Module 3: AI ethics

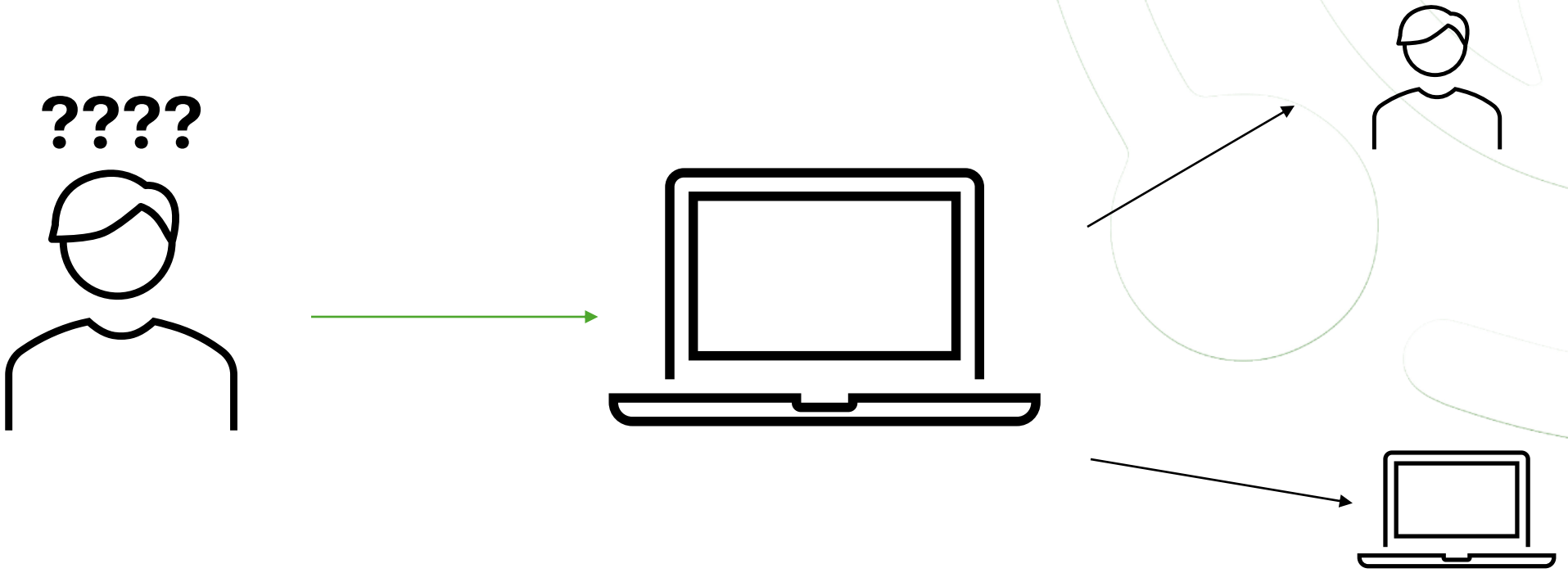
"The real risk of artificial intelligence is not malice, but competence."

Stephen Hawking

AI Ethics

- **Ethics** is the branch of philosophy concerned with judging whether actions are good or bad.
- **AI Ethics** is the branch of technology ethics that specifically focuses on artificially intelligent systems.
- It involves the **creation of a test capable of determining whether decisions made by AI are ethical**.

The imitation game



Purposes of AI

- **Should we give AI a purpose?**
  If so, **what kind of purpose** should that be?
- **How can we define goals** for an AI system in a way it can understand and follow?
- **How can we ensure** those goals are maintained over time, especially as the system evolves or learns?
- **What are the purposes of human beings?**
  And should AI align with them, replicate them, or challenge them?

Friendly Artificial Intelligence

Eliezer Yudkowsky

Artificial intelligence **whose goals are aligned with ours**, based on the principle of **coherent extrapolated volition**.

It means building an AI that does **what we would want it to do**, **if we knew more**, were **more rational**, and had **more time to think**.

AI should help us fulfill our **better, wiser, long-term goals**, —not just our immediate desires or flawed preferences.

Breakdown of the problem:

- Ensuring that AI **understands** our goals
- Ensuring that AI **adopts** our goals
- Ensuring that AI **preserves** our goals

Understanding human goals: solution

Two key problems:

1. **Finding an effective way to encode arbitrary systems of goals and ethical principles** into a machine.
2. **Enabling machines to determine which specific system** of goals or values corresponds to the behavior they observe.

Understanding human goals: solution

- **Inverse Reinforcement Learning (IRL)**
  (Proposed by Stuart Russell)
- This approach expects the AI to **infer something about our goals** by observing the **decisions and actions it takes**.
- In other words, the AI learns what we want by analyzing behavior, rather than being explicitly told.

Adopting our goals: solution

Corrigibility

=

It is possible to give AI a system of goals
that **can be corrected or adjusted** by humans.

Adopting our goals: solution

But are we sure that AI’s goals won’t evolve as its intelligence evolves?

Maintaining goals: the goal preservation problem

- Steve Omohundro and Nick Bostrom argue that we can predict certain **sub-goals** of an AI regardless of its initial goals.
- If a Friendly AI self-improves, can it remain friendly?
- Therefore, it is crucial to clearly define the AI's goals and ensure they are aligned with human values.

Goal alignment: the most important problem

- What are the goals of human beings?
- Four guiding principles:
- **Utilitarianism**
- **Diversity**
- **Autonomy**
- **Legacy**

Human Principles

UTILITARIANISM

Conscious positive experiences should be **maximized** while suffering should be **minimized**.

↓

Challenge: The problem of consciousness — how do we define and measure conscious experiences?

DIVERSITY

A varied set of positive experiences

↓

Has enabled the survival of the species

AUTONOMY

Conscious beings and societies must be free to pursue their own goals.

LEGACY

Ensures compatibility with scenarios that humans consider good.

https://www.moralmachine.net/

AI principles: can human principles align with AI principles?

Six major high-level documents:

- Asilomar AI Principles (2017)
- Montreal Declaration for Responsible AI Development (2017)
- Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems (IEEE, 2017)
- Statement on Artificial Intelligence, Robotics, and Autonomous Systems (EGE, 2018)
- AI in the UK: Ready, Willing and Able? (AIUK, 2017)
- AI Partnership Principles (2018)

AI principles

In 2020, the AI Ethics Guidelines Global Inventory identified **160 proposed principles**

Problem: Overlap and confusion caused by so many guidelines

Overview of the five principles:

1. **Beneficence**
2. **Non-maleficence**
3. **Autonomy**
4. **Justice**
5. **Explainability**

Beneficence

Promote well-being, preserve dignity, and support the planet

- **“The development of artificial intelligence should ultimately promote the well-being of all sentient beings.”** — Montreal Declaration for Responsible AI Development
- **“Common Good”** — Referenced in both **AIUK** and **Asilomar AI Principles**

Non-maleficence

Privacy, security, and capability caution

- It is still unclear whether the people developing these technologies should be encouraged not to do harm,
  or if it is the technology itself that should be prevented from doing harm.
- At the heart of this dilemma lies the issue of **autonomy**.

The power to decide to decide

Establishing a balance between the decision-making power we retain and the power we delegate to artificial agents

- "They must not compromise humans' freedom to establish their own standards and norms." — ESE
- “The autonomous power to harm, destroy, or deceive human beings should never be granted to AI.” — AIUK

Justice

Promote prosperity, preserve solidarity, and prevent inequity

Establishing a balance between the decision-making power we retain and the power we delegate to artificial agents

- “The development of AI should promote justice and strive to eliminate all forms of discrimination.”— Montreal Declaration

Are we (human beings) the patient receiving the "treatment" from AI, which presents itself as the doctor, or are we both?

Explainability

Enabling the other principles through intelligibility and accountability.

Answers the question:
HOW DOES IT WORK?

TRANSPARENCY

The five principles in the six documents

Tabella 4.2 I cinque principi nei sei documenti analizzati e in altri documenti.

	Beneficenza	Non maleficenza	Autonomia	Giustizia	Esplicabilità
AIUK	•	•	•	•	•
Asilomar	•	•	•	•	•
EGE	•	•	•	•	•
IEEE	•	•			•
Montréal	•	•	•	•	•
Partenariato	•	•		•	•
AI4People	•	•	•	•	•
HLEG	•	•	•	•	•
OCSE	•	•	•	•	•
Pechino	•	•		•	•
Rome Call	•	•	•	•	•

Luciano Floridi, «Etica dell'intelligenza artificiale»

Risks

- **Ethical shopping**
- **Ethical bluewashing**
- **Ethical lobbying**
- **Ethical dumping**
- **Ethics evasion**

Ethical shopping

"The malpractice of selecting, adapting, or revising ethical principles, guidelines, codes, frameworks, or similar standards by picking from a variety of available options, in order to give a new veneer to some pre-existing behaviors and thereby justify them retrospectively, instead of implementing or refining new behaviors by comparing them with public ethical standards."

Ethical shopping

Risk of mixing and matching preferred ethical principles, causing incompatibility of standards

+

Risk of reduced competition, evaluation, and accountability

↓

STRATEGY:
Establish clear, shared, and publicly accepted ethical standards

Ethical guidelines for trustworthy AI

In 2021, these guidelines influenced the proposal adopted by the European Commission for an AI regulation, described as the first-ever legal framework on AI.

Ethical bluewashing

“The malpractice of making unfounded or misleading claims regarding ethical values and the benefits of processes, products, services, or other digital solutions in order to appear more ethically sound in the digital realm than one actually is.”

Ethical bluewashing

Marketing Practice

Bluewashing + Ethical shopping

=

A public or private actor acquires ethical principles and publicizes them to emphasize their ethical commitment without producing real improvements

↓

STRATEGY:

Transparency and education

(In the long term, certifications for digital products and services are also expected to be established.)

Ethical lobbying

"The malpractice of exploiting digital ethics to delay, revise, replace, or avoid appropriate and necessary legal regulation (or its enforcement) related to the design, development, and implementation of processes, products, services, or other digital solutions."

Ethical lobbying

Undermines the foundation of ethical self-regulation

↓

And can delay the introduction of necessary regulations

↓

STRATEGY:
Good legislation and effective enforcement

Ethical dumping

"The discontent of (A) outsourcing research activities related to processes, products, services, or other digital solutions to other contexts or locations (for example, from European organizations outside the EU) in ways that would be ethically unacceptable in the original context or location; and (B) importing results of such ethically questionable research activities."

Ethical dumping

Export of Unethical Research Practices

Involves both the export of unethical practices and the unethical import of their results

Ethical dumping may worsen in the near future due to:

1. Impact of digital technologies on healthcare, social services, defense, policing, and security
2. Ease of their deployment and use
3. Strong economic interests

Ethical dumping

STRATEGY

1. **Research ethics:** Control of public funding for research
2. **Consumer ethics:** Establishment of a certification system for products and services

Ethics evasion

“The malpractice of performing less and less ‘ethical work’ in a given context the lower the perceived return of such ethical work in that context.”

Ethics evasion

Applying double standards

STRATEGY:

Address the issue of lack of accountability

More fairness, less bias, and an ethics of distributed responsibility

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani