?

## LECTURE 8: DEBATE ON ARTIFICIAL AGENCY AND RESPONSIBILITY (PART I)

### RECAP: AI AND RESPONSIBILITY

When we delegate to autonomous agents, we must expect something to go wrong for various reasons: errors in implementation, malfunctioning, ... How to manage these situations?

AI = autonomous tool → may cause harm.

Who is responsible – legally, morally, politically?

We cooperate in the functioning of the machine and in achieving the goal. The relationship between us and the machine is collective. But there are many human and technological actors involved → *problem of the many hands*.

- Programmers/Designers/Manufacturers
- Producers/Directors/Companies → they create the conditions under which an intelligent agent can be produced, sold, spread.
- Users/End-Users/Operators → they use these technologies in real contexts.
- Technologies themselves
- Nobody.
- Some of them.
- All of them.

Modello collettivo di distribuzione di responsabilità è una delle soluzioni possibili che coinvolge tutti, anche le macchine. Assegnare la responsabilità nel modello collettivo è però poco pratico, difficile nel sistema della società. Nella realtà non funziona.

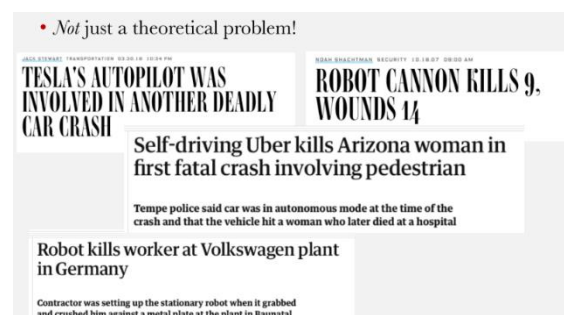
Costo associato alla responsabilità → per esempio la sanità è pagata da tutti a prescindere quanto coinvolga ciascuno nel particolare. Allo stesso modo, tutti paghiamo la stessa quota del canone Rai, a prescindere se e quanto ciascuna persona ne fa uso.

Nel caos degli umani spesso associamo la comprensione di un'azione con l'attribuzione delle conseguenze: se so e capisco cosa sto facendo allora è mia responsabilità le conseguenze. Per le macchine non è lo stesso: hanno modelli di predizione, quindi sono responsabili di qualcosa ma non è detto che ne abbiamo comprensione. Capire cosa stanno facendo è un altro discorso. → **Forciability**

Anche se sono stati introdotti come problemi teoretici non sono più tali: abbiamo già avuto a che fare con problemi in cui qualcosa è andato storto durante il funzionamento delle macchine e dobbiamo capire chi è responsabile. Esempio più comune: *driving cars*. Ma anche *industrial robotic*.

Sono problemi non teoretici ma molto pratici.

Più questi agenti artificiali saranno sviluppati e più avremo a che fare con questi problemi.



Abbiamo quindi bisogno di strumenti per gestire queste situazioni e analizzare problemi di attribuzione della responsabilità.

### ANALOGICAL THINKING

New technology that creates new problematic situations from the point of view of responsibility, but there were no concepts and arguments to deal with them precisely because they were new. We need new mechanisms, new concepts to analyze them.

Very helpful tool to analyze new situations in which IT responsibility is involved:

**Analogical Thinking** → way of thinking that compare a new situation to a situation that we are familiar with, and try to analyze the new one with reference to the familiar one.

In questo modo possiamo capire quali concetti ci servono per gestire le nuove situazioni, quali argomenti possono essere utili.

Ma il paragone deve essere convincente! Perché naturalmente nell'*analogical thinking* ci sono somiglianze ma anche differenze! Dunque la difficoltà in questo modo di ragionare è che abbiamo bisogno di un equilibrio tra somiglianze e differenze e dobbiamo assicurarci che il paragone che abbiamo selezionato sia uno strumento valido per esplorare la nuova situazione nel modo più appropriato.

Analogical Thinking → Useful research tool **in Computer Ethics**:

Notice «what happens when ethicists are presented with situations involving computer technology (*quelle che noi abbiamo chiamato “situazioni nuove”*). In order to determine whether there is an ethical issue and what is, an attempt is made to connect the computer situation with a familiar situation and with a familiar moral concept or principle. This often involves reasoning by analogy (...). The process here involves using what we know about a familiar situation to understand another, less familiar, situation, and we do this by connecting the new situation to a familiar moral concept or principle. It seems that human beings would not be able to recognize a new situation as having ethical implications unless the situation connected in some way or another to what we already understood to be an ethical matter» [D. Johnson, Computer Ethics]

Unless there is a connection between what is new and what is known, we cannot understand the ethical implications connected to the new situation. This is the base on which we can **build *analogical arguments***.

*Computer Ethics* field is full of *analogical arguments* and proposals for when **we explore new situations using frameworks that belong to cases that have nothing to do with information technology.**

Examples: **applications of *analogical thinking***

Nuove situazioni su cui ragionare in termini di attribuzione della responsabilità	Paragone con una situazione analoga, conosciuta e non appartenente al mondo informatico	Utilità
Hacking	Breaking and Entering	“Forzare ed entrare” non ha nulla a che vedere con l’informatica! È un tipo di situazione che accadeva già prima, ma ci è utile perché offre il <i>teoretical framework</i> (i.e. la concettualità e le argomentazioni) per capire quali sono le condizioni per cui l’hackeraggio è sbagliato oppure permesso.
Undisclosed Cookies	Hidden Surveillance Cameras	pensiamo ai cookies come a delle telecamere nascoste e questo ci dà la possibilità e il quadro di riferimento per poter ragionare su argomenti etici (e anche norme legali)
Copying Software or content in general) without Permission	Stealing Properties (but!)	MA c’è una differenza tra copiare un software e per esempio rubare una bici: “It doesn't disappear on the other side”. Il software rimane al proprietario, lo può ancora usare, ma io lo sto usando senza pagare (pagare il diritto di usare il software). Non sto esattamente sottraendo una proprietà a qualcuno: il proprietario sta perdendo la mia quota. ↓

Analizziamo dunque anche i **limiti** di questo tipo di ragionamento!

→ **Analogical thinking is useful but It’s risky!** Lo dobbiamo usare attentamente e tenendo conto delle differenze.

Il paragone tra situazione nuova e situazione analoga non è sempre valido.

Abbiamo molte analogie e paragoni per pensare alle situazioni in cui *artificial agents* sono la causa dell’accadere di qualcosa di male e quindi ne hanno responsabilità.

A cosa possiamo paragonare le nuove situazioni? Inquadramento dei tipi di analogia:

- Natural Events → responsabilità riguarda solo la reazione all’evento, non alla causa (es. terremoto)
- Products → legal accountability and monetary compensation
- Minors and Animals → in grado di agire ma non di comprendere ciò che stanno facendo. La responsabilità ricade su un genitore o sul padrone. Situazione in cui qualcun altro è responsabile dell’agente.
- Full Person → piena responsabilità. Interessante esplorare analogia tra un *artificial delegatee* e *human delegatee*. Anche l’agente artificiale a cui abbiamo delegato i nostri compiti è pienamente responsabile di ciò che fa, proprio come un essere umano?

NO APPUNTI SECONDA PARTE DELLA LEZIONE (videolezione non registrata) → argomento: dibattito sulla responsabilità (no slide)

## LECTURE 9: ARTIFICIAL AGENCY AND RESPONSIBILITY (PART II)

### SUMMARY

- Responsibility: a complex notion. Most of the time its meaning depends on the context.
- Passive/Active, Role, Moral, Professional Responsibility.
- Passive R: Justification & the 4 Conditions of Blameworthiness.
- Active R: Ideals, Virtues, Duty to Care.
- Action, Mediation, & the Problem of Many Hands.
- Distribution of R Models: Hierarchical, Collective, Individual.
- Analytical Tool: Analogical Thinking.

### RESPONSIBLE ARTIFICIAL AGENTS

Andreas Matthias, *The Responsibility Gap. Ascribing Responsibility for the Actions of Learning Automata*, 2004.

It is a very controversial paper but it opens very important points of discussions.

Focus on human agents act through the use of Artificial Agents → delegation.

The point is that when we delegate tasks to AA based on ML it becomes unfair to hold any human being involved responsible for some consequences (those that cannot be avoided by humans).

**Postulate of control:** we can be responsible only of the processes we can control (this is a mix of some of the four blameworthiness conditions: causal contributions and foreseeability - and perhaps also freedom of action).

«For a person to be *rightly*<sup>2</sup> held responsible, that is, in accordance with our sense of justice<sup>2</sup>, she must have *control* over her behaviour and the resulting consequences “in a suitable way”. This means that the agent can be considered responsible only if he knows the particular facts surrounding his action<sup>3</sup>, and if he is able to freely<sup>4</sup> form a decision to act, and to select one of a suitable set of available alternative actions based on these facts».

Technology and no-responsibility situations:

«In situations where the operator has reduced control over the machine he also bears less or no responsibility. (...) But who could be held responsible instead? In fact, *nobody*. In such cases of accidents

---

<sup>2</sup> This is a direct reference to the *moral fairness* requirement. But there is an implied assumption: it exists a common sense of justice, equally recognized by everyone. In the reality this is not true. We can consider “rightly” here into a social-dependent framework so into the dominant ideas of responsibility allocations, so for ex. according to law or guidelines, etc... So keep in mind that this definition is relative (valid into some specific contexts), not absolute. **What is right or wrong is not absolute but depends on the social and cultural context (it's relative).**

<sup>3</sup> Reference to the *foreseeability* condition

<sup>4</sup> Reference to the *freedom of action* condition

that occur through no fault of a specific person, society refrains from ascribing responsibility, and collectively bears the costs resulting from the accident's consequences».

→ **Theorem:** Technological Autonomy ↑ = ↓ Human Control

Matthias focuses on cases when there is **no human control** and no responsibility applies. These situations occur when the operator has reduced control over the machine (ex. when we **delegate** tasks to AA).

**When artificial agents act autonomously (so with no requirement of human supervision and intervention) then the chain of events that allows us to trace harms back to those responsible breaks down.**

This tracing back from machine to humans becomes impossible and cannot be substituted by anything else! Nobody is responsible.

We have already seen situations where R. is bypassed, and the issue is framed with reference exclusively to cost.

Matthias' idea is that when the chain of R. breaks down, R. must be taken out of the frame and the only important thing is how to bear the cost resulting from the harmful consequences. **The entire discussion is shifted from a R. allocation framework to a cost distribution framework. There is no more moral component** but only an economical matter. Find the best way to distribute cost over society.

This could be read as a proposal, an introduction of a new model to handle harm caused by the functioning of AAs. This model is applied to situations that we already know.

Problem: if we do that, those who are actually responsible for something have the possibility to avoid paying.

Autonomous agents & delegation:

«Autonomous agents are artificial entities that fulfil a certain, often quite narrow purpose, by moving autonomously through some 'space' and acting without human supervision»<sup>5</sup>.

«presently there are machines in development or already in use which are able to decide on a course of action and to act without human intervention. The rules by which they act are not fixed during the production process, but can be changed during the operation of the machine, by the machine itself. This is what we call **machine learning**».<sup>6</sup>

«Connectionist systems lack an explicit representation and the contained information can only be deduced from their behaviour».

---

<sup>5</sup> Definition of AA

<sup>6</sup> Matthias introduces ML topic

→ Machine behaviour is not predictable before → **foreseeability is no more possible!**



ML systems are “black boxes” → **control is delegated / lost.**

Who is in control?

«Thus we can identify a process in which the designer of a machine increasingly loses control over it, and gradually transfer this control to the machine itself. In a steady progression the programmer role changes from coder to creator of software organisms.<sup>7</sup> (...) Essentially, **the programmer transfers part of his control over the product to the environment.** This is particularly true for machines that continue to learn and adapt in their final operating environment».

→ In the R. chain, the machine is now the closest element to the harmful consequences.

Therefore, automated behaviour «must be attributed to the machine itself and NOT to its designer or operator».

At least for now, Matthias is NOT talking about R. attribution but only about behaviour.

N.B.

BUT remember that even if we are not responsible of *what* the machine has learned we are responsible of the *learning process* because we are the programmes that write the code telling the machine *how to learn*.

## RESPONSIBILITY GAP

«Traditionally we hold either the operator/manufacturer of the machine responsible for the consequences of its operation, or ‘nobody’ (in cases where no personal fault can be identified).<sup>8</sup> Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral frameworks of society because nobody has enough control over the machine’s actions to be able to assume the responsibility for them. These cases constitute what we call the **responsibility gap**».

Control is the most important element for R. allocation: only if it is possible, it is morally fair to held someone responsible.

But we STILL have the need to find someone or something that is responsible for something. But we cannot point someone ore something that is convincedly and fairly responsible.

This is the problem that we need to face when we analyse how R. is connected with delegation to AAs. This is the problem of R. distribution in case where actions are mediated by AAs.

**The problem remains: WHO is responsible?**

---

<sup>7</sup> This is **analogical thinking**: he says “creator” referring to God that cannot foresee human behaviours (Adam and Eve).

<sup>8</sup> Here Matthias is referring to the **model of “Natural events”** (see “Analogical Thinking”).

Following Matthias, humans are not:

«If we want to avoid the injustice of holding men responsible for actions of machines over which they could not have sufficient control, we must find a way to address the responsibility gap in moral practice and legislation. The increasing use of autonomously learning and acting machines in all areas of modern life will not permit us to ignore the gap any longer».

He is not saying that machines should be held responsible.

It is time to face the issue: this situations (note that he wrote the paper in 2004) is already happening.

### REACTIONS TO MATTHIAS' ARGUMENTS

#### BUT this conclusion is highly CONTROVERSIAL!

Because the way in which problems of R. allocations are solved in the AI and robotics debates is by finding humans responsible: there is a strong will regarding the necessity of keeping humans involved. This is one of the most insisted point of the European proposition of legislation concerning AAs.

#### 1. ONLY HUMANS ARE RESPONSIBLE

This emerges in the paper:

J.J. Bryson, *How do we hold AI itself accountable? We can't*, 2018

«if you build an AI system and allow it to operate autonomously, it is essential that the person who chooses to allow the system to operate autonomously is the one who will go to jail, be fined, etc. if the AI system transgresses the law. There is no way to make the AI itself accountable».

Two important points here:

1. We cannot hold the AAs itself accountable
2. There must be a human involved: even there is no direct control it doesn't matter because there will be always some person that decided in the first place to deploy the machine. Even though there is no relation between the designer or the user and the machine, we will be always able to find some other connections between the machine and a human: machines do not deploy themselves (still! :D). We can always trace back to those people.

This is why Matthias paper is so controversial: he is stick to the idea that ONLY direct control is a necessary requirement for R. allocation. We need to move beyond the direct control presupposition, beyond this way of reasoning and explore different ways because what we want is that there is always a human being that can be held responsible.

Again in Bryson (this is a quote about the possibility of shifting R. directly to machines, so keep human out of the question):

«there is no way to ensure that a synthetic person can be held legally accountable. It does not matter whether you mean a 'synthetic person' to refer to a robot, or to the legal fiction that is used to make a



corporation<sup>9</sup> appear like a person. The only way to ensure that law is stable is to have a human to be accountable for the actions of an artefact, and the same human be the one in control of the artefact's behaviour».

Here it is assumed that there is some significant form of control that humans can exercise on the functioning of the artefact, while for Matthias this is not possible any more.

The word “control” is differently used by Matthias and Bryson:

- For Matthias, control is relevant only if it is direct
- In Bryson's idea, direct control is just one of the many possible forms of control.

→ Conclusion: for Bryson only humans are responsible entities. And the only important thing is that we find WHO has decided to deploy the machine in that specific circumstances: he/she must be held a contribute for the consequences, even if he/she cannot exercise direct control on machines' acts.

## 2. PRODUCT LIABILITY

Another possible conclusion: reframe all issue and **think the machines not as subject of law but they're object of law, because they are technological products.**

So we should apply to them the legislation that already regulates situations when harmful consequences comes from the use of products.

**AI = technological product = object, not subject, of law.**

«so long as robots do not achieve self-consciousness they cannot be deemed moral agents or autonomous—in a strong sense—beings. Short of that capacity there is no logical, moral, or philosophical—and thus not even legal—necessity to consider them subjects of law and bestow individual rights on them. Therefore, all existing robots up to a point are to be deemed objects—more precisely, artefacts created by human design and labour, for the purpose of serving identifiable human needs, otherwise known as products».

(A. Bertolini, *Robots as products*, 2013)

→ Human stakeholders are responsible: producers, programmers, users (in addition, there are product liability rules).

Bertolini says that even though ML agents are able to change behaviour and understand autonomously it doesn't matter because they are still products.

We don't need to change laws and our philosophical frameworks in order to find ways to attach R. to robot themselves. We nly have to be mor ehonest concerning the future of these technologies and treat them like any other tech product.

---

<sup>9</sup> A corporation is a collective

## SUMMARY

Two reactions to Matthias 'paper:

- Explore some other forms of control in order to assign R. to other humans → Bryson
- Challenge the particularity of the things we are discussing about (ML robots) → Bertolini

**Projection of human characteristics onto robots must be resisted** to understand that robots are just products (for what concern R. attachment):

«Our solution is to resist, as strongly and explicitly as possible, the tendency to assign any moral agency whatsoever to the “autonomous” robot and, meanwhile, be as explicit as possible (ideally in advance!) about the appropriate distribution of responsibility between robot operator, manufacturer, those giving the orders on the battlefield, those giving the orders higher up the chain of commands<sup>10</sup>, and other implicated parties. One must consciously acknowledge and consciously oppose any impulse to assign even partial responsibility to the robot or leave the distribution of responsibility unclear».

(J. Parthemore & B. Whitby, *Moral Agency, Moral Responsibility, and Artefacts*, 2012)

What is actually important is not the philosophical problem of understanding whether the machines can be held responsible or not (or when) but is that, right now, there is no way to hold machine responsible so in advance, before we are involved in situations where R. is difficult to ascribe and to be traced back to humans, we need to have clear mechanism to DoR. So that it would be clear WHO will be held R. for something, and people will behave accordantly before things go wrong.

So the point is to decide in advance how to socially and legally distribute R. so that any human being involved in that possible future situations (when bad things happened) can foresee what will happen to him/her when something goes wrong and therefore will act accordantly.

The point here is to shift foreseeability from the actual functioning of the machine to what will happen to it if something goes wrong. Foreseeability does not apply to the consequences itself but to what will happen when a consequence will present itself.

J. Parthemore & B. Whitby are here trying to reframe the idea of foreseeability in order to accommodate for the situations and the issue concerning the unpredictability of machine behaviour. **They are shifting foresee from machine behaviour (which is unpredictable because of ML) to the social reaction to harmful consequences** caused by AAs. This is a smart way to fair the issue.

## COMBINING OR DISCHARGING RESPONSIBILITY

«We take the main case of the abuse of legal personality to be this: natural persons using an artificial person to shield themselves from the consequences of their actions. Recognition of robot legal personhood could present unscrupulous actors with such “liability management”<sup>11</sup> opportunities».

(J. Bryson et al., *Of, for, and by the people*, 2017)

---

<sup>10</sup> Reference to hierarchical DoR

<sup>11</sup> Expression used in corporations R.: all branch of law that has to do with shifting R. from managers to the corporation itself. This is a very important problem.

This is not a combination but a case of dischargement: allocate R. to machine means shifting it from humans to them.

But there are also people who think that R. can be combined between humans and AAs.

«The criminal liability of an AI entity does not replace the criminal liability of the programmers or the users, if criminal liability is imposed on the programmer and/or the users by any other legal path. Criminal liability is not to be divided, but rather, combined».

(G. Hallevy, *The criminal liability of artificial intelligent entities*, 2010)

---

## LECTURE 10: DEBATE ON ARTIFICIAL AGENCY AND RESPONSIBILITY (PART II)

Today we will present to very controversial topics:

1. Can we punish AAs? Is it meaningful? → Analogical Thinking plays an important role here.
2. How to think human R. beyond the paradigm of direct control → some ideas to ascribe R. to human beings where computer scientists and programmers have a very important role.

### WHO'S RESPONSIBLE

The artificial agent itself? Conditions:

«criminal liability for a specific offense is mainly comprised of the external element (*actus reus*) and the internal element (*mens rea*) of that offense. (...) If an AI entity is capable of fulfilling the requirements of both the external element and the internal element, and, in fact, it actually fulfills them, there is nothing to prevent criminal liability from being imposed on that entity. As long as an AI entity controls<sup>12</sup> a mechanical or other mechanism to move its moving parts, any act might be considered as performed by the AI».

(G. Hallevy, *The criminal liability of artificial intelligent entities*, 2010)

Here we are thinking AAs as full persons (from the p.o.w. of criminal law we can do this because the conditions fully apply). We can **reason analogically** and institute comparison between humans and artificial agents for what concerns criminal liabilities. This is the first step for an argument that leads us directly to the idea of punishing AAs exactly like we punish human beings. The same scheme can be applied to AAs as long as we interpret punish in a very specific way (see below).

---

<sup>12</sup> Essential **link between direct control and responsibility** allocation → the argument is really similar to that of the *Responsibility Gap*.

In order to apply criminal liability we need two elements:

- an external one → if **what happens is wrong**
- an internal one → if **there was the intention** in doing that

In Hallevy's mind there is no problem in extending these two conditions to AAs.

Since the AA has **direct control** on its moving part, then R. must be allocated to it.<sup>12</sup>

## PUNISHMENT

Can we punish artificial agents? What is the analogous element of punishment for humans if the model is now applied to AAs? This is an open question and must be solved.

As in every *analogical thinking* we have to move from the familiar situation to the unfamiliar one so we need some adjustment considerations.

Hallevy explains these adjustment considerations:

1. What is the fundamental **significance of the specific punishment for a human**? We need to understand **WHY we are choosing a specific punishment for a specific felony, so which is the function of the punishment.**
2. **How does that punishment affect AI entities?**
3. What practical punishment may achieve **the same significance** when imposed on AI entities?

(G. Hallevy, *The Criminal Liability of Artificially Intelligent Entities*, 2010)

Role of these three questions: **make the transfer from human domain to artificial domain applicable**. In fact when we switch domain, something may become unreasonable or unaffacting because the two domains are analogical, not equal (analogical = some aspects are similar and some are different).

We have to frame punishment with an externalist p.o.w., so as a process where input are transformed in output.

Punishment means "providing the rights inputs".

→ "same significance" = same results = same outputs

→ Analogical thinking + Externalism!

Capital punishment:

«the significance of capital punishment for humans is the deprivation of life. The "life" of an AI is its independent existence as an entity. Considering capital punishment's efficacy in incapacitating offenders, the practical action that may achieve the same results as capital punishment when imposed on an AI entity is the deletion of the AI software controlling the AI entity. (...) The deletion eradicates the independent existence of the AI entity and is tantamount to the death penalty».

These two punishments are considered analogous because they lead to same results (**methodology fully externalist, input-output**).

Incarceration:

«the significance of incarceration for humans is the deprivation of human liberty and the imposition of severe limitations on human free behaviour, freedom of movement, and freedom to manage one's personal life. The "liberty" or "freedom" of an AI includes the freedom to act as an AI entity in the relevant area. (...) Considering the nature of a sentence of incarceration, the practical action that may achieve the same effects of incarceration when imposed on an AI entity is to put the AI entity out of use for a determinate period. During that period, no action relating to the AI entity's freedom is allowed, and thus its freedom or liberty is restricted».

The same goes for suspended sentencing, fines, and even community service!

Not very convincing... is it really meaningful to punish AAs?

Two objections:

1. «only a moral agent can be reformed, which implies the development or correction of a moral character—otherwise it is merely the fixing of a problem. (...) robots do have bodies to kick, though it is not clear that kicking them would achieve the traditional goals of punishment. (...) there is little sense in trying to apply our traditional notions of punishment to robots directly».

(P. Asaro, *A body to kick but still no soul to damn*, 2012)

He refers to the fact that **if we take an externalist perspective (behaviouristic methodology) to the problem of extending punishment to AAs we lose track of the most important element on punishment that is the correction of the moral character**. It is meaningless to say that we are reforming AAs.

When we move from the human dimension to the artificial one through analogical thinking we are building a very "unstable bridge": because the difference between the two domains are too strong and similarity are too superficial to sustain analogy.

2. «What matters is that none of the costs that courts can impose on persons will matter to an AI system in the way they matter to a human. (...) Law has been invented to hold humans accountable, thus only humans can be held accountable with it».

(J.J. Bryson, *How do we hold AI itself accountable? We can't*, 2018)

Again the analogical think does not hold because it is too weak. Punishments applied to people have not the same effect when applied to artificial systems.

Who's responsible?

EU AIHLEG, *Draft Ethics Guidelines for Trustworthy AI*, 2018.

«systems should be in place to ensure *accountability* and *responsibility*. (...) Ultimately, human beings are, and must remain, responsible and accountable for all casualties. (...) Organisations should set up an internal or external governance framework to ensure accountability. (...) Trustworthy AI also requires responsibility mechanisms that, when harm does occur, ensure an appropriate remedy can be put in place. Knowing that redress is possible when things go wrong increases trust».

## HUMAN RESPONSIBILITY

We have to **think R. beyond** foreseeability and **the postulate of direct control** to preserve our sense of moral R. and our legal accountability.

Technological Autonomy ↑ = ↓ Human Control → How to address the gap?

→ Günther Anders to Claude Eatherly:

«The 'technification' of our being: (...) today it is possible that unknowingly and indirectly, like screws in a machine, we can be used in actions, the effects of which are beyond the horizon of our eyes and imagination, and of which, could we imagine them, we could not approve—this fact has changed the very foundation of our moral existence. Thus, we can become 'guiltlessly guilty', a condition which had not existed in the technically less advanced times of our fathers».

(G. Anders & C. Eatherly, *Burning Conscience*, 1961)

Eatherly was that who released the atomic bomb in Japan. → We have today new technologies, new situations (like the atomic bomb at that time) and we must remain responsible.

Responsibility beyond direct control:

«the human being is responsible (...) not only the relationships it initiates, or the criteria on which these relationships are made, but, primarily, the same relational context, that is, the environment within which it operates. And this can be done by the human being even if that environment does not depend on him or her, even if one does not have full control over it, even if it is already activated by the action of other subjects: natural or artificial that they are. In other words, if the human being recognises this context as the context of his action, as an area in which to interact, also in order to modify it, one consciously assumes responsibility for what he or she is not responsible».

(A. Fabris, *Ethics of Information and Communication Technologies*, 2018)

**The fact that we lose direct control should not be the presupposition for an argument that free us of R.** but it is the base to say that we need to assume R. for what happens even though we are not directly connected to that.

Taking responsibility for AI-mediated actions!

«The point here is that robots, and their underlying control systems, depend on human intervention. The robots may be "set loose" to make unpredictable decisions, *but the decision to allow them to do so is a human and societal one*. Any decision made by the robot will still depend on their initial design. Even if the robots are "trained" or "evolved" to make decisions, their training or fitness regime will still have involved human intervention at some point, and it is imperative that human responsibility is assumed and recognised».

(A. Sharkey, *Can robots be responsible moral agents and why should we care?* 2017)

Even if the execution of functions is autonomous, the system, the robot built in a specific way and its deployment in a specific context and for specific purposes, all this facts link the robot to the human. **We need to enlarge our perspective (go beyond) and see that even if there is not direct control, there is anyway a link.**

**Any (autonomous) decision of the robot still depend on the initial design.**

## LECTURE 11: ANTHROPOMORPHISM

### TERMS, DEFINITIONS AND ETYMOLOGY

Interpretation of AAs as they were human beings.

Meaning of the term “**anthropomorphism**”: psychological tendency to project human qualities (elements that constitutes human form) onto non-human entities (do not belong initially to the entity itself). It is a **projection**.

Word made by combining two terms → Ancient Greek: *Anthropos* + *morphe*.

Antrop. is a limiting way of expressing what is interesting in robotics because most of the time it's just not the human qualities projected onto a robot but a more general set of qualities that describes not just human lives and human existence but more broadly the existence of animals (zoomorphism) or, more broadly, anything that is living (biomorphism).

- If we shift the focus from the human dimension to a more extended one we can speak of zoomorphism. **Zoomorphism**: psychological tendency to project animal-like qualities onto non-animal entities → Ancient Greek: *Zoon* + *morphe*.
- An ulterior extended way: **Biomorphism**, psychological tendency to project life-like qualities onto non-animal entities → Ancient Greek: *Bios* + *morphe*.

These are all **analogical process**: we use a model that we are already familiar with (human, animals) to **understand and interpret something we cannot frame**, we are not already familiar with.

Antrop. is not new. It is a very old psychological tendency that characterize the way in which human understand things from the very past.

Xenophanes of Colophon, 570-475 BC.

“But mortals suppose that the gods are born (as they themselves are), and that they wear man's clothing and have human voice and body.”

→ **Religion** is a very good context in which to inquire antrop.: God is an entity very difficult to describe and understand because we have not direct experience with it and so we use the methodology of antrop.

We have no reason to think that God has human form but it is easier for us to think it in that way.

“But if cattle or lions had hands, so as to paint with their hands and produce works of art as men do, they would paint their gods and give them bodies in form like their own-horses like horses, cattle like cattle.”

Antrop. is a **bias psychological phenomenon**: what we are familiar with depend on what we are so there is no assurance that what we describe in a certain way is really in that way.

What do we project?

When the line that separates human and robot becomes more blur, we project to robots human features. Some of the most common projected features:

- **Animacy** (in religious context = “having a soul”) → see P. Asaro’s title paper.
- **Agency** (framing robotic systems as “agents” is a form of antrop.)
- **Mental States** = concepts we use to describe the life of the mind → we project them in order to understand the behaviour of the concept.
  - Emotions
  - Feelings
  - Dispositions
  - Preferences
  - Intentions
- **Personality**
- **Gender**
- and so on...

This is a model, a strategy that we apply to frame objects that we are experiencing and interacting with.

## HOW DESIGN CAN EXPLOIT ANTHROPOMORPHISM

Operative side of the issue: antrop. has to do with the way in which we design AAs.

Anthropomorphism = **perceptive bias / cognitive bias in the elaboration of mental models.**

Noi cerchiamo di interpretare ciò che non sconosciamo per dargli un senso ma se lo facciamo usando modelli e oggetti che già conosciamo introduciamo un *bias*: il bias è **una prospettiva che ci permette di capire qualcosa in modo chiaro ma allo stesso tempo esclude qualcos’altro. Ci influenza**. Ha ruolo attivo nel modo in cui comprendiamo l’oggetto.

Quando interpretiamo il comportamento di un robot umanoide di cui non sappiamo nulla usiamo lo stesso modello mentale di quando cerchiamo di interpretare i comportamenti di un amico, un conoscente, qualcuno. → *razionalizzare* comportamenti nuovi o difficili da interpretare.

**Trasferiamo così un certo bias al fenomeno: lo guardiamo attraverso una prospettiva ben precisa che ce lo rende chiaro ma allo stesso tempo esclude qualche aspetto.**

È come indossare degli occhiali attraverso le cui lenti osserviamo i fenomeni.

Questa tendenza psicologica non è necessariamente sbagliata: può essere molto utile, funzionale ed efficiente. Può esserlo tanto quanto *l’analogical thinking*.

- Why? To quickly make sense of situations or rationalise behaviours → natural interpretative schemes (shortcuts).
- Anthropomorphism can be studied in its main processes and elements.
- Knowing how we anthropomorphise → knowing how to trigger anthropomorphism (know → exploit)
- Anthropomorphism by design: include design elements that encourages anthropomorphism.



Possiamo vederlo come un processo: così possiamo studiarne le leggi, il funzionamento e capire come influenzarlo, gestirlo per ottenere degli obiettivi. Possiamo studiarlo per riprodurlo. Sapere come antropomorfizziamo le cose ci fa capire come poterlo sfruttare per certi scopi.

Conoscere una cosa rende possibile sfruttarla a proprio vantaggio; potremmo scoprire che antrop. ci può aiutare a raggiungere degli obiettivi. Sfruttare antrop. per raggiungere *design goals*.

Reasons why anthropomorphic design is pursued:

1. In some cases, anthropomorphic design is inevitable. **Our environment is designed in a human-centric way**: it is shaped for us to live in it. If we want robots to blend in, they must be endowed with human-like capabilities. When the env is shaped precisely for us, if you want a robot to blend in it you must make it anthropomorphic. It is not a matter of how the robot look.
2. Has to do with the way in which device is perceived by the users.  
Anthropomorphism boosts **attachment** and **likeability**. → users become more open → devices are easier to understand and so to interact with.  
People develop **emotional bonds** with human/animal-like things. Consumers positively welcome anthropomorphic-zoomorphic products. → this has a lot to do with animations (Disney cartoons for ex.).

Anthropomorphism triggers **sympathy** and **empathy**, which are very important elements in social interactions.

### RESEARCH FIELDS

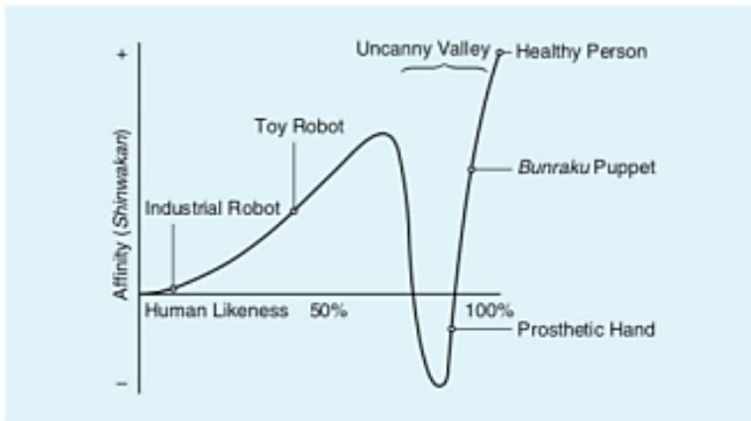
- **HRI** (Human Robot Interaction): multidisciplinary research field that studies interactions between humans and robots. Specific prospective focused in interaction in order to frame and transform it in effective and successful interaction.
- Closest relative: **HCI** (Human Computer Interaction) – multidisciplinary research field that studies interactions between humans and computers (interface and software design). Find regularity in human behaviour w.r.t. computers, useful to improve these technologies.
- **Social robot**: intelligent machines specifically designed to entertain social-like relationships with users (also/exclusively). **Implement social qualities** in AAs to improve interactions with users, in many different contexts (hospitals, home, kids, sexuality, ...).

Zoo/anthropomorphic design facilitates, simplifies, boosts HRI and provides engaging interactions and enhances familiarity/acceptability. These are the main reasons why antrop. design is pursued.

## NEGATIVE ASPECTS

Anthropomorphism in robot is very effective and powerful but... it has to be handled carefully!

Masahiro Mori, The Uncanny Valley (1970) → shows some psychological limitations.



**Figure 1.** The graph depicts the uncanny valley, the proposed relation between the human likeness of an entity, and the perceiver's affinity for it. [Translators' note: *Bunraku* is a traditional Japanese form of musical puppet theater dating to the 17th century. The puppets range in size but are typically a meter in height, dressed in elaborate costumes, and controlled by three puppeteers obscured only by their black robes (see front cover).]

There is a similarity between human and robots that actually is an obstacle for the interaction! After a certain degree of similarity, passing a certain threshold, antrop. becomes an **obstacle**. Human user perceived the robotic entity as a human that cannot be alive (different from him/her) and this generate a feeling of anxiety and he/she rejects it.

## CONCLUSIONS

Antrop. can become also confusing and disturbing:

- [Boston Dynamics Testing Robustness](#)  
Se usiamo il modello mentale sbagliato, non adeguato al contesto, ragioniamo sul robot in modi che non sono i più appropriati. Way of framing the issue that might become ethically problematic. We develop expectation that might led us to unethical consequences.

... but also funny!

- [DARPA challenge - Robot fails](#)

## REFLECTIONS

- Antrop. is a tool that we as designers can use to generate emotional reaction in users → ethical responsibility!
- How you as a person are affected by antrop. in product that you use every day.

## LECTURE 12: DEBATE ON ANTHROPOMORPHISM

Today's main aims

- Understand specifically anthrop design is useful
- In what sense anthrop design might be harmful and not help us to reach design goals. It depends on the goal and to the extent to which we want to apply anthrop.

Frame both issues from an ethical perspective and explore advantages and risks (ethically and socially speaking).

### SUMMARY OF PREVIOUS LECTURE

From the perspective of the user there are some elements that obstacle the development of interaction.

Humans' positive attitude towards social robots: pivotal.

Elements that hinder likeability: gross failures, communication hurdles, strangeness... These bring the user to distrust the system.

How to enhance HRI? This is the most important problem. We can do this in two ways.

- Endow machines with social skills. Isolate what social skills are important for the product, find a way to translate them in computer language and implement them in the program. The machine will display the same conditions displayed by other humans → create interaction and relation. But this is a difficult strategy.
- New domain of issue: not implement features but trick users into thinking that machines master social skills or are endowed with social characters. Here we are interested in the mental model of the user's mind. It is all about fiction. If the user is truly convinced that the machine actually displays social skills necessary for a good interaction, then a good interaction occurs.

Two very different options.

### SOCIAL ROBOTS

Most effective anthrop. factors to be reproduced to conduce the user to an effective interaction and relationship with the robot:

- Face, voice, gesture, body language (nodding, shaking of the head), physical embodiment, presence.
- System autonomy. If the AA display truly and good autonomy (without supervision and control by human) the user is comfortable in rely and trust it.
- Natural-like motion (otherwise it will appear strange).
- Natural language usage and conversational skills.
- Facial expressions and eye contact.
- Emotions (expressing and understanding). If the machine is able to grasp this kind of unexpressed and implicit information, it will accomplish a better interaction.

- Social skills: taking turns, mimicking human ways of talking, inflection, small talks, attentive cues (the machine display attention, shows that it is paying attention) ... Again, important to create a natural relation with the user.<sup>13</sup>
- Typically, human traits such as making mistakes (!). Not trivial mistakes but those that are very insignificant in the functioning. This build trust! The quality of the relationship improves. Intuitively this is strange, but it actually works and helps in build reliability.
- Robot personalities or identities. Elements taken from the human interaction domain; we want to reproduce them for a better relation with the robot.

Example: [Google Duplex](#).

It is actually a passed Turing test!

Observation: the "ehm" is inserted to mask processing time of the model

### DESIGN ISSUES

Main challenges:

- Balance between comfort in use and human likeness (this one must not be used *per se* but always in relation to the goal and context).
- Balance between functionality and likeability.
- Balance between user expectations (what design choices suggest to the user), functions executed and actual capabilities of the technology. If the functioning seem simple, the user will apply his/her entire mental model to the robot and so expectations are very high.

→ **Optimal anthropomorphism:** anthropomorphic design that clarifies the functions carried out by the robots, specifies expectations, and boosts usability. → these are the specific goals we want to accomplish in antrop. design. We want the user to develop a good mental model.

**Antrop. is not a value in itself** for design! It must not be used *per se*: **it is a mean to an end**. We adopt antrop. **strategies in order to** get something specific, i.e. **create a relationship between users and robot in which users have a good mental model w.r.t. it.**

### SOCIAL & ETHICAL ISSUES

Now we shift from a design perspective where antrop. is just an asset to reach goals, to an ethical one where we evaluate the impact of these design practise onto users.

Antrop. now in relation not to design goals but to another goal: making sure that our technology are acceptable from an ethical perspective.

First issue: **deception**. Anthropomorphism is actually deceptive: we want to assume deception in order to reach specific goals, to influence the mental model of the user in order for them to behave in specific ways.

But in case we don't want to influence users directly in this way, and instead we implement social skills into the robot it is hard to claim that the robot actually has this social skills as a human have.

---

<sup>13</sup> We have a mental model that link facial expression to the corresponding emotion. This mechanism are studied to be reproduced and projected in robots (for ex. in the eyebrow movement and position)

Depending on what you think about deception, from an ethical perspective this might lead to different positions.

### 1<sup>ST</sup> POSITION: ANTHROPOMORPHISM IS UNETHICAL

Anthropomorphism is deceptive → so anthropomorphic design is **unethical**. No matter the circumstances or the aims we are trying to achieve using deception.

Deception is always unethical because it has *per se* an harmful impact on something that we think that is valuable and must be respected no matter what. Values that are harm according to this perspective:

1. Fails to respect users' **dignity**. You are "using" users to accomplish so goals so you are reducing them to mere tools, they are no more people. It is kind of an offense to their self-determination. These perspective does not allow anyone to use antrop. for a greater good: the fact that you are not respecting dignity *now* is more important that a goal that might be beneficial *in the future*. This kind of reasoning works in this way.
2. Fails to respect users' **cognitive rights**. Since antrop. has to do with the development of "good mental model" of the users w.r.t. robots, so it is an obstacle to the cognitive status of the user because he/she is strict into thinking that something is what it is not. Users have the right to be put in the position to develop the most adequate model of artifacts.<sup>14</sup>
3. Fails to treat users **fairly** → fairness is another important value. When you deceive someone through design you can **influence the behaviour of the user** with the technologies: **the designer has a position of control → has a role in R. → R. includes no deception** (of any kind). The presupposition of this argument is that **users do not have the means no react** this kind of deception. So you are actually exploiting a weakness of users and doing this means **taking advantage from a position of dominance** and this of course is unethical.
4. **Patronises** or **infantilises** users. → most well-known ad used argument. Paternalism is when a **small group of people decide** what is valuable, ethical, etc and what is not, **and impose this set of beliefs and values** to others. They do this kind of imposition because they think that others do not have necessary requirement to decide by themselves (asymmetrical relationship → members are not equal). So deception can be seen as a from of paternalism: asymmetry between designers and users. There is the risk that designer choose what is right or wrong and impose it to users.
5. Provides only the external aspects of what it imitates, depriving it of its *essence* [e.g. pet robots or robot carers].

In this first position → **Deontological approach** (*deon* = "duty"): **duty, rights, unconditional values, respect.**

---

<sup>14</sup> We saw that "optimal antrop." point is exactly this. SO perhaps this second point is an objection for suboptimal forms of antrop. (while the first -dignity- applies also to this optimal antrop.).

### PROBLEM

It is not necessarily true that deception is always harmful. In some occasion this can be desired by the user: he/she is in control of deception (for ex. when we go to the theatre/cinema and participate to what happens in the fiction but at the same time we are aware of it).

Problem: is there a clear line between “as-if” ascription of anthropomorphic qualities and actual ascription thereof? Is there a difference between projecting human or animal qualities on a robot in a fictional way or just ascribing these qualities to the robot itself?

**“There are relevantly different results between abstract considerations and actual behaviour”** (Fussell et al., *How People Anthropomorphize Robots*, 2008).

This kind of ascription seems to be correlated with **Vulnerability** vs. **Acceptability** of the users: people in need of social contact (vulnerable: elderly, children) are found to anthropomorphise more readily.

### 2<sup>ND</sup> POSITION: THE ETHICS DEPENDS ON THE CONSEQUENCES.

Entirely different position from an ethical perspective: do not focus our attention to antrop. design *per se* (attention to which values must be respected no matter what), but focus our attention the consequences of antrop. design and see how the use of antrop. solutions might lead us to consequences and results that are good in an ethical perspective. We now frame antrop. design as a tool to reach goals.

**Even though anthropomorphic design is deceptive, it doesn't mean it is unethical. The ethics of anthropomorphic (deceptive) design depends on its consequences.**

This is a more flexible perspective. Explore argument with a more open mind:

- Deceptive design may maximise the users' well-being, welfare, happiness (the final goal is important so we want to find a balance between them and the deception).
- In the long run, non-anthropomorphic design may generate more harm than good.
- Anthropomorphic design may help to reach shared social goals.
- Fictional experiences sometimes are morally productive.

Users are not in an inferior position w.r.t. designers: they are perfectly aware that it's a fiction and they *want* to participate in it and take the advantages that comes from the fiction itself.

In this position → **Consequentialist approach**: trade-offs, predictions. Whether something is ethical or not depends on the future of the actions.

## SUMMARY AND CONCLUSIONS

Comparison between two opposite ways of reasoning:

<u>Deontology</u>	<u>Consequentialism</u>
<ul style="list-style-type: none"> <li>• Rigid, abstract, precautionary.</li> <li>• Provides clear limitations, red lines and strong guidelines.</li> <li>• Issues: consistency and definition.</li> </ul>	<ul style="list-style-type: none"> <li>• Flexible, concrete, proactive.</li> <li>• Uncertain and arbitrary since it is based on <u>future</u> (hypothetic) consequences.</li> <li>• Issues: value definition.</li> </ul>

→ practical judgement needed!

## OPEN ISSUES:

1. Ethics of Self-Deception: is it always unethical to engage in fictional experiences?
2. Can we really control anthropomorphism, as deeply wired, complex and context-related a phenomenon as it is? Can we really exploit antrop. design as an asset or perhaps it is too delicate and we should avoid to use it at all?
3. Are there other social initiatives we should explore (and fund) before turning to social robots? From a strictly social perspective, is it correct to invest social money in it? Or should we explore more "traditional" methodologies?
4. How does anthropomorphic design impact on user social expectations regarding AI & Robotics? Responsibility, rights, resource allocation, and so on.

We design antrop. robots → we influence the mental model of the user → this has consequences on the technical but also social understanding of the artefact → we become more open to the idea of ascribing R. to machines that looks like humans. Just changing the external aspects we can push people to think that the machine can reasonably be responsible, and to act accordingly.

Antrop. for some reasons can also be confusing and can lead to issues:



## LECTURE 13: AI ETHICS

### SUMMARY OF PREVIOUS LESSONS

- What is an autonomous agent: *autonomous, purposeful*, *intelligent* artefacts → delegation/loss of direct control → issues of responsibility.
- Artificial Agents directly responsible, morally and/or legally? Not very convincing.
- Only human beings are literally responsible agents? It seems so.
- Still: responsibility gap. How is it to be bridged? Open question.
- We must take responsibility: what does it mean? What does it entail?

### THE ETHICS OF AI

All the questions about responsibility, anthropomorphism, ecc... can be summarized in one big question:

**What can we do to keep AI ethical?** → **Ethics of Artificial Intelligence = the ethics of//for the humans who build AI systems.**

When we talk about AI ethics we are talking about what *humans* should do in order to keep AI ethical → **focus on the human element**. This focus is exactly what difference AI ethics from and Machine ethics (lecture 14).

Similar theoretical and practical fields: Computer Ethics, Robot Ethics, Roboethics Nanoethics and so on → **applied ethics** (main aim: **keep technology aligned to moral values**).

### APPLIED ETHICS

AI Ethics is a form of Applied Ethics. What is Applied Ethics?

New trend (from Sixties): **concrete issues** VS **ethics as a purely theoretical & abstract discipline**.

Moving from an abstract to a concrete level → social practices, professions, ... → Rediscovered the ethical side of professional and social practices connected to applied ethics.

Challenges the distinction between ethical theory and practice. There is a strong connection between them and if it is lost, both becomes useless. So it is meaningful to analyse from an ethical perspective some practices only if you really use theoretical construct and, theoretical constructs can be built only in relation to the analysis of some practical case (a sort of feedback loop).

Also the **target** of applied ethics changed: practitioners, laymen, society (not just philosophy professors).

Also **language** and **values** have changed.

**Aims** becomes more practical: design strategies, problem-solving procedures, easy-to-follow techniques + offer concrete advice/guidance to people on the field. → face and manage concrete moral problems.

Examples of applied ethics: Bioethics, Medical Ethics, Environmental Ethics, Animal Ethics, Business Ethics, Engineering Ethics and so on.

### DIMENSIONS OF APPLIED ETHICS:

- Academic (the original one): research (articles, books, conferences).
- Social (new): foundations and research centres.
- Political (new): expert committees and public organs.



### MISSION OF APPLIED ETHICS:

- Definition of fundamental principles.
- Guidelines for their applications.
- Promoting, incentivising & monitoring activities and support initiatives (social and political side).

### TECHNOLOGY AND ETHICS

Common trait in all applied ethics: the role of technology.

Technologies:

- Transforming impact on social practices
- Wide-ranging effects
- Ubiquitous presence (there is *always* technology in our life)
- New powers and possibilities (give us tools for doing new things in new ways)

→ Ethical reflection needed → that's why we need applied ethics.

### DESIGN ISSUES

We used to think design only as a **technical aspect**: "Is the artefact efficient? Does it execute the function it is built for? Does it do this in the best possible way?" → Important criteria = *Efficiency, Effectiveness, Reliability*.

BUT the more we get entangled with technologies, we find out that technical aspect is not enough. There are many other sides included in design that are important!

**Social aspects:** Is the artefact's function aligned with the general purposes pursued by society? Is it safe? Is it legal? Does it respect habits and shared beliefs? → *Social acceptance, trustworthiness*.

**Ethical aspects:** Is the artefact's function aligned with the shared beliefs about what is good and what is evil? Does it enforce our values and minimise ethical harm? → the *Good*.

Engineering activism:

«Engineers (...) face an unfamiliar obligation to perceive not only the usual set of properties that the systems they build or design may embody, but those systems' moral properties as well: bias, anonymity, privacy, security, and so on. The challenge of building computer systems is transformed into a forum for activism—engineering activism. Not only is such activism a calling for which many may feel unfit, it is also a difficult one»

(Helen Nissenbaum, *How Computer Systems Embody Values*, 2001.)

It has become a hot topic in software engineering: products that embody values of their stakeholders. This becomes a problem when they have different beliefs, different ideas of what is good and what is bad (ex. about privacy). → problem: **alignment** (a unique solution that satisfies all different beliefs).

### (NOT JUST) DESIGN ISSUES

Is AI Ethics only a matter of designing ethical technologies? NO.

Ethical Design → Ethics ok but also **Policy** and **Regulation**. Why?

- Addressing issues of *delegation* and *use*
- Defining social aims and ethical values
- Providing frameworks for ethical innovation
- Providing clear and viable *regulative guidelines*
- Establishing technical standards and auditing mechanisms
- Enforcing users' awareness and education
- Incentivising good AI, penalizing "bad" AI
- Supporting stakeholders towards Ethical AI

### AIMS OF AI ETHICS:

1. Study the ethical impacts of AI technologies
2. Devise principles for good AI
3. Design strategies, measures, and policy to reduce harm and maximise benefits from AI (practical and easy to follow functional indications to how to translate principle in practice).
4. Provide guidance to practitioners, stakeholders, lawmakers, and the like

→ Ethics for *human beings*

### IMPORTANT INITIATIVES

- [Oxford Internet Institute](#) → Director: Luciano Floridi.
- [The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems](#) → Chair: Raja Chatila.
- [High-Level Expert Group on AI](#) → 52 members (they have proposed a document that becomes one of the most important so far: the *EU AI Ethics Guidelines for trustworthy AI* – see below).

### IMPORTANT APPROACHES

- [Ethically aligned Design](#)
- [AI for Good](#)
- [Responsible AI](#)

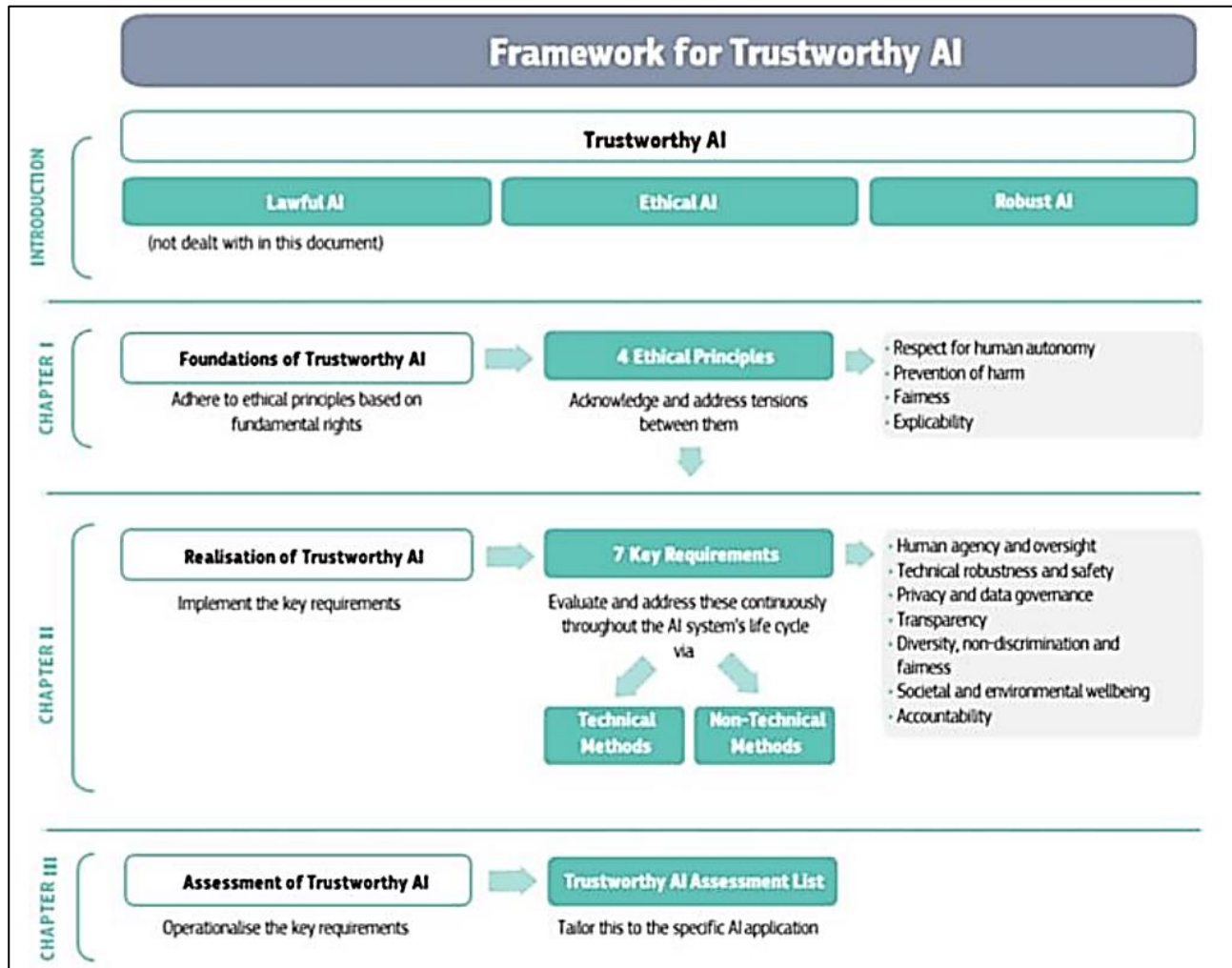
### AI ETHICS PRINCIPLES

1. [Asilomar AI Principles, Future of Life Institute](#)
2. [Montreal Declaration for Responsible Development of AI](#)
3. [IEEE Ethically Aligned Design](#)
4. [EU Statement on AI, Robotics and Autonomous Systems](#)
5. [AI in the UK: ready, willing, and able?](#)
6. [The Tenets of the Partnership on AI](#)
7. [EURON Roboethics Roadmap](#)
8. [EPSRC Principles of Robotics](#)

→ So many principles! How to comply with all of them?

Simplification is needed → *EU AI Ethics Guidelines for trustworthy AI*.

Scheme of the content of the document:



Main focus: the value of **trust**.

Ch.1: How to carry out trade-off between the 4 principles.

Ch.2: Translation of principles in requirements. Alignment between artefacts and these requirements should be addressed continuously.

## THE ETHICS OF AI – ISSUES

THEORETICAL ISSUES:

- Gap between principles and recommendations or policies
- Gap between universal principles and specific cultures

PRACTICAL ISSUES

- Proliferation of so many principles: AI4People group found no less than 47 principles in literature!
- **Ethics washing**: exploit ethical discourse for marketing (whether a company makes ethical choices or *marketing* choices - selecting some values that you know your public might like and present them as ethics).



### AI ETHICS EPIC FAIL

Google's Ethics Panel: ATEAC (Advanced Technology External Advisory Panel). Shut down after a week.

Why?

Members:

- CEO @ [Trumbull Unmanned](#) Interest in AWS
- President @ [Heritage Foundation](#) Conservative Think Tank Vs. LGBTQ, immigrants...

### ETHICS WASHING

ATEAC: Just a public relation operation? Unpaid positions, no veto right (toothless), only 4 meetings scheduled for 2019. → so this group was just a marketing operation.

Same considerations on AI\_HLEG: it released the final version of *Ethics Guidelines to Trustworthy AI* after a draft document was released for the public to comment upon. They collected comments and tried to address the issues raised in the comment in the final version. And there were some problematic changes:

- "red lines" (hard expression: it gives you the idea that ethic really matters) → deleted or changed in "critical concerns".
- "non-negotiable" → completely removed.

Why? Probably pressures from industry representatives. Industries were in fact highly represented among the group. HLEG Composition: 4 ethicists + 48 non-ethicists, and among those 48 the majority were industry representatives.

See articles by [VOX](#) and [Thomas Metzinger](#).

---

## LECTURE 14: MACHINE ETHICS

Artificial agents are not immoral (i.e. evil) but non-moral: they are not able to handle moral informations, morality is not a dimension they take into account.

Main aim: **transform AAs in good moral agents**. Build good robots. → very technical and practical matter.

Example: [self-driving cars](#).

### LITERATURE REFERENCE

The **idea of machine ethic belongs to the literature domain before it was assumed in computer science!**

→ Isaac Asimov's, *I Robot*: he proposes 3+1 Laws of Robotics.

- 0th Law: A robot may not harm humanity, or, by inaction, allow humanity to come to harm. → **before the others because it's the most important**
- 1st Law: A robot may not injure a human being, or, through inaction, allow a human being to come to harm. → **protect human beings**
- 2nd Law: A robot must obey orders given it by human beings except where such orders would conflict with the First Law. → **assure that delegation is effective**

- 3rd Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. → **conservation of the robot**

The story of the book is about the robot that cannot act without *contradict* these laws → **coherence** in machine ethics is very important.

Imagination is a really important aspect for technological design.

### AI ETHICS VS. MACHINE ETHICS

Remember that they are different!

- **AI Ethics**: of/for the *humans* who build AI systems.
- **Machine Ethics**: ethics of/for *machine* functioning.
  - Building machines that reproduce ethical behaviour (what does it mean *reproduce*?)
  - Implement moral skills into AI systems (exactly like we implement other skills)
  - Represent moral experience computationally (find a possible representation in computational language of what moral experience is).
  - Teach robots right from wrong (Wallach and Allen, *Moral machines*)

→ but the most important notion when we talk about what machine ethics is trying to accomplish is **delegation**. What we are trying to achieve is to delegate (partially or entirely) moral choices to autonomous agents.

Beware! Terminology is not consistent: AI ethics, ethics of AI, robot ethics, machine ethics, machine morality, artificial morality, artificial ethics, and so on.

Why do we need machine ethics (in addition to AI ethics)?

There is an observable trend in AI research towards **higher degrees of autonomy and flexibility**.

**More autonomy = more delegation** → wider array of tasks can be delegated to machines.

Some of these tasks **require moral agency/judgement**. When these tasks are delegated to machines, they still require moral agency to be adequately carried out.

→ We need to endow artificial agents with *moral subroutines*

### ARTIFICIAL MORAL AGENCY

A proposal of classification of Moral Agents (James H. Moor, *The nature, importance, and difficulty of Machine Ethics*, 2006):

- **Ethical-Impact Agents**: products whose introduction has important ethical effects (non-necessary a digital technology)
- **Implicit Ethical Agents**: products that incorporate fix elements specifically designed to avoid ethically undesirable outcomes (ethics now is directly implemented in the artifact). Ex. an algorithm that avoid discriminations.
- **Explicit Ethical Agents**: products that imitate human moral reasoning by representing ethical categories, perform analysis, compute ethical judgements and justify them.
- **Full Ethical Agents**: ethical judgment, autonomy, intentionality, consciousness, free will (standard: human beings).

At least by now, machine ethics has nothing to do with Full Ethical Agents: the aim is find a way to build an implicit or explicit ethical agents.

### THE MACHINE ETHICS PROJECT

Tasks:

- imitate human moral agency by means of computer science/robotics → tools: *behaviorism*, *analogical thinking*, etc
- elaborate a computable model of human moral agency. We don't want a copy, an imitation, of human morality but a methodology that associates inputs and outputs in a similar way as humans do. And there two main approaches to do so (see below).

- James H. Moor, *Is ethic computable?*, 1995 → **not if but how.**

### DESIGN STRATEGIES

<u>Top-Down Approaches</u>	<u>Bottom-Up Approaches</u>
Based on symbolic AI (GOFAI).	Based on Machine Learning techniques.
<b>Pros:</b> secure-predictable, transparent.	<b>Pros:</b> flexible, trainable.
<b>Cons:</b> rigid, demanding.	<b>Cons:</b> unpredictable, opaque.

→ How to balance both approaches? **Hybrid Approaches/Modular AI architectures** (but it's speculative)

### MACHINE ETHICS ISSUES

1. **Implementation:** how can we implement values in algorithms (technical)? What is the best way to translate values in computational language? How does this feedback on our understanding of human ethics (philosophical)?
2. **Value Determination:** which values should be implemented in which technology? Why? Who should decide this? How to avoid ethnocentrism, paternalism, technocracy(->the fact that which value is important is decided by technicians)
3. **Moral Deskillling:** if we delegate the management of moral situations to machines, how will we learn to be ethical? If **moral behaviour comes from exercise and experience**, delegation might lead to moral dependence and inadequacy.

## LECTURE 15: NEURAL NETWORKS AND MACHINE LEARNING

There is a lot of **confusion surrounding the many terms** and scientific definitions about NN and ML. These are the technologies towards which we experience a mix of feelings: we are fascinating but we also feel obsolete w.r.t. them.

These are the **technologies that mediate the feeling between human and AI**: what people feel about AI is regulated and mediated by ML and all these concepts. When we talk about NN and ML, that **the most difficult thing is to build a bridge between technological expertise and knowledge and what people actually understand about this technologies** (what is their mental model, which is not based on an understanding of the knowledge behind but on the way on which these technologies are commonly presented).

Is there something that these technologies can't do?

Technical questions that are asked by politicians, policy makers, regulators, ... to computer scientists. Ex. what is a NN, how does it work, what it can do. These problems have been extended to a larger part of the population because digital now is a common dimension. This is no more something specific for computer scientists: **is a feature of the relationship between expert knowledge and common sense**.

**The roleplay of experts in this scenario is fundamental: they have the *responsibility* to translate their expert knowledge in a way that is clear and understandable to general people.**

Es. make the concept of NN more understandable → analogy between the code and the neural structure (Analogical Thinking is a very strong tool for communication).

It is difficult to explain to people that algorithms are not always right because of statistical errors: they can produce false positive or false negative results. So a machine is not a calculator: not always exact and reliable answers. We have to be critical concerning to a machine's results.

"Neural Networks are Black Boxes": opacity, unpredictability, unexplainability.

**NN today are the best example of AAs.**

Data are very important. There are social and ethical impacts on the way in which we build and use datasets. There are other problems related to data: who own them, who produces them, privacy, consent, ...

Big Data, Big Issues:

- Bias
- Discrimination
- Delegation
- Transparency/Explainability
- Privacy
- Consent
- Manipulation
- Exploitation
- Ownership

## LECTURE 16: BIG DATA ETHICS

Three problems:

### BIAS

Bias = **prejudice, opinion, preferential belief** or tendency that normally lacks rational or logical or scientific support. NO rational and logical reasons to justify this believes (lack justifications). So it exists in human psychology: is not a feature that belongs to NN.

Implicit or explicit: you might (or not!) be aware that your argument or behaviour is subject to bias.

Examples: instincts, cultural preferences, discrimination on the base of skin colour, race, gender, ...

But bias is not necessary a negative thing in human: actually it is a “gift” of evolution. There is a reason why our mind works in this way. Bias is a practical shortcut for example to handle situations that are very complex. Some biases are good because they help us to navigate complex environments.

Typically human character: evolutionary advantage, quick (and often appropriate) responses in situation of potential dangers or high complexity. In every situation we are embedded in some kind of context which influences us in the form of pre-existence knowledge: it's a non-avoidable precondition. There is no new beginning. This is something given from the environment.

In the form of pre-existing knowledge is a *necessary* condition of any action. What is more, since human beings are historical beings (philosophical movement of existentialism), i.e., always embedded in historical cultural context, bias is a necessary condition of existence, not just an error.

Biases are just there, we can't eliminate. Should we do something about them?

We should distinguish between biases we should not worry about (because they do not lead to any problematic effects) and biases that might be harmful.

Four types of biases:

1. **Neutral Bias**: bias that does not have any worrisome effects and, as such, do not particular issues (e.g., we associate pleasantness to flowers more often than to bugs).
2. **Positive Bias**: bias that serves desirable ends such as selfpreservation (e.g., fixed reactions to particular perceptions) or enhancing self-esteem (self-serving bias). Practical shortcuts to achieve goals that we value for some reason.
3. **Veridical Bias** (or true opinion): bias that we cannot justify but still accurately represents a state of things.
4. **Problematic Bias**: bias that leads to harmful behaviour, discriminate people on unethical basis, disrespects human dignity... → *Ethical task: reduce them.*

Biases play a role in determining human thought and action → Human agency *projects* biases → **Human artefacts (technologies) incorporate biases** - e.g.: language corpora contain biases (semantic biases influence statistical contexts of words).

Technologies contain biases.

Data are collections of human-generated contents (immediate or mediate) → Data are human artefacts → Datasets incorporate bias! → Neural Networks are trained on data → Data are biased → **Algorithmic Bias (transmission of the bias from data to the model).**



Another notion of bias linked to Information Theory: before human interpretation bias is the lack of entropy, meaning that data is not all equal (in language, character and words have not the same probability); how data deviates from a meaningless uniform random distribution. So bias is exactly what makes data valuable: it gives probabilities, connections, relationships between data. Bias is the valuable part: we don't have to avoid or eliminate it in this context but find it, exploit it in the best possible way.

So we give to it different names depending if we want to give it a bad or good connotation: we call it "information" or "knowledge" if we want a positive connotation or "bias" if we want a negative one. But it is the same phenomenon.

If data contain problematic biases → **"unethical" algorithms**: unethical discrimination, prejudiced outcomes, spreading inequality, offense to human dignity, reiteration of unethical patterns, and so on.

Avoid this situation not only for ethical and social reasons but also for financial reasons: a company having this kind of algorithm would be in trouble.

**The fact that this bias is contained into data should not justify an eventually bad output.**

Countermeasure:

- Mathematical formulations of non-discrimination in decision-making.
- Modular AI architectures: mix implicit learning of statistical regularities + explicit instructions or rules of appropriate conduct (a way to ensure that what is taking into consideration from the machine is socially acceptable and aligned with our moral sensitivity).

But also: NNs can help *detecting* biases. Not by themselves, but in cooperation with human users. Why? Not all biases are problematic, it depends on the culture. Only with cooperation with social and technical reasoning we can build a technology on which we can rely entirely on.

Whether a bias is problematic or neutral is an ethical and social issue. The "wrongness" of a bias cannot be calculated absolutely but is always **relative to the cultural context** in which it is active. For example, a NN trained with data collected in Italy may not be acceptable in Japan.

Framing an algorithmic bias as a *problematic* bias cannot precede but only follow its ethical and social determination → logical primacy of the social over the technical. We need to put social and ethical reflection first and then the technical practices.

Different sources of Bias:

1. **Data-driven bias**: incomplete/incorrect/poorly labelled datasets → technical dimension.
2. **Bias through interactions**: learned from interactions with other users or agents → technical and social dimensions mix.
3. **Similarity bias**: echo-chambers and filter bubbles (ex. recommending algorithms).
4. **Conflicting goals bias**: narrow AI creates negative consequences in lateral or secondary applications.
5. **Emergent bias**: reinforce prejudices and questionable behaviour that are *already* in society.

## DELEGATION

Fact: NNs outperform humans in many tasks that involve *recognition*.

**Recognition:** very wide concept. Includes *subsumption*: classify a case under its proper class.

Traditional performance of *practical judgement* (*phronesis*) → recognition cannot be executed by following rules but it requires something else:

- implies logically non-representable or ill-defined features such as: expertise, talent, intuition, tact, genius ... we don't have a way to formally represent what goes on, therefore we use notions that are not precise but helps us to indicate the phenomenon and study it.
- This performance is commonly attached to socially highly esteemed individuals: the sage (ethics), the judge (verdicts), the doctor (diagnoses), the politician (choices for the common good).
- Displays a relevant social and ethical status: personal engagement and responsibility.

Problem: can NNs *substitute* humans in domains where practical judgements are required?

Fact: **full delegation of practical judgment to machine is particularly resisted in:**

- **Life/death scenarios:** healthcare, warfare, death penalties in court judgments...
- **Life-changing scenarios:** mortgage assignments, hiring, firing, criminal court judgments...
- **Ethically loaded scenarios:** care, companionship, warfare, resource allocations...

→ Why?

Because:

- On a technical level: *success rate*. If NNs success rate > human experts' success rate, why not? There is no rational reason, from a technical p.o.w. to resist full delegation.
- On a social and ethical level: we realize instead that something is missing. We attach values (talent, genius, ... but also engagement and responsibility) to those who are able to perform practical judgment in a good way. When they are substituted by a machine, all the social dimension does not apply any more. What corresponds to human personal engagement and responsibility? *Nothing*.

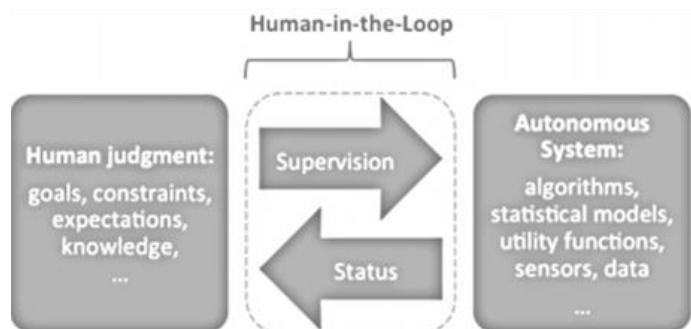
Function of personal engagement and responsibility is correlated to the idea of *respect human dignity*.

Politicians, doctors, ... that behaves responsibly show respect for those who depend on their judgement → they address social and ethical needs. Such needs remain unaddressed when full delegation to NNs occurs.

→ delegation to NNs of decision-making that has effect on human dignity is *ethically troubling*.

Solution: keep humans in the loop → **from substitution to cooperation**: NNs as useful tools in the hands of human practitioners → both 'technical' and 'social' needs are addressed.

What does this mean in practice? Hard to figure out. *Taking* responsibilities for NNs-mediated decisions?



## BLACK BOXES

ML: algorithms autonomously determine how to process information.

Black Boxes: input is known, output is known, information processing is unknown (unknowable?).

We do not have *control* on information processing and decision-making carried out by NNs.

We do not have *proof* that algorithms 'reason in the right way' or 'look at the right thing'.

Problematic: correlation is not causation.

### Case study: Wolves vs. Huskies

(Ribelio et al., *Why should I trust you? Explaining the prediction of any classifier*, 2016)

Data set: photos of wolves with snow in the background and photos of huskies with no snow in the background.

A NN is trained on this data set and works fine; NN can also provide explanation.

Explanation: NN had classified images on the basis of the snow on the background.

→ Data set is important. → Correlation is not Causation.

Issues:

- Can we reasonably rely on NNs whose functioning we cannot doublecheck or control?
- How opaqueness, unintelligibility and unexplainability would impact on user trust?
- How can we make sure that no discrimination or other problematic biases influence decision-making?
- Does the use of black box algorithms respect the dignity of those interested by the decisions they take?

Fundamental ethical value: **explainability**

- AI must be *understandable* (epistemological need).
- *Accountability* must be clear (ethical need).
- GDPR: right of explanation.
- Other terms: transparency, intelligibility, interpretability.

Can NNs comply with explainability? Open issue.

## OTHER ISSUES

- Privacy
- Confidentiality
- Transparency
- Autonomy and manipulation (recommending algorithms)
- Informed Consent
- Data Ownership
- Surveillance and security
- ...

### III. THIRD PART: PRACTICAL CASES AND STUDENT PRESENTATIONS

---

#### LECTURE 17: SEX ROBOTS

What are sex robots/sexbots? Robotic products specifically designed to simulate sexual interactions. They can be anthropomorphic or not.

Different degrees of AI: simulation of personality traits, sexual tastes, sexual attitudes. Some of these robots are integrated with learning techniques to memorize users' preferences and behave accordingly in future interactions!

Complex technology → Expensive machines! → Also available for rent (for ex. in Japan).

Sex robots but also robot companions: different *but* potentially integrated technologies. Love and sex aspect can work together. Sexual dimension can be integrated or taken apart. The user can develop a psychological bond!

Example:

**Hardware** → Realbotix [US]: robotic talking heads on RealDoll bodies

**Software** → Harmony AI App: personality, user preferences, sex chats.

Other companies and projects:

- TrueCompanion – Roxxy & Rocky [US]
- Synthesia Amatus – Samantha [Spain]
- Doll Sweet [China]
- Z-onedoll – Silicone Robot [China]
- AI Tech [China]

Which notions and mental models should we use to frame these technologies?

Which notion should be used to understand what intelligent sex robots are? Are they tools or more than that? Should we use the conceptuality that we apply w.r.t. sex workers? Or to sex toys?

Do we need to refer to sex as a practice between human beings or masturbation?

Sex robots maybe can be interpreted in the middle between sex toys and sex workers and prostitute.

There exist rights and laws that regulate sexual interactions between people that of course are meaningless applied to sex toys. But what about sex robots?

Issues:

- Will sex robots have a *beneficial* or *detrimental* effect on society?
- Sexual intercourse with robots is *just* simple masturbation or *much more* than that?
- Consent is important or irrelevant? How much is relevant in this point the anthropomorphic form?
- Robot and Love:
  - Emotional bonds with robots, being unidirectional, are *dangerous* and must be *avoided*
  - Emotional bonds with robots, though unidirectional, may be *beneficial* and should be *permitted*
- Marriages between humans and robots should be legally *prohibited* / *permitted*
- Suppose that you live in a society where paying for sex is legally permitted, morally accepted, and everybody does it. Would you switch from human sex workers to robots? Would it be the same?

Psychological dimension.

People fall in love and perceive their feeling as love with a lot of tools (for ex. with a car), also non anthropomorphic and not so much complex tools. The level of complexity and anthropomorphic behaviour and appearance is not so important in the process of developing feelings toward a robot or tool.

---

## LECTURE 18: SELF DRIVING CARS

*Guest Lecturer: Professor Guglielmo Tamburrini*

---

## LECTURE 19: AUTONOMOUS WEAPON SYSTEMS

One of the oldest debate in ethics about AI.

Why it is a difficult topic: our imagination has been exposed to war robots (it's natural to link to what we knows from movies, comics, videogames, ...). We think to anthropomorphic robots (eg. Terminator) because it is very common in cultural representation.

Reality: AWS has nothing to do with robotic soldiers.

**Automatize some war functions**, NOT recreating a human soldier.

AWS is a label for very different kinds of technologies.

## DEFINITIONS

From two important institutions (two different perspectives, due to different roles of this institutions):

- ICRC (**International Committee of the Red Cross** – institution for our safety so it has a very conservative p.o.w. - ) → Work for peace.

“Weapons that can independently select and attack targets, i.e., with *autonomy* in the ‘critical functions’ of acquiring, tracking, selecting, and attacking targets.”

Concept of **autonomy**: they do NOT need human intervention.

- US DoD (**Department of Defence of USA** - institution directly involved in AWS - ): → Interested in war.

“Systems capable, once activated, to select and engage targets without further intervention by a human operator.”

Again, concept of **autonomy**.

Many definitions exist, but ONE general agreement (inclusive, not exclusive):

Alternative name → **(L)AWS: (Lethal) Autonomous Weapon Systems, or Killer Robots.**

This definition is not neutral, but it adds a **value label!** → Interpretative operation: adding a new information to the name of the technology that suggests a specific interpretation.

From the first definition:

Now we focus on clarify what functions are called “critical” and why.

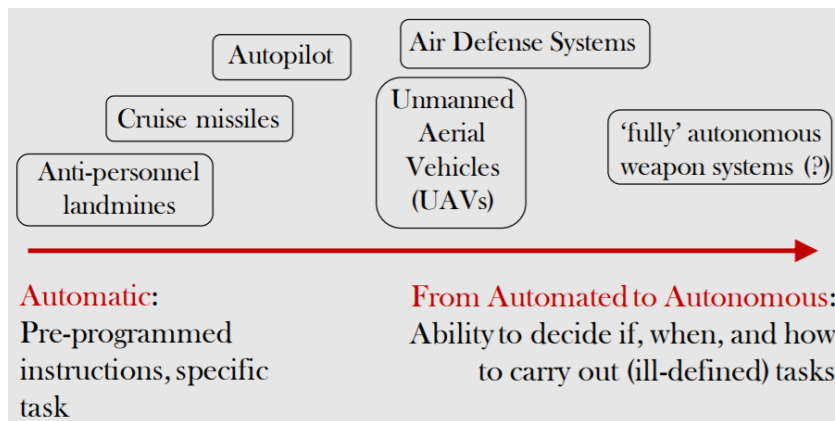
## APPROACHES TO AUTONOMY OF WEAPONS

### 1. FUNCTIONAL APPROACH

<b>Non-critical functions</b> (do not directly put in jeopardy human life)	<b>Critical functions</b> (endangers directly human life)
Mobility → move around a selected area autonomously	<b>Target identification</b> → recognize an element as a target or not ( <u>political aspect</u> : who decides what is a target or not)
Intelligence	<b>Target tracking</b> → follow it during its movements
Interoperability	<b>Target prioritization</b> → chose among different targets (depending on their military importance)
Health management	<b>Target selection and engagement</b> → kill it

We recognize as AWS also weapons that do NOT perform critical functions.

## 2. CAPABILITY APPROACH



## 3. HUMAN CONTROL APPROACH

Relationship between the system itself and the humans that deploy it.

- **Humans in-the-loop** (or, system **semi-autonomy**): once activated, the weapon system identifies and attacks only targets predetermined by humans. The machine executes tasks autonomously but there is always a human in control.
- **Humans on-the-loop** (or, **supervised system autonomy**): the system is autonomous in identifying, tracking, and attacking targets but human intervention and override is always possible.
- **Humans out-of-the-loop** (or, **system autonomy**): total delegation. Strongly opposed by many institutions.

Aim: **balance** between system autonomy and human control.

→ But there is tension between the two: autonomy is supposed to exclude supervision.

It's a technological task: we need instruments to balance them.

Ideal: **Meaningful Human Control** must be maintained on machine functioning. It means in practice delegating all the non-critical functions to the machine while keeping in human hands all that functions that have an ethical meaning.

- How much autonomy is acceptable?
- How can we specify what meaningful human control is?
- How can we maintain control at high level of complexity and speed?

Today's AWS:

- Direct human control on critical functions is maintained.
- Highly constrained in the task carried out (mostly defensive operations).
- Limited types of targets (vehicles, objects—but: 'sentry weapons').
- Limited contexts (simple, static, predictable environments)

→ Developmental trend: increase autonomy.

### ETHICAL FRAMEWORKS

Divided in two disciplines:

- Theory of the Just War
- Ethical way of fighting a war (which values should be encouraged)

This moral and philosophical reflection reflects itself into regulations:

### LEGAL FRAMEWORKS

International Humanitarian Law [IHL]:

- Rule of **Distinction** (military target must be distinguished by civilians)
- Rule of **Proportionality** (the intensity of an attack must be proportional to the purpose of the military operations, in order to keep under control the use of force)
- Rule of **Precautions in attack** (victims among the civilians and damage to their structure must be kept on a very low threshold)

Problems:

- do we need new, specific laws for AWS? Or do we need to just apply those that we already have...
- little information available due to confidentiality and classification.

### POLICY & SOCIAL INITIATIVES:

- US and UK have developed publicly available policies.
- NGOs: ICRC, UN...
- Campaigns: [Stop Killer Robots](#).

### EXAMPLES OF AWS DEPLOYED TODAY:

- Air Defense Systems (oldest kind of AWS): intercept incoming missiles and neutralize it.
- Land weaponry (es. self driving systems with sensors)
- Smart bombs/Loitering weapons
- UCAVs (Unmanned Combat Air Vehicles)

A disturbing simulation by Stuart Russel, Future of Life Institute:

<https://www.youtube.com/watch?v=KqoGacUu07I>



## LECURE 20: DEBATE ON AUTONOMOUS WEAPON SYSTEMS

### ISSUES

Reasons in support:	Reasons against:	Recommendations:	Responsibility Mechanisms:
Technical	Technical	Permitted or banned?	Operators?
Military	Military	Which constraints?	Generals?
Ethical	Ethical	National or International?	Chief of military forces?
Social	Social		Producers?
			Developers?

Moral imagination experiment. How would you feel and what would you think if you were

- An operator of a semi-autonomous AWS?
- A general deploying a fully autonomous AWS?
- A civilian exposed to an AWS in a war zone?
- An enemy soldier of a technologically advanced country exposed to an AWS?
- An enemy soldier of a technologically underdeveloped country exposed to an AWS?
- A programmer or designer in an AWS company?

### UTILITY OF AWS

- Reduce risks and death toll (military and civilian).
- Increase military capability.
- Reduce operating costs and personnel requirements.
- Increase soldiers' efficiency (AWS do not have fear of being killed and resistance to killing).
- Simplicity (one systems, all functions: ultrared vision, flying, shooting, and so on).
- Decrease reliance on communication links.
- Dull, dirty, dangerous, and deep missions will be delegated to machines: no human life involved.

→ interest in increasing autonomy of weapon systems.

### ISSUE 1: TECHNICAL OBSTACLES.

Relevant limitations in respect to the delegated function:

- Low adaptability.
- Low environmental awareness.
- Incapability of complex decision making and reasoning.
- Highly dependency on task specification (not so flexible).
- Unpredictability.
- Fragility: easily break down.
- Heavy reliance on human inputs.
- Lack of standard for testing and validation.

Open problems:

- How can AWS reliably distinguish between combatants and non-combatants or between engaging and surrendering combatants? (mental state ascription problem vs. IHL rule of distinction)
- Could AWS comply with the IHL rule of proportionality and precautions in attack? They require fine judgment or practical judgment (uniquely human? We can build machines that follow our values but not machines that are able to immediately relate to these values).
- How can AWS be properly tested and validated?

→ Challenges: versatility, adaptability, recognition and reasoning skills, predictability, benchmarking...

## ISSUE 2: RISKS.

- Increased unpredictability.
- Cyber-attacks and sabotage.
- Proliferation.
- Sense of injustice and retaliations: terrorism.
- Lowering of the threshold for the use of force.
- Arms race.
- Research and development costs potentially enormous.
- Social backlash.

→ Problem: why invest on full autonomous LAWs instead of sticking to supervised weapon systems?

## ISSUE 3: ACCOUNTABILITY.

Who's responsible for LAWs violating IHL?

- States
- Military commanders
- Manufacturers
- Programmers
- LAWs themselves (punishment vs. fix)?

→ Problem A: is it fair to hold humans accountable for decisions autonomously (unpredictably) taken by AI systems? Accountability gap.

→ Problem B: how can we ensure the possibility of meaningful redress and just punishment?

## ISSUE 4: HUMAN DIGNITY.

- Is it ethical to *delegate* life/death decisions to machines? Common answer: NO. Life/death decision must remain under human control. → strictly refers to the "autonomy" part of weapons.
- Does the use of LAWs respect *human dignity*? Common answer: NO. Respect for dignity commands that no human life is taken through automated decision-making.

Vs. Dehumanisation of war.

→ Problem A: ideally, being void of *emotions*, LAWs may perform better than human soldiers and be more *transparent*. Are emotions just hurdles to 'fair fighting', or their very condition?

CONCLUSION: THE 'MARTENS CLAUSE'.

"Recalling that, in cases not covered by the law in force, the human person remains under the protection of the principles of *humanity* and the dictates of the public *conscience*" (Geneva Convention, Protocol I and II)

---

## LECTURE 21: MACHINE ART

*Guest Lecturer: Professor Mario Verdicchio*