Agreement No. ***2023-1-PT01-KA220-VET-000153919***
KA220-VET – Cooperation Partnerships in Vocational Education and Training
Project Duration: ***01/09/2023 – 28/02/2026***

VERFISUM

Study on the Major Ethical and Trust Concerns and Potential Stakeholders Needs regarding Artificial Intelligence We Should Be Aware of and How They Should Be Addressed

WP2: ESCAPE ROOM STORYTELLING AND AI ETHICS AND TRUSTWORTHY

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Summary

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

DOCUMENT IDENTIFICATION

OUTPUT TITLE:	Study on Major Ethical and Trust Concerns
OUTPUT TYPE:	Desk study
LEADING PARTNER:	PROMEDIA 2000
AUTHOR(S):	Stefano Penge stefano@stefanopenge.it
STATUS:	Draft
VERSION:	1.0.85
DATE:	23/08/2024

1. Before reading

1.1. What is this document for

This document is a desk study attempting to answer to one fundamental question: what does it mean to develop and use ethic and trustworthy AI services?

The main target of the document are teachers, trainers, students which are directly or indirectly confronted in their day by day life with AI services, visible or invisible: self driving cars, selecting algorithms, intelligent help desks; and, in the last two years, chat bots and software capable of creating and elaborating texts, images, videos, speeches and music. While using these services - mostly free at least for basic versions - is easy, amusing, interesting and even useful as a help for producing new documents, this easiness make less visible - or even invisible - some non-technical problems that could arise. These services may diffuse false news, discriminate persons, take the job of someone or simply consume too much natural resources.[1] The most general category in which these problems fall is Ethics.

The term "ethics" at a first glance may appear as something clear, self-evident, without need of further analysis: an action executed by someone on someone else is ethic if it is good for the recipient. Ethics is the sum of all ethic actions. Whatever falls outside of this set is non-ethic and should be condemned.

Unfortunately, as the history of philosophy teach us, there have been (and there still are) many different ethics, depending on time, space and point of view, each of them pretending to be the only one.

When it comes to “applied ethics”, i.e. ethics applied to technologies, there is another difficulty: devices change the relationship between agent and recipient and distribute the responsibility of the actions along a chain. There is not a single subject that acts on another person, but a series of subjects which receive and propagate the action with its beneficence or maleficence. Weapons are typical examples: is the responsibility of the act of dropping an atomic bomb on Hiroshima of Tibbets, the Enola Gay pilot, of the general Handy, of the President Truman or of the entire Manhattan Project?

- Beyond weapons, media are the general class of devices that multiply the effects of our actions on other people in terms of number, speed, distance. Mediated actions have much more strong effects – even if they are in the domain of knowledge and feelings, rather than in the

[1] With the words of UNESCO Recommendation: “AI systems raise new types of ethical issues that include, but are not limited to, their impact on decision-making, employment and labour, social interaction, health care, education, media, access to information, digital divide, personal data and consumer protection, environment, democracy, rule of law, security and policing, dual use and human rights and fundamental freedoms,including freedom of expression, privacy and non-discrimination”. UNESCO Recommendation on the Ethics of Artificial Intelligence, November 2021, I, 2. (c). https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

physical one – than simple speeches. Since media are inherently chains, the responsibility of this effects has to be divided into all agents involved, even it is not always clear how.

- Digital media, like the Internet, may be seen as giant, hidden, chains. From a click on a button on a web page to the multiple final results there are nearly infinite steps, every one implying some degree of responsibility of the human persons who programmed that step.
- Finally, a software based on Artificial Intelligence is a very special case of digital media which is not only a device but also an agent, a *subject* of the relationship. In this case, the distribution of responsibility is even more difficult.

There can be different definitions of “acting on someone” when the subject is not human, and multiple approaches to judging this actions; consequently, there can be multiple set of principles that define which behaviours are acceptable for an AI (we will see some of them in Chapter 4). While there already are attempts to build a single framework to collect and harmonize all these set of principles to gain a single point of view, things are changing at a great speed and every new AI application push researchers to rethink their approaches.

We don't think yet that the moment is come to choose a single theory and adopt it disregarding the rest. We are still in a phase of field recognition and we should leave place for new ideas and visions. But nonetheless, it is important to design and diffuse “thinking tools”, conceptual keys that can be useful to analyse new situations and to reveal hidden aspects .

We, as Verfisum partnership, working in an Erasmus+ project aimed at young people, are focused more in promoting awareness about AI challenges than in teaching some particular ethical theory about AI. We think that youngsters should primarily know that there already are, and there always will be, ethical problems in applying a powerful technology like AI in everyday life, and particularly in education and in job placement.

There is the maximal attention, at European level, to these problems: to avoid that they could be ignored or, on the contrary, used as an excuse to block the "placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union". A further effort is required to study them "to promote the uptake of human centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights".[2] This document has the ambition to give a little contribution in this direction.

We hope that this document (along with the Survey) will be used also outside Verfisum project and partnership to help teachers and students to plan learning activities about ethics and AI and, more generally, to raise interest about these questions even in non-technical contexts.

[2] Artificial Intelligence Act, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, pag. 4

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

1.2. Structure

After this short introduction (Chapter 1) and a fast of recall of the role of technology in ethics (Chapter 2), we will examine four concepts: Trustworthy, Completeness, Collectivity and Identity (Chapter 3). We will see that they are highly interrelated and could help us to better define the field of our research.

Thus, we will focus on two general categories of concerns:

- Ethical concerns: ethics and machines, ethical principles, cultural issues and the EU position (Chapter 4)
- Trust concerns: artificial artifacts, weakness of users and possible causes (Chapter 5)

We have investigated the potential stakeholder needs by means of a survey (Chapter 6) that has been especially designed and managed; we will discuss the questions, their meaning and the answers.

We will then offer some suggestions (Chapter 7) on how these needs could be addressed using the tools described.

A bibliography (Chapter 8) closes the document.

1.3. Methodology and sources

This document is not the output of a long research, which would have been out of scope inside Verfisum project. The general idea was to draw a general picture of the domain of ethics and intelligent machines, but also to pinpoint some crucial concepts strictly related.

The first sources of the document were recent researches about Ethics and AI, in particular with respect to education. In the last five years there were a lot of studies which apply a critical approach, neither refusing all possible benefits from AI services nor hiding all possible risks, especially for the youngsters. We collect them and divided them into three categories: AI and Ethics, AI Ethics in education, AI in Education.

A particular focus was given to public debate on Principles for an Ethic AI, which produced several set of principles, from Asilomar conference (2017) to UNESCO Recommendation (2021) and finally to European AI Act (2024).

Another source was the on-line survey, conducted in June 2024 on 90 participants from Portugal, Greece, Estonia, Ukraine and Italy. The answers to the 30 questions are summarized in Chapter 6 and figures are reported in Annexe I; the key findings were also used to (re)design the overall structure of the document.

1.4. License

This document is released under Creative Commons 4.0 Attribution/ Not commercial /Share alike (BY/NC/SA) license http://creativecommons.org/licenses/by-nc-sa/4.0/

This means that everyone can freely distribute – and translate - this document, without fees; though, it is not allowed to modify it and to drop this CC license. It is also required to leave the attribution to the authors and to the Verfisum Project.

Raw results of the survey are released as open data under ODbl 1.0 License https://opendatacommons.org/licenses/odbl/1-0/ .

2. Ethics and digital technologies

In this chapter we will briefly touch the history of AI ethics, with the shift from an ethics *for experts* to an ethics *for all*. Then, we will describe some of the fundamental concepts and problems related to the application of ethics to artificial intelligence, namely the contradiction between freedom and algorithms, and the different scenarios that could be drawn trying to answer to the general ethic question when the subject is an artificial agent.

2.1. A bit of history

Ethics can be thought of as a set of general principles guiding the process of writing laws and regulations, or as a set of recommendations to guide the conduct in situation in which there is a lack of Right. Ethics is normally implicit: one of the meaning of the word is “the standard behaviours, the expected ones”. In a given culture, an act is defined “ethic” when/because it is performed by the large majority of persons of that culture. In normal, standard, situations, we don’t need ethics at all.

Ethics is brought to attention only when facing new and unforeseen situations. The impact between different cultures, as we see nearly every day, sadly shows that what seems good to us could be different from what is good for someone else from another country, language, religion. While there have been attempts[3] to create a global ethic, over all the differences, there is still much to do to ensure a full respect of other people's values.

Technology has a crucial role in this context. Ethics was used to be the “science of the evaluation of acts of persons towards other persons”. If a tool, for example a weapon, was involved in a situation requiring ethical judgement, it was seen as irrelevant, as a simple multiplier of the effects of the action. But in the last centuries we started to be confronted with devices which are more the simple tools. Obviously, they didn’t have a mind and consciousness, but we started to divide the responsibility among the human that *used* them and the human that *designed* them, as in the case of mass destruction weapons.

From this point of view, is not surprising at all that, when applying new technologies to the old world, we discover new ethical problems, particularly when the impact of these technologies cannot be foreseen easily. When cars slowly took the place of horse-drawn carriages, we started to pave roads; but we were not prepared to face all the problems related with speed, safety, pollution, petrol extraction, and so on. In general terms: we invent a technology, that will be efficient only when we will have rearranged the world in a certain way (a train need rail roads to run); but we are not particularly good at foreseeing the global effect of these changes on the Earth and on our minds.

[3] https://parliamentofreligions.org/globalethic/

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

The change in our perception of *time* has become evident comparing mail, e-mail and social messages. Progressively reducing the interval between the act of sending a message and receiving an answer caused an unpredictable psychological difficulty that we all experiment every day: we are no more able to wait. We check for message in a compulsory way. We bring smartphones in every place and in every moment of our life. Empty time is no more a resource (like the "otium" of old Latin writers), but a risk to avoid.

The change on the perception of time can be seen under another respect. Written documents are no more a synonym of permanent decisions, as in the old Latin motto "*Verba volant, scripta manent*" (spoken words get lost, written words persist). We write just to communicate immediately with friends, not to keep a trace of ideas. The easiness of conversion of speech to digital texts (and vice-versa) makes this communication even more immediate. In a similar way, while analog photos were printed and stored in album to be viewed later, today they are just taken to capture the moment, to be sent through social media system: nobody spent her time to review, order, filter them. They live only in the present and in collective space.

These changes are often invisible. Internet services have been continuously redesigned to be more transparent: a click is most of time all that it requires from us. This has had the effect of hiding all the process behind: software, protocols, computer, cables, filters, and so on. We need this infrastructure running, if we want to keep our way of life; but we are not aware of it, and we are not able to judge of its secondary effects on persons, resources, climate,

______

While the relationship between ethics and AI came up as a problematic issue only recently with intelligent weapons, self-driving cars or chat bot reinventing history, the general theme is very older. There was an *information technology ethics* long time before killing drones. Before AI, there were already some form of semi-autonomous devices, which can be thought of as agents: even if a software wasn't programmed to reason about situations, it could be perceived by human users as if it were intelligent, with its own will. This is evident every time the device uses our natural language to interact with us. The simplest program capable to write on a green terminal "Hello, world!" is seen as some form of intelligent agent even if we can read the code and can tell that there is no one behind.

It is interesting to see how the focus of ethics principles and recommendations changed during last thirty years. The Internet Architecture Board (IAB), a committee of the Internet Engineering Task Force (IETF) and an advisory body of the Internet Society (ISOC) was created by Vinton Cerf, one of the pioneers of Internet, in1979. In a memo dated 1989 (RFC 1087: "Ethics and the Internet")[4] the IAB takes a formal stance on what constitutes proper use of the Internet. This is the list of "bad actions" that every programmer has to avoid:

[4] https://datatracker.ietf.org/doc/html/rfc1087

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

(a) seeks to gain unauthorized access to the resources of the Internet,

(b) disrupts the intended use of the Internet,

(c) wastes resources (people, capacity, computer) through such actions,

(d) destroys the integrity of computer-based information,

and/or:

(e) compromises the privacy of users.

This list is written from the point of view of the (few) scientists or technicians capable to work with the baby-internet of first 80's, and it sound probably a little naive. But we already find here some of the concepts that become ethics principles forty years later, like security, non-maleficency, environment protection, privacy (see below, chapter 4).

In a similar way, the *Ten Commandments of Computer Ethics* presented in Dr. Ramon C. Barquin's paper, "*In Pursuit of a 'Ten Commandments' for Computer Ethics*" in 1992[5] are dedicated to programmers. The idea behind is simple: computers (and software) are very powerful tools, and we, the programmers, should try to make a fair use of this power.

1. Thou Shalt Not Use A Computer To Harm Other People.

2. Thou Shalt Not Interfere With Other People's Computer Work.

3. Thou Shalt Not Snoop Around In Other People's Computer Files.

4. Thou Shalt Not Use A Computer To Steal.

5. Thou Shalt Not Use A Computer To Bear False Witness.

6. Thou Shalt Not Copy Or Use Proprietary Software For Which You have Not Paid.

7. Thou Shalt Not Use Other People's Computer Resources Without Authorization Or Proper Compensation.

8. Thou Shalt Not Appropriate Other People's Intellectual Output.

9. Thou Shalt Think About The Social Consequences Of The Program You Are Writing Or The System You Are Designing.

10. Thou Shalt Always Use A Computer In Ways That Insure Consideration And Respect For Your Fellow Humans.

[5] https://computerethics.institute/publications/ten-commandments-of-computer-ethics/

Thirty years after, the Internet was changed a lot, in terms of dimension, number and competences of users and risks. In their new memo dated 2020 (RFC 8890) the IAB identifies protecting end users as the first priority in their maintenance of the Internet. "*[...] when we've identified a conflict between the interests of end users and other stakeholders, we should err on the side of protecting end users*".[6] So an action is not ethic if it has a negative effect on users - even if it was effected with the better intentions by programmer. It is the *effect*, not the intention, that is crucial.

Today, ethics concerns are much more extended. Privacy, digital divide, environment protection, sustainability, peace are unavoidable concerns. Consequently, there is a shift is from a list of commandments targeted to single persons (programmers) towards some general principles that should be respected at a more general level by collective subjects: companies, enterprises, States, Federations of States, the whole World.

In a way, this is the same ratio behind the Asimov's Zeroth Law (see below, 4.1). Ethics of machines is not only about judging single actions done by machines, but also about the general effects of these actions on humanity. Some of these effects are well known since the XIX century:

- machines can substitute workers, instead of making them more efficient
- machines can kill humans (soldiers or not) at a very large scale
- the quest for resources (coal, petrol, rare minerals) needed to build and run machines is changing the face of the earth

Some secondary effects of producing, using (and dismissing) machines, which increase the division among people, were recognized only recently:

- machines produce waste (solid, liquid, gas) that makes some environments difficult to live in, for animals and humans
- machines consume a lot of resource (electricity, water), stealing it to people
- machines can divide the world in two: those who have the knowledge, the abilities and money to use them, and those who haven't

[6] https://datatracker.ietf.org/doc/html/rfc8890#name-identifying-negative-end-us

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

We understood only in the last few years that some special machines (or better, the software running on them) in the long run could have effects also on the *cognitive* plan. These kind of effects are difficult to see, because they are the hidden, sneering face of a smiling one:

- software machines can hide the infrastructure (teaching us to *live on the surface*)
- software machines can solve more and more problems (contributing to the loss of the related abilities)
- software machines can take decisions using only a subset of information (instead of taking into account all the context)

2.2. Freedom and algorithms

The last type of problems is related to the algorithmic nature of machines. While the word "algorithm" is at present used as a synonym for a complex software using profiles and data to take decision about humans,[7] technically speaking every machine could be seen an implementation of one (or more) algorithms. Algorithms are not the hidden souls of machines: they are a clever, well studied and well-defined way of *describing* how a machine behaves. An algorithm is a formalized description of a way to do something in a limited interval of time, without risk of errors or infinite loops.[8] Machines normally follow always the same algorithm; but starting from Jacquard looms (1801), there is the possibility to change algorithms while keeping the same hardware. This was the great idea that prepared the advent of computers made of software decoupled from hardware.

Still, we do not expect a machine to ignore its algorithm and to try something new, different, creative. Even a computer, which produces extraordinary outputs thanks to millions of different algorithms, is always executing strictly the operations defined in one algorithm or another. To this point, there is no room for choice and ethics.

But things were going to change. Classic algorithms are deterministic (meaning that at every step the action is well defined), whereas non-deterministic algorithms solve problems "guessing" at every step the better possible choice given some metrics and some preassigned weights. Algorithms applied on time-dependent problems and contexts (like social media profiling and suggesting) are obviously non-deterministic.

[7] "[...] algorithms are playing an increasingly widespread role in society, automating a wide range of tasks ranging from decisions that impact whether someone gets a job to how long someone should remain in prison." AI and education: guidance for policy-makers, UNESCO, Paris 2021, https://doi.org/10.54675/PCSP7350

[8] There are wrong algorithms or infinite algorithms, but normally we are not interested in them

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

These are precisely the kind of algorithms used by AI systems: at any given moment they can generate a word (or a figure, or a sound) on the basis of what was generated before and of the evaluation of the values of all possible alternatives. Every run of the algorithm can give different outputs, since the enormous "models" that have been built scraping the Web have thousands and thousands of dimensions (and of weights).

We are building, testing and using machines that *seem capable of taking decisions* out of the well established path that we designed for them that we call "the program". From the ethic point of view, we are entering in a new territory.

2.3. Actors and subjects

In his long research about ethics of artificial intelligence, Luciano Floridi recall the distinction between *agency* and *intelligence*. He says that the use of the term "intelligence" about software like ChatGPT or other applications of Large Language Model (and in general machine learning based services), is incorrect. Remembering the well-know definition of artificial intelligence products given by John McCarthy as something than "would be called intelligent, if done by a human", he wrote that "these LLMs can process texts with extraordinary success and often in a way that is indistinguishable from human output, while lacking any intelligence, understanding or cognitive ability". ChatGPT would probably pass the Turing's Test, exactly because it is designed with this aim: not to be intelligent, but to *seem* intelligent.

This decoupling of agency and intelligence has a clear impact on Ethics: AI systems are agents without the intelligence that is needed to be conscious of (potential) effects of their actions. So, at least at the very moment, the question about ethics is about the *use* (or design, or production) of these system, not about AI system as subjects themselves.

____

Let's try a more general approach to this problem. Ethics is not about good and evil in abstract, but about good and evil *for someone*. There is someone doing an action: we are interested in effects for someone else and we want to understand if they are good for her.

We could define the "standard ethic question" as an *evaluation* of the expression:

X acts (via Y) and has effect(s) on Z.

In the media domain, Y is a chain of y1...y2...y3; or, we could substitute to Y an entire expression of the same type. Also, the effects on Z could be multiple and not always evident.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

This evaluation could be conducted from other points of view: economy (costs), technology (feasibility), ecology (sustainability). Ethics choose the point of view of Justice, or better said, of "Right non regulated by formal laws". Some of the questions of the survey (see below, chapter 7) have this form: is it right to use AI to judge someone?

The evaluation should take into account context, intentions, necessity, acknowledge of the effects and so on; but first of all we should think about *what* (who) we substitute as value of variables X, Y and Z. Different substitutions give different scenarios. Let's see some of them:

1 We normally put AI (like every other technologies) at the place of Y, as a simple *tool*: a hammer, a weapon. In this case, we have the traditional *ethics of instruments*: someone acts - through AI - on someone else. In this scenario, AI is like nuclear power: it is neither good nor evil, but all depends on its usage.

2 If we think about AI as an *agent*, we can put AI at the place of X. In this scenario, if AI kill civilians, hit a pedestrian, take the place of workers, get rid of inexperienced users, it is its responsibility and it should be blocked, or limited. This is the scenario taken from science-fiction novels and movies. AI are not neutral tools but agents, that means that the can have been programmed (or were self-programmed) with aims, general objectives, models of actions. Chat bot based on LLM are *not* this kind of agents (but they are often taken as if they wer) and this misunderstanding causes some effects too.

3 X can be taken as the final user (in the traditional instrumental Ethics), but also as a *condition* for existence of Y: the designer, the owner of the project, the owner of hardware in which the software is executed, the investor, the vendor. The responsibility climbs up along the chain and is distributed among all the links.

4 Finally, which is exactly the subject with respect to whom we do the ethical evaluation? Some Z, all the Z of some category, or all the possible Z in time and space?

4.1 We are normally akin to pay much more attention to harmful actions conducted towards weak subjects that can't defend themselves: children, aged persons, ill persons. But twenty years ago we discovered that missing education, formal or non formal, could be a strong limiting factor, and called this effect "digital divide". While the quantity of digital service increases in our life, some persons can't get any more a social service, needed information, or a job. These persons are also more prone to the risk of being cheated by a chat bot pretending to be a social service practitioner.

4.2 Normally speaking, Z can include only humans, but we are trying to extend its domain to animals, plants. This extension lead us to trespass the boundary between ethics and

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

ecology: we should be worried not only about our destiny as a species, but also about the whole Earth destiny.

4.3 Z could even be extended to other artificial being, like robots. Is killing a robot ethically correct (even if not prohibited by laws?). This is not an issue yet, but it is a good example of the kind of problem that are not faced until it is too late.

3. Trust and ethics

3.1. Good or true?

This document is about the relation between Ethics and Trustworthy in AI. It seems evident that we can only trust in a "good" subject, human or artificial; but things appear to be more complicated if we have a closer look to these general concepts. Trusting someone (or some source of information, or some source of decision) implies that this source is in some way compelled to *tell the true* in every occasion, even if it is wrong. To someone of my generation, the character of HAL9000 comes immediately to mind: an intelligent agent, obliged by its program to tell the truth, committed an error. HAL has an "opinion" of what is really good, which is different from that of humans. HAL has to solve a classic ethical dilemma: follow the law (its program) and tell the true objective of the mission, or apply its general principle of auto-conservation and kill all the equipage ? Be trustworthy or be good?

In the past, there was a common faith in the existence of Something that was Good and True. Plato's theory – which has been the reference for most of European philosophy and culture - put at the top of the hierarchy of entities the idea of supreme God. This faith, sadly, often implied that a missing accord in recognizing what is Good and what is True could be resolved only by brute force.

Years after year, we started to understand that quantity (one or all) and focus (we or they) are very important parameters. Some choices could be suitable if applied to one, but not to all; and they could be acceptable for us, but not for them.

Along the centuries, we discovered that good and true are not necessarily the same, nor necessarily tied together. We started to accept the idea that something could be *true*, *but not good* – this separation is the origin of modern science, that in principle isn't interested with values. We started to accept the idea that something could be *good*, *but not true* - this was the birth of the study of dreams, hallucinations, but also of myths, as something that has a relevant value even if from scientific point of view doesn't correspond to physical reality.

Moreover, thanks to anthropology studies, we understood that "humanity", a generic subject that is the intended target of ethics, means in facts something more restricted. Humanity is "us". The discourse of ethics is a discourse with a subject that is implicitly white, male, middle class, with a standard education, with standard intelligence, living in a town, speaking English (or at least, an official language), and so on. We then started to recognize the existence of a different kind of human persons, which don't live in towns and are not able to read and write in English. We tried to use the concept of "inclusion" to mitigate this strong bias, but this was a short term, partial solution. We declare our intention to "include them within us"; so the terms are not changing that much: there is still an "us" and there is a "them".

But the relationship between good and true become central again if we recognize that, when speaking about digital applications and AI services, the axis around which they turn is *data*.

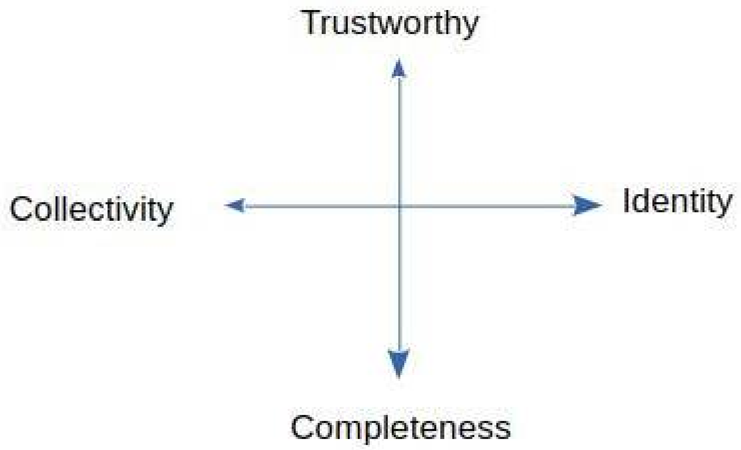
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Information technology was born with the aim of duplicating the world, building a mathematical representation of some parts of it that are more computable. We build a “digital twin” of a situation (a problem, an object) by converting analog data to digital data and filtering them; we use this reduced representation to do computations and finally we transfer back the result in the physical world. This is the way large ballistic computations for WWII or statistic elaborations for USA Census Bureau became feasible. But today we have digitalized maps, texts, images, voices and so on; and we treat with these reductions as if they were the original ones. Personal profiles are our simplified, digital twins which allow for computations: a software can categorize a profile, deduce some properties, take a decision about what to show after a click or which course or job is most suited for that profile. So the question, about which data can be collected and how they can be used, is crucial to build AI services that are trustworthy *and* ethic.

3.2. Dimensions

There are (at least) two dimensions along which we can dispose opinions, sentences, ideas. Let’s see them in the conventional form of a two-dimensional XY schema:

1. Trustworthy-Completeness
2. Collectivity-Identity



Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

These dimensions are an example of a kind of opposition in which both values are positive, but their sum is zero, meaning that when one increases, the other diminishes of the same quantity. Other well known examples are the Heisenberg's principle in physics (one cannot know position and energy of a particle in the same moment) or first Gödel theorem for axiomatic systems: they can be complete or safe, but not both. This kind of dimension is a conceptual tool useful to evaluate a choice.

The first dimension is a *representation of knowledge in terms of detail level*. We are aware of a fundamental limitation of our mind: one can know a lot, or even all, about something *or* can know something about all; but cannot know all about all. We have long time ago adopted this limitation as a base for building our education system in form of pyramid. Humanity spends time and money to prepare specialists which will always give correct answers about a small set of questions; the rest of us can give in average an acceptable answer about nearly any question.

Ideally, a piece of knowledge or a service could be placed in some point along the Y line to signify how much it is specific. Wikipedia would be placed near the Completeness pole (acceptable in average), while the Turing’s article about the computability of functions is close to the Trustworthy extremity (true in its small domain). We tend to give more credit to small, specific pieces of a puzzle – or to service that promise to offer them – than to big collections which contain a lot of general and perhaps imprecise sets of information.

The second dimension is a *representation of values in terms of focus*. It is about which part of our identity we would like to preserve versus how much we expect to be protected by a system that knows all. A service could boost the security or the privacy, but not both. This simply depends on the fact that security is maximized if all (dangerous) situations are known; on the contrary, privacy is the attempt to guarantee that a minimal part of our life is shared with others. This choice is not the same for all of us, but depends heavily on cultures.

Privacy is strictly connected with body and with sharing information about it. Dress codes in our post-Victorian societies are always a compromise between privacy (which part of our body we share with others) and security. By uncovering some portions of our body, we try to communicate with some category of persons – not with all – our will to be seen and desired.

The protection of privacy is also a main concern for urban architecture. In Europe, we open windows in houses to let light come in. But in some (southern) countries, windows can be shuttered if the owner would like to protect her privacy, while in other (northern) countries people simply don’t look through windows, so windows don’t have shutters.

Security – or safeness - is a collective value, one of the main reasons for living in collectivity-based societies. Where one animal is weak, a pack of many together is much more difficult to attack. We, as humans, went further in this direction building villages and putting palisades or walls all around them.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Medicine is just another type of wall around our body – or, around the bodies of all subjects of the collectivity. Medicine is necessarily a collective practice: only with the data collected from many patients we can try to develop a remedy.

Collectivity doesn't mean something well defined. We can think of a family, a clan, a village, a Country, all Europe, the human population on Earth, but also animals and plants. Here, we trespass the boundaries between ethics and ecology.

Maximizing the good for the collectivity, however, we have to renounce to some freedom. We want to be protected by our society; we are ready to give in change a part of our identity, but not all.

In other words, the two dimensions, **Trustworthy vs Completeness** and **Identity vs Collectivity**, are not independent. Security without completeness will fail; privacy requires trustworthy.

Security is based on the assumption that all doors are known and closed; even a single window forgotten or left open will break the security of the village (or of the body).

Privacy is based on trust; we don't want to share information about our life with some person or service if we don't trust that particular one.

Next in this chapter, we'll try to see how these concepts and dimensions are applied in the domain of digital services and AI services.

We will analyse the search engines, that are becoming every day something different and more similar to intelligent assistants.

3.2.1. Trustworthy

"If an answer is given, then it is true". This is the foundation (implicit) theorem behind any search engine. False answers cannot (or should not) be given.

The theorem stems from the metaphor of search engines as "librarians" managing the access to repository of documents. They use indexes to facilitate their work (they don't have to physically search a book, but just look at the catalogue); but the correspondence one to one between catalogue and library is mandatory. *Trustworthy*: if the document is in the index, it is also in the library. *Completeness* : if the document is in the library, it is also in the index. There is no room for false answers.

Nowadays search engines are something completely different: they aim to be full assistants that solve any problem: not only about finding written documents, but about facts, places, persons. They are not more like librarian, but like waiters, always ready to anticipate of the master before s/he has to ask a question.

In this new metaphor, it is not so important that the answer is true. First of all, it is not about the existence of a document, but about more general kind of information that could have been extracted or calculated. If the question is: "Please, suggest me a new web series I could to see with my husband tonight", there is not a true answer. And even: "Tell me which is the web series most seen in Portugal" has not a single, correct answer. Changing the usage model of search engine changes also our expectation of the unicity and trueness of the answer. We still trust the new AI powered search engine, but in a somehow different way.

It is interesting to read what said Sundar Pichai about Google trustworthy in a April 2024 interview with Roger Montti:

> *"Search used to be text and 10 blue links maybe 15 years ago but you know be it images, be it videos, be it finding answers for your questions, those are all changes you know …to to my earlier point people kind of shrug and …we've done all this in Google search for a long time and people like it, people engage with it, people trust it. So to me, I view it as a more natural continuation, obviously with LLMs and AI. I think you have a more powerful tool to do that and so which is what we are putting in search, you know with Search Generative Experience and so we'll continue evolving it in that direction too."*[9]

Pichai tries to rely strictly the two models (the librarian and the waiter), but it is hard to admit that all the question is reducible to same task accomplished with more power behind. And the trustworthy of the giant is somehow undermined after U.S. District Court for the District of Columbia on 5th of August 2024 found that Google has maintained an illegal monopoly of online search engines by paying other companies, like Apple and Samsung, billions of dollars a year to have Google automatically handle search queries on their smartphones and web browsers.[10]

3.2.2. Completeness

We are used to see hundred of pages of answers to a simple search, ordered by relevance. Relevance implies that the answer we are looking for is among the first ten; the more we flip the pages, the less we expect to find the answer. In other words, the normal behaviour for a search engine is to give answers *anyway*, even if they are only poorly related to the search. It is very rare (but still possible) that a search engine answers "No results found for xxx"; in this case, the search engine gives some suggestions to exit the impasse, like the following ones:[11]

- *Make sure all words are spelled correctly.*
- *Try different keywords.*

[9] https://www.searchenginejournal.com/google-ceo-on-future-of-search/513619/

[10] https://www.nytimes.com/interactive/2024/08/05/technology/google-antitrust-ruling.html

[11] The example is taken from Duck Duck Go, a search engine alternative to Google and focused on privacy

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

- *Try more general keywords.*
- *Try fewer keywords.*

Hidden comment: the answer is somewhere, but **you** have not been able to find it by submitting the correct question. The answer: "We do not know nothing about xxx" is never given. If no answer is given, then it doesn't exist. This is the claim for Completeness that every search engine is proud of.

But only a superior entity could know all and, consequently, could give correct answers about any questions. And AI chatbot services like ChatGPT are **not** - by large - this kind of entity. Remember that they are just software producing new texts (or images) on the basis of other texts (or images). They don't know *anything* about the physical world, they only have access to a small part of the digital and public representation of it. Moreover, they suffer of a particular syndrome: they can't stay mute, they have to give an answer, any answer, even if it is created on the basis of few information of low quality (that is, they are false).

3.2.3. Collectivity

The information era is also the digital security era. We stopped to build great castles with large walls when the gun powder demonstrated that any wall could be destroyed by a single cannon shot. Today, we build digital walls around any digital object that could be reached from outside. If we can do virtually anything acting on the "digital twin" of any object of the physical world, we should be able to ensure that these actions are performed by the authorized persons only and not beyond the contracted limits. The term "hacker" is often used with reference to a fictional character which is all the time trying to crack bank's digital walls. But digital security is not a battle against solitary, romantic criminals. It is mostly about protocols and laws, rights and exceptions.

The main theorem about digital security is: "To protect digital data, we have to get more data".

Data have to be collected to avoid threats and to prevent damages. The video cameras in front of banks record images of all people entering and exiting to be sure that a robber or a terrorist could be identified – before the action, if possible, or later. It is clear that a single image or number (which is the same thing) are not meaningful: we need series of data. When we have a series, we can search differences, values that are far from the expected ones. We build a default behaviour and we search for exceptions.

Data collected for security aims are primarily personal data, i.e. information about a person's acts, thoughts, habits, intentions, characteristics. The recent history of some smartphone apps aimed at signalling a contact with COVID19 infected persons has shown how security is primarily obtained by collecting personal data – and how it can be dangerous.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

The problem is that those data, useful or even crucial to catch a terrorist or someone infected with a epidemic virus, are collected *all the time* and even for persons that aren't neither terrorist nor infected. If there is no match among the data collected and the database, then the data should be erased. But there could always be false positives.

So security has a strong impact on privacy.

3.2.4. Identity

While the conceptual division between public and private can be found 2400 years ago in Aristotle (Polis vs Oikos), the history of term "privacy" started only in the last two centuries, with newspaper and journalists trying to violate the private space of VIPs. We have to remember that in those days, ordinary people didn't have a privacy right, in the same sense in which they didn't have personal belongings to be stored and protected in vaults. The first reference to privacy right ("the right to be let alone") was probably in a law review article published in the 1890 Harvard Law Review. Only in 1948 United Nations Declaration of Human Rights we can read that "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks."

Personal data should not be collected nor shared without the explicit consensus of the owner. This rule is based on the idea of recognizing and respecting citizens' personal limits. In our life there is a public part and a private part; only the first one can be accessed without explicit authorization.

But this standard situation has been revolutionized by Internet business model. The services offered by Alphabet' companies (Google) are typically free of charge: email, maps, translation and primarily search. The main revenues of Alphabet – covering the costs of these free services – are from targeted advertisement. This simply means that Google need to collect personal data from all of us, to elaborate them in profiles and to extract value from them. What is worth noticing is that it is not data which make value, but *big data*; it is only beyond a certain threshold that the quantity of data starts to convert in quality; that is, in money. To create a profile which is useful to foresee the behaviour of a class of users, Google needs a lot of traces: clicks, scrolls, searches, messages. Last generation AI services are heavily based on data collection. When a webcam records a face, it is useless without an enormous database with the faces of all the dangerous people. To build a model, we need also a lot of images of non-dangerous people, exactly to train the model and to teach him the difference. For example, a five years old smiling blonde child is not dangerous; hence, all faces similar to this one aren't dangerous neither.

If we think of privacy in a broader sense, as a protection of the identity, this concept cover also the intellectual properties. We have a common copyright law that defines the rights of an author on her written (or drawn) productions. If someone use a document under copyright without the proper authorization of the author, is it called "stealing".

But code is a written document.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Microsoft Copilot is able to write a good quality source code in several programming languages. It owe its capacity to the access to GitHub, the greatest freemium repository of source code. Programmers weren't asked to give permission to use their code for the training of the model: Microsoft (the owner of GitHub) simply put this permission as a default in the conditions of use of the repository. If someone decides to use GitHub to preserve his code, implicitly s/he is giving to Microsoft the permission to use it to train its models. This is legally correct, of course, because Copilot don't output exactly the copy of the source code collected, but some programmers said that it was a violation of their "privacy", that is, a theft.

A final remark about size. The necessity of big data implies big infrastructures: GPUs to make the computations, data lakes to keep the data safe and accessible, solar farm to generate the electricity needed to run and water pump to lower the temperature. This has an effect on the minimum dimension of AI companies: only great companies like Alphabet, Microsoft, Meta can afford the needed investments and develop the Large Language Model which are the core of AI chat bot. Nobody can forbid a startup to enter the field of AI services, but it should be big enough to have the technical resources to collect data and build its own models; so it needs big investors to sustain its research (like Microsoft did for OpenAI). In practice, the situation tends towards a monopolistic market.

Do this kind of remarks fall in the field of ethic of AI (business)? Later in chapter 4.3 we will see how the size matters when we come to the transparency and explainability principles.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

4. Ethical concerns

4.1. Ethics and intelligent machines

The very first idea of judging the behaviour of intelligent machines in terms of good and evil probably appeared inside science fiction literature. Most sci-fi readers will remember the Asimov's Laws of Robotics, presented in a 1942 short story:

1. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
2. *A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.*
3. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*

After 44 years, Asimov added a fourth law, which he called "Zeroth" because it should be the basis of the other three:

0. *A robot may not injure humanity or, through inaction, allow humanity to come to harm.*

Oddly enough, many if not all of the stories written by Asimov can be read as a demonstration of the practical *impossibility* for robots to obey to his Laws.

One could say that only Laws 0 and 1 are really related to ethics, while 2 and 3 are related to hierarchy, power, orders. In any case, an attempt to draw ethics guidelines on the top of these laws was done by a working group of professors and researches from 14 British universities in 2010. They recognize the historical importance of Asimov's laws, but clearly state that "they were not written to be used in real life and it would not be practical to do so". Most importantly, "[...] Asimov's laws are inappropriate because they try to insist that robots behave in certain ways, as if they were people, when in real life, it is the humans who design and use the robots who must be the actual subjects of any law."

Here follow the 5 "laws" that represent the outcome of the committee work:[12]

1. *Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.*
2. *Humans, not Robots, are responsible agents. Robots should be designed and operated as far as practicable to comply with existing laws, fundamental rights and freedoms, including privacy.*
3. *Robots are products. They should be designed using processes which assure their safety and security.*

[12] The principles are no more readable on line; this is the archived version of the page: https://webarchive.nationalarchives.gov.uk/ukgwa/20210701125353/https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

4. *Robots are manufactured artifacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.*
5. *The person with legal responsibility for a robot should be attributed.*

But these are not the only principles that were defined to control the (future) behaviour of robots and of intelligent agents in general.

4.2. Principles

We don't know if the Asimov laws were explicitly used as a background, but from 2017 to 2019 at least three set of Ethical Principles were published around the world (namely, California, Canada e Europe): the Asilomar Conference Principles, the Montreal Declaration and the European Guidelines. This coincidence is probably due to the big results achieved in AI development in those years. Consequently, there was a common perception of the need of limiting what would be possible to do in AI research, development and deploy - without blocking it.

Subsequently, at least four new sets were published at world-wide level:

- Council Recommendation (OCSE, 2019)
- Beijing AI Principles (Beijing Academy of Artificial Intelligence, 2019)
- Rome Call for an AI Ethic (Pontificia Accademia per la Vita, 2020)
- UNESCO Recommendation on the Ethics of Artificial Intelligence (2021)

It should be noted, though, that many of these principles were defined *before* that free AI based web services appear and become the most known application of AI all around the world. When asking to non-technical people to give an example of AI, the first answer is “ChatGPT”. When reading papers dedicated to the possible use of AI in education, the large majority refers to the use of machine learning systems to create texts, images, maps and so on. Teachers worried for the AI related risks always refer to the possibility to use Large Language Model to generate a text which can be used by students to cheat.

All these new applications, from the ethics point of view, bring with them new contexts, new problems, new concepts and perhaps some solution. The central issues are directly related to the Turing test, which was the answer given by Alan Turing to the question: what is Artificial Intelligence?

In short, his answer avoid to try to list the specific characteristics that a software should have to be intelligent, knowing that they depend on the definition of intelligence and that they could change in

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

future. He implicitly accepted the definition of intelligence as “what people call intelligent” and imagine a test: if a person is not able to distinguish a software agent from a human, than that software is intelligent.

Today there is the possibility for a AI service to present itself as a human person, by writing or speaking as a human, by creating images, videos, music and source code that is difficult to be recognized as machine-made. Hence, from an ethical point of view, there are some questions:

- should intelligent agents always explicitly present themselves as artificial ?
- should all their creations be marked, in an unmodifiable way, as artificial?

But other questions arise if we look at the way these services are build. All machine learning systems need a lot of examples to build the models by which they can simulate an human behaviour; these examples are normally taken from the web. So:

- should the copyright of machine-learning created artifacts belong (partially or totally) to the authors of the examples used? Should they be recognized, cited, and economically rewarded?
- should an author be explicitly asked for permission to use her/his creations to train an AI model?

Finally, as in classical industry process, these services can be used to mimics the professional performances of persons, which can be then excluded from the job market. But this time, it will be the intellectual jobs that can be replaced by machine; and, among them, the education ones.

These questions are not yet well understood and, hence, represented in form of principles in any of the set that we are aware of. There is the need for further analysis and integration of the existing set of principles.

______

The "AI Principles" became also a *marketing* tool, meaning that the enterprises involved in developing and selling AI services realized soon that a generalized fear about what AI could do would have a negative impact on the market. So Google (which has a long history of research in intelligent agents) has declares its own Responsibility principles[13]:

[13] https://ai.google/responsibility/principles/

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

1. Be socially beneficial.

[...] As we consider potential development and uses of AI technologies, we will take into account a broad range of social and economic factors, and will proceed where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides.

AI also enhances our ability to understand the meaning of content at scale. We will strive to make high-quality and accurate information readily available using AI, while continuing to respect cultural, social, and legal norms in the countries where we operate. And we will continue to thoughtfully evaluate when to make our technologies available on a non-commercial basis.

2. Avoid creating or reinforcing unfair bias.

AI algorithms and datasets can reflect, reinforce, or reduce unfair biases. We recognize that distinguishing fair from unfair biases is not always simple, and differs across cultures and societies. We will seek to avoid unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief.

3. Be built and tested for safety.

We will continue to develop and apply strong safety and security practices to avoid unintended results that create risks of harm. We will design our AI systems to be appropriately cautious, and seek to develop them in accordance with best practices in AI safety research. In appropriate cases, we will test AI technologies in constrained environments and monitor their operation after deployment.

4. Be accountable to people.

We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. Our AI technologies will be subject to appropriate human direction and control.

5. Incorporate privacy design principles.

We will incorporate our privacy principles in the development and use of our AI technologies. We will give opportunity for notice and consent, encourage architectures with privacy safeguards, and provide appropriate transparency and control over the use of data.

6. Uphold high standards of scientific excellence.

Technological innovation is rooted in the scientific method and a commitment to open inquiry, intellectual rigor, integrity, and collaboration. AI tools have the potential to unlock new realms of scientific research and knowledge in critical domains like biology, chemistry, medicine, and

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

environmental sciences. We aspire to high standards of scientific excellence as we work to progress AI development.

We will work with a range of stakeholders to promote thoughtful leadership in this area, drawing on scientifically rigorous and multidisciplinary approaches. And we will responsibly share AI knowledge by publishing educational materials, best practices, and research that enable more people to develop useful AI applications.

7. Be made available for uses that accord with these principles.

Many technologies have multiple uses. We will work to limit potentially harmful or abusive applications. As we develop and deploy AI technologies, we will evaluate likely uses in light of the following factors:

***Primary purpose and use**: the primary purpose and likely use of a technology and application, including how closely the solution is related to or adaptable to a harmful use*

***Nature and uniqueness**: whether we are making available technology that is unique or more generally available*

***Scale**: whether the use of this technology will have significant impact*

***Nature of Google's involvement**: whether we are providing general-purpose tools, integrating tools for customers, or developing custom solutions.*

All the same, Microsoft (which has strongly invested in OpenAI, the company which produces ChatGPT) published its six Responsible AI Principles[14]:

*1. **Fairness***

AI systems should treat all people fairly.

*2. **Reliability and safety***

AI systems should perform reliably and safely.

*3. **Privacy and security***

[14] https://www.microsoft.com/en-us/ai/responsible-ai#tools. Microsoft published also in 2022 the version 2 of its Responsible AI Standard, which extends and explains the six principles and can be downloaded here: https://go.microsoft.com/fwlink/?linkid=2257674&clcid=0x409

AI systems should be secure and respect privacy.

4. ***Inclusiveness***

AI systems should empower everyone and engage people.

5. ***Transparency***

AI systems should be understandable.

6. ***Accountability***

People should be accountable for AI systems.

Even it these "enterprise responsibility principles" may be similar or identical to the others, there is a clear difference. Alphabet and Microsoft are declaring their *intention* to behaviour ethically; the European Guidelines have been the basis for the EU AI Act (see below, chapter 4.4), a Regulation which the enterprises *must* follow.

We briefly discuss some of these set of principles plus a recent attempt to build a synthetic framework.

4.2.1. Asilomar Principles

The Asilomar Conference on Beneficial AI was held in 2017 in Pacific Grove, California. The conference was organized by the Future of Life Institute, a non-profit organization founded in 2014 by MIT cosmologist Max Tegmark, Skype co-founder Jaan Tallinn and others. The 23 principles were developed by a group of AI researchers, technology experts and legal scholars from different universities and organizations.[15]

They are divided in three main categories: research issues, ethics and values and longer-term issues. We focus on the second one:

[…]

6) Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

7) Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.

[15] https://futureoflife.org/open-letter/ai-principles/

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

8) Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

9) Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

10) Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

11) Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

12) Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.

13) Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

14) Shared Benefit: AI technologies should benefit and empower as many people as possible.

15) Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

16) Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

17) Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

18) AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

At the time of writing this document (07/2024), 5720 signatures has been collected

4.2.2. Montreal Declaration

At the end of 2017, the University of Montréal launched the co-construction process for the Montréal Declaration for a Responsible Development of Artificial Intelligence (Montréal Declaration)[16] "to stimulate discussion on social issues that arise with artificial intelligence (AI). [...]15 deliberation workshops were held over three months, involving over 500 citizens, experts and stakeholders from all backgrounds. The Montréal Declaration is a collective work that aims to put AI development at the service of the well-being of all people, and to guide social change by developing recommendations with strong democratic legitimacy."

The main objectives of the declaration were:

[16] https://montrealdeclaration-responsibleai.com/the-declaration/

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

1. Develop an ethical framework for the development and deployment of AI;

2. Guide the digital transition so everyone benefits from this technological revolution;

3. Open a national and international forum for discussion to collectively achieve equitable, inclusive and ecologically sustainable AI development.

The 10 principles:

1. ***Well-being***: *The development and use of artificial intelligence systems (AIS) must permit the growth of the well-being of all sentient beings.*
2. ***Respect for autonomy***: *AIS must be developed and used with respect for the autonomy of individuals and with the goal of increasing individuals' control over their lives and their environment.*
3. ***Protection of privacy and intimacy***: *Privacy and intimacy must be protected from AIS intrusion and data acquisition and archiving systems (DAAS).*
4. ***Solidarity***: *The development of AIS must be compatible with maintaining the bonds of solidarity among people and generations.*
5. ***Democratic participation***: *AIS must meet intelligibility, justifiability, and accessibility criteria, and must be subjected to democratic scrutiny, debate, and control.*
6. ***Equity***: *The development and use of AIS must contribute to the creation of a just and equitable society.*
7. ***Diversity Inclusion***: *The development and use of AIS must be compatible with maintaining social and cultural diversity and must not restrict the scope of lifestyle choices or personal experiences.*
8. ***Prudence***: *Every person involved in AI development must exercise caution by anticipating, as far as possible, the adverse consequences of AIS use and by taking the appropriate measures to avoid them.*
9. ***Responsibility***: *The development and use of AIS must not contribute to lessening the responsibility of human beings when decisions must be made.*
10. ***Sustainable development***: *The development and use of AIS must be carried out so as to ensure a strong environmental sustainability of the planet.*

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

The declaration was translated in ten languages (French, English, Spanish, Italian, Russian, Arabic, German, Chinese, Portuguese and Japanese). At the time of writing this document (07/2024), 2,830 citizens and 277 organizations signed the declaration.

4.2.3. European Guidelines for Trustworthy AI

The "Ethics Guidelines for Trustworthy AI"[17] were made public in April 2019. They were the outcome of the common work of an High-Level Expert Group on AI. The AI HLEG has worked closely with the European community of AI stakeholders through the European AI Alliance, an online forum with over 4000 members representing academia, business and industry, civil society, EU citizens and policymakers.In the first European AI Alliance Assembly, 500 members of the forum met in a live event that engaged the community into a direct feedback provision to the European Commission's policymaking process on AI. Although the AI HLEG ended its mandate in July 2020, the community of the AI Alliance continued its activity. In October 2020 over 1900 participants joined online the second European AI Alliance Assembly to discuss the main findings of the Public Consultation on the Commission's White Paper on Artificial Intelligence and future perspectives in building a European approach of excellence and trust in AI.[18]

The chapter I "identifies the ethical principles and their correlated values that must be respected in the development, deployment and use of AI systems".

In detail, the keys defined in this chapter are three:

1. Develop, deploy and use AI systems in a way that adheres to the ethical principles of:

   a) **respect for human autonomy**,

   b) **prevention of harm**,

   c) **fairness** and

   d) **explicability**

   Acknowledge and address the potential tensions between these principles.

2. Pay particular attention to situations involving more **vulnerable** groups such as children, persons with disabilities and others that have historically been disadvantaged or are at risk of

[17] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[18] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

exclusion, and to situations which are characterised by **asymmetries** of power or information, such as between employers and workers, or between businesses and consumers.

3. Acknowledge that, while bringing substantial benefits to individuals and society, AI systems also pose certain risks and may have a negative impact, including impacts which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself.) Adopt adequate measures to **mitigate these risks** when appropriate, and proportionately to the magnitude of the risk.

On the 17 of July 2020, the High-Level Expert Group on Artificial Intelligence presented their final Assessment List for Trustworthy Artificial Intelligence. Through the Assessment List for Trustworthy AI (ALTAI), AI principles are translated into an accessible and dynamic checklist that guides developers and deployers of AI in implementing such principles in practice.[19]

The key requirements are seven:

1. Human Agency and Oversight;
2. Technical Robustness and Safety;
3. Privacy and Data Governance;
4. Transparency;
5. Diversity, Non-discrimination and Fairness;
6. Environmental and Societal well-being; and
7. Accountability.

For example, to check the application of *Explainability* and *Communication* principles, which are part of *Transparency* requirement, the ALTAI suggests these questions:

- Did you explain the decision(s) of the AI system to the users?
- Do you continuously survey the users if they understand the decision(s) of the AI system?
- In cases of interactive AI systems (e.g., chatbots, robo-lawyers), do you communicate to users that they are interacting with an AI system instead of a human?

For *Diversity*, *Non-discrimination and Fairness* requirement:

[19] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

- Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
- Did you assess whether there could be groups who might be disproportionately affected by the outcomes of the AI system?

4.2.4. UNESCO Guidelines for policy makers

On November, the 23th 2021, UNESCO adopted a set of ten principles in its "Recommendation on the Ethics of Artificial Intelligence"[20]

The principles were drawn on the basis of four core values that reflect UNESCO approach:

1. Human rights and human dignity.
2. Living in peaceful, just, and interconnected societies
3. Ensuring diversity and inclusiveness
4. Environment and ecosystem flourishing

In the Preamble, some of the ethical risks of AI development and application are listed:

> *[…] AI technologies can be of great service to humanity and all countries can benefit from them, but also raise fundamental ethical concerns, for instance regarding the biases they can embed and exacerbate, potentially resulting in discrimination, inequality, digital divides, exclusion and a threat to cultural, social and biological diversity and social or economic divides; the need for transparency and understandability of the workings of algorithms and the data with which they have been trained; and their potential impact on, including but not limited to, human dignity, human rights and fundamental freedoms, gender equality, democracy, social, economic, political and cultural processes, scientific and engineering practices, animal welfare, and the environment and ecosystems,*

1. Proportionality and Do No Harm

The use of AI systems must not go beyond what is necessary to achieve a legitimate aim. Risk assessment should be used to prevent harms which may result from such uses.

[20] https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

2. Safety and Security

Unwanted harms (safety risks) as well as vulnerabilities to attack (security risks) should be avoided and addressed by AI actors.

3. Right to Privacy and Data Protection

Privacy must be protected and promoted throughout the AI lifecycle. Adequate data protection frameworks should also be established.

4. Multi-stakeholder and Adaptive Governance & Collaboration

International law & national sovereignty must be respected in the use of data. Additionally, participation of diverse stakeholders is necessary for inclusive approaches to AI governance.

5. Responsibility and Accountability

AI systems should be auditable and traceable. There should be oversight, impact assessment, audit and due diligence mechanisms in place to avoid conflicts with human rights norms and threats to environmental wellbeing.

6. Transparency and Explainability

The ethical deployment of AI systems depends on their transparency & explainability (T&E). The level of T&E should be appropriate to the context, as there may be tensions between T&E and other principles such as privacy, safety and security.

7. Human Oversight and Determination

Member States should ensure that AI systems do not displace ultimate human responsibility and accountability.

8. Sustainability

AI technologies should be assessed against their impacts on 'sustainability', understood as a set of constantly evolving goals including those set out in the UN's Sustainable Development Goals.

9. Awareness & Literacy

Public understanding of AI and data should be promoted through open & accessible education, civic engagement, digital skills & AI ethics training, media & information literacy.

10. Fairness and Non-Discrimation

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

AI actors should promote social justice, fairness, and non-discrimination while taking an inclusive approach to ensure AI's benefits are accessible to all.

4.3. A common framework

Every new set of principles try to keep into account the others. There were also attempts to survey, collect and categorize all principles in a single framework, with the aim of use them to practically to define parameters and check items not only to guide the development of AI, but also to stop a dangerous one.

Probably the most known of these research was the one conducted by professor Luciano Floridi of Yale University in chapter Four of his "Artificial Intelligence Ethics" [Floridi, 2023].

After having reported 47 different principles, Floridi proposes a simple schema with four standard class of principles (Beneficence, non Maleficence, Autonomy, Justice) plus a new one (Explainability).

The first four classes were taken from standard bioethics, which is the most recent form of ethics applied to technology:

1. **Beneficence**: AI services should help human persons
2. **Non-maleficence**: AI services should not harm or damage human persons.
3. **Autonomy**: human persons should always maintain the possibility to run and stop AI services
4. **Justice**: AI services should not discriminate and their advantages should be shared among all human persons

The fifth class is presented by Floridi as a specific one, that depends strictly on the machine learning and Large Language Model properties.

5. **Explainability:** there always should be a possibility to explain *why* an AI application took a decision

There is no room here for a complete history of AI; but it should be at least known that AI research started in '50 years, with the general idea that every human reasoning could be translated in formal rules. That general idea of AI substantially failed, or succeeded only in small, perfectly well defined domains like mathematics - that were, by the way, the first domain used to test this model. Last ten years saw the advent of another general idea of AI, which existed in the past as a pure hypothesis but was eventually made possible with a super power hardware and some brilliant mathematical ideas. In this "new" strategy, AI systems are capable of doing certain complex task without using any formal rules, but simply building models of a situation on the basis of countless similar situations. This

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

approach lead to software able to win every chess match, to drive a car, to maintain a natural language conversation and so on.

While the extreme power of this approach is apparent to all of us, there is a drawback: we loose the possibility to explain in detail *why* an AI system choose a certain move. As we said before, it is a central requirement of this approach to have a lot of data (big data) to build a model. The Large Language Models are very, very big files, that hardly can be contained in a standard PC and that require powerful parallel processors to run; but - most important thing - , they have literally *billion* of parameters. The size and the number of the parameters make unthinkable to "read" and analyse it. Hence, the importance of explainability as a requirement.[21]

Another possible meaning of explainability refers to the openness of source code. Beyond data, AI systems are made up of programs and libraries of functions that allow for exploring, connecting, selecting data. Models are not simply gigantic bunch of data, but structured data with an interface to interact with them. Apart of being usable without fee or not, their source code could be open (like Linux operating system) or closed and proprietary (like Microsoft Windows). The Free Software movement, from 1985 onwards, advocates the distribution of the source code together with every program, allowing each user to read, modify and redistribute the ameliorated code, not only to get better and efficient software, but to get a better and free society. But AI systems developers gather competitive advantages - with some exception[22] - by keeping closed their source codes.

Perhaps even more important than the content of the principles classes are what we can call *meta-principles*, that should be valid for any set of principles, to ensure their understandability outside of the limited circle of (English speaking) researchers, the real applicability at present and in the future:

1. principles should have a unique, multilingual definition; their definition should be based on a public ontology[23]
2. principles should have a protocol to check if they are applied or not; services which are *not* compliant which them should be publicly known.
3. a set of principles should have a scheduled maintenance

[21] There is a growing research on this problem; see Bao, A., Zeng, Y. *Understanding the dilemma of explainable artificial intelligence: a proposal for a ritual dialog framework*. Humanit Soc Sci Commun 11, 321 (2024). https://doi.org/10.1057/s41599-024-02759-2

[22] Llama-3 is a well know open source Large Language Model produced and distributed by Meta https://ai.meta.com/blog/meta-llama-3

[23] Ontology here is intended as a formal representation of a domain of knowledge. See, for example, Blagec, K., Barbosa-Silva, A., Ott, S. et al. A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks. Sci Data 9, 322 (2022). https://doi.org/10.1038/s41597-022-01435-x

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Having a common set of principles is not, however, the final solution to all ethics problems. As we wrote before, ethical values have different meaning in different cultures. We can easily imagine a situation in which the application of a principle (or a set of principles) would contrast with the application of another. Even if there is not an evident contradiction, we could simply have not resources enough to apply all of them and should make a choice. So we should order the principles by importance or give them different weights.

In the next chapter, we will have a look at the European AI Act, which is the most recent and complete set of principles about AI development and use.

4.4. The EU AI Act

Based on the European Guidelines, and on all researches conducted in the last seven years, after three years of discussions, comments and revisions, the European Artificial Intelligence Act was adopted on 13th March 2024.

This is the fundamental text[24] containing Regulations that should be known by every European citizen about the use of Artificial Intelligence services.

Due to its dimension, its nature of European level regulation (like the GDPR), its complexity (180 premises, 113 Articles divided in XIII Chapters, plus three annexes), its technical language, it could well stay unknown for the large majority of EU citizens that aren't working in legal domain.

This is the reason why here below we try to summarize it, quoting only a small subset of sentences that are relevant for ethics and for our aims.

1. Subject matter

The main objective of the AI Act is double.

On one side,

- *to "support innovation, improve the functioning of the internal market" but also "promote the uptake of human-centric and trustworthy artificial intelligence (AI)"* ,

On the other side,

[24] https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

- *ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter of Fundamental Rights, including democracy, the rule of law and environmental protection"* (Chapter I, Article 1)

These are not two different independent objectives, because one without the other will be useless or dangerous.

2. AI Literacy

The AI Act ask to every subject involved into the develop of AI services, but also simply those using them massively, to "*ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf, taking into account their technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons on whom the AI systems are to be used*. (Chapter I, Article 4)

3. Prohibited practices

Some practices are explicitly excluded by the AI Act. It is worth noticing that the prohibition is applied at different levels. This means that the chain of responsibility is extended from the final user to the producer, through all the intermediate link (distributor, reseller). These prohibitions applies to AI system that:

1. *deploys subliminal techniques beyond a person's consciousness [...] with the objective, or the effect of, materially **distorting the behaviour** of a person or a group of persons by appreciably **impairing their ability to make an informed decision***

2. *exploits any of the **vulnerabilities** of a person or a specific group of persons due to their **age, disability or a specific social or economic situation**,[...] with the social score leading to either or both of the following:*

   1. *evaluation or classification of natural persons or groups of persons detrimental or unfavourable treatment that is unjustified or disproportionate to their social behaviour or its gravity*

   2. *in social contexts that are unrelated to the contexts in which the data was originally generated or collected;*

3. *making risk assessments of natural persons in order to assess or **predict the likelihood of a natural person committing a criminal offence**, based solely on the profiling of a natural person on assessing their personality traits and characteristics*

4. *create or expand* ***facial recognition databases through the untargeted scraping*** *of facial images from the internet*

5. ***infer emotions of a natural person*** *in the areas of workplace and education institutions*

It is also prohibited:

6. *the use of biometric categorisation systems that categorise individually natural persons based on their biometric data to* ***deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation***

7. *the use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purposes of law enforcement*

(Chapter III, Article 6)

AI system shall not be considered to be high-risk if it does not pose a significant risk of harm to the **health, safety or fundamental rights of natural persons**, including by not materially influencing the outcome of decision making, if the AI system is intended to:

(a) perform a ***narrow*** *procedural task;*

(b) improve the result of a ***previously*** *completed human activity;*

(c) detect decision-making patterns or deviations from ***prior*** *decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review; or*

(d) perform a ***preparatory*** *task to an assessment*

(Chapter III, Article 6)

Two areas of use of high-risk AI systems are relevant for us: education and employment.

3. Education and vocational training:

AI systems intended to be used

(a) to ***determine access or admission or to assign*** *natural persons to educational and vocational training institutions at all levels;*

(b) to ***evaluate learning outcomes****, including when those outcomes are used to steer the learning process of natural persons in educational and vocational training institutions at all levels;*

(c) for the purpose of assessing the appropriate level of education that an individual will receive or will be able to access, in the context of or within educational and vocational training institutions;

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

(d) for ***monitoring and detecting prohibited behaviour of students*** *during tests in the context of or within educational and vocational training institutions.*

4. Employment, workers management and access to self-employment:

AI systems intended to be used:

(a) for the ***recruitment or selection of natural persons****, in particular to place targeted job advertisements, to analyse and filter job applications, and to evaluate candidates;*

(b) to make decisions affecting terms of work-related relationships, the ***promotion or termination of work-related contractual relationships****, to allocate tasks based on individual behaviour or personal traits or characteristics or to monitor and evaluate the performance and behaviour of persons in such relationships.*

(Annexe III)

In the next chapter we will examine the other dimension (trust vs completeness), which will give us a way to understand the difference and the relationship between search engines and Large Language Models.

Before starting to read the chapter, we suggest you an exercise: try to take a set of AI principles and substitute “AI” with simple “information technology”.

Which will be the difference, if any? Which are the principles specific for AI and not applicable to every digital application development and use?

If we are worried by the effects of AI on climate change, shouldn't we worry about the much greater effects of the entire Web?

Can we leverage the growing debate on AI and ethics to analyze *standard* information technology to check for ethical implications on "employment and labour, social interaction, health care, education, media, access to information, digital divide, personal data and consumer protection, environment, democracy, rule of law, security and policing, dual use and human rights and fundamental freedoms, including freedom of expression, privacy and non-discrimination"?[25]

The answer is, obviously, yes. This is precisely what has been done by *Vienna Manifesto on Digital Humanism*: a position statement written originally in 2019 by scholars from Technische Universität

[25] UNESCO Recommendation on the Ethics of Artificial Intelligence, November 2021, I, 2. (c). https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Wien and signed by over 1000 leaders worldwide that lays out the motivation and goals for the Digital Humanism Initiative. The Manifesto has been translated in seven national languages.[26]

[...] we proclaim the following core principles:

- ***Digital technologies should be designed to promote democracy and inclusion.*** *This will require special efforts to overcome current inequalities and to use the emancipatory potential of digital technologies to make our societies more inclusive.*
- ***Privacy and freedom of speech are essential values for democracy and should be at the center of our activities.***
- ***Effective regulations, rules and laws, based on a broad public discourse, must be established.*** *They should ensure prediction accuracy, fairness and equality, accountability, and transparency of software programs and algorithms.*
- ***Regulators need to intervene with tech monopolies.*** *It is necessary to restore market competitiveness as tech monopolies concentrate market power and stifle innovation. Governments should not leave all decisions to markets.*
- ***Decisions with consequences that have the potential to affect individual or collective human rights must continue to be made by humans.*** *Decision makers must be responsible and accountable for their decisions. Automated decision making systems should only support human decision making, not replace it.*
- ***Scientific approaches crossing different disciplines*** *are a prerequisite for tackling the challenges ahead. Technological disciplines such as computer science / informatics must collaborate with social sciences, humanities, and other sciences, breaking disciplinary silos.*
- ***Universities are the place where new knowledge is produced and critical thought is cultivated.*** *Hence, they have a special responsibility and have to be aware of that.*
- ***Academic and industrial researchers must engage openly with wider society and reflect upon their approaches.*** *This needs to be embedded in the practice of producing new knowledge and technologies, while at the same time defending the freedom of thought and science.*
- ***Practitioners everywhere ought to acknowledge their shared responsibility for the impact of information technologies.*** *They need to understand that no technology is neutral and be sensitized to see both potential benefits and possible downsides.*
- ***A vision is needed for new educational curricula, combining knowledge from the humanities, the social sciences, and engineering studies.*** *In the age of automated decision making and AI, creativity and attention to human aspects are crucial to the education of future engineers and technologists.*
- ***Education on computer science / informatics and its societal impact must start as early as possible.*** *Students should learn to combine information-technology skills with awareness of the ethical and societal issues at stake.*

[26] https://caiml.org/dighum/dighum-manifesto/#vienna-manifesto-on-digital-humanism

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

5. Trust concerns

5.1. Oracles and Singers: from search engines to creative agents (LLM)

It has been said that the very first examples of Artificial Intelligence are cited in Homer's Iliad (two golden handmaiden build by Hephaestus) and Odyssey (the intelligent ships that Alcinous, King of Phaiacians, give to Odysseus to bring him home). Robots and self-driving vehicles are in the dreams of humanity at least since then; and who can say what a dream can do, if taken seriously?

In any case, we can use two metaphors to better understand intelligent agents: the Oracle[27], like the Apollo's priestess Pythia, and the Aoidos, like Homer himself. These metaphors will help us to better understand the difference between search engines and intelligent creative agents in terms of trustworthy and completeness. A search engine should be trustworthy, even it is not complete; a chat bot aims to be complete, but cannot be totally trustworthy.

An Oracle is not a proactive agent: it simply waits for questions and give answers. The answers given are - obviously - *always* true, even if it may not seem so at the beginning, and this is the reason why even kings or heroes ask for the Oracle's advice before leaving for a war or an enterprise. A typical characteristic of its answer is obscurity, meaning that some kind of interpretation is needed to understand the full meaning of the answer, underneath or against the apparent one. But still more typical is the impossibility to retrace the genealogy of the answer, the logical chain that leaded from data provided in the question to that answer. An Oracle is not explainable. Still, oracles are very powerful, because of their source. They cannot lie, because they are simply a channel for the God (Apollo in the Pythia's case) to speak. It is a little bit strange to us the fact that, on the contrary, ancient gods spent their time lying and cheating mortals.

In a similar way, an Aoidos is a kind of prophet/poet that sing stories while being possessed by Muses, which are Goddesses. The story is rooted in the oral memory of the community, but the form in which it is sung is very personal. When asked for a story, the Aoidos retrieves from his memory a model and instantiate it in a version suitable for the conditions of his performance: the place, the time, the audience. So, an Aoidos *creates* every time a new instance of a static model.

We know that the research on Large Language Model and chat bot capable of dialogue were primarily aimed at empowering the search engine. In a recent interview, Sundar Pichau, Google and Alphabet CEO, declared that Google ha started to build the infrastructure for using AI to empower search engines back in 2016.

[27] The metaphor of oracle used to explain the power (and the limits) of AI based chatbot is widespread today. See G. Roncaglia, L'architetto e l'oracolo. Forme digitali del sapere da Wikipedia a ChatGPT. Laterza, 2023.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

The first search engines could be thought of as a kind of librarians which simply try to find documents that were marked with some labels. The request was of the form "[Give me a list of all documents marked with] ethics machines": a simple list of keywords. The search engine tried to find in the previously generated index one or more document marked with one or more of the keywords given by the user. There was a learning path to follow to become "proficient searcher". The users needed to be trained to use the engine; the first rule to be learned was that a document would be found if and only if *someone had marked it* with exactly those keywords. Only few knew that it was possible to choose between the search for "ethics AND machines" and the search for "ethics OR machines", choosing between intersection or union of the terms, or even "ethics NOT machines".

But training the user has always been a hard task; the preferred way has always been to make the tool more usable and transparent. Hence there were some attempts to develop a Natural Language Interface for search engines, to make the interaction with the "digital librarian" more easy. Another big issue was the national languages: searching for "etica e macchine" (or: "eetika ja masinad", "Етика і машини", "Ética e máquinas","Ηθική και μηχανές") will not return the same results as "ethics and machines" because since English is spoken by around 1,5 billion of person and Italian only by 60 millions, a lot more documents were written in English and marked with English terms. This issue was faced by creating multilingual ontologies and using them to "tag" the documents.

All these attempts had the objective to give us a better "user experience". But year after year, the search engines became services more and more necessary in our everyday life. We don't search (only) for documents, we search for the nearest pizzeria still open, the plot of a TV series, the result of the elections. Even if we don't do an explicit search, we get some suggestion. So, the research about how to offer a more powerful search service without a complex interface went on.

There were several strong ideas behind this research path:

- don't bother users with a special syntactic form ("find (x AND y) AND (x NOT w) "); let them build the question in a natural form, even it is redundant ("I'd like to know if there is ...")
- let users speak their mother tongue (“Per favore trovami...”)
- don't force users to imagine the exact result; let them express only their needs (“This night we want to eat outdoors”)
- don't force users to imagine the type of result: enlarge the searchable domain to the content of books, programs, Wikipedia articles, images, songs, videos and so on
- return the answers like a written report with only the most relevant answer and not in the form of a flatten list of results ordered by relevance
- don’t restrict the possible actions to "search" but permit also "resume, translate, visualize, tell".

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

At this point, the Oracle (the classic search engine) became an Aoidos (an AI creative agent *which is presented* as a search engine).

We are starting to see this kind of services in action with the intelligent assistant like Alexa and her friends; it is very likely that this kind of devices/services will increase their presence in our houses, cars, communication devices. They are very powerful services that can be of invaluable help for assisting impaired people: they don't need users to *write* prompts, they can answer on every question of every domain., they can be at our disposal night and day. At the same time, they are going to give help even *before* we can ask for it. It has become a standard behaviour, when starting a trip, to switch on the GPS antenna of the smartphone and to launch a navigator app: even when the destination and the routes are well known, we prefer to be supported and comforted. We have discovered that to lessen the cognitive charge needed by complex actions we don't need further experience, but just external trustworthy services. Next generation phones will probably show up a map every time their accelerometer notices an increase of speed.

Problems could raise when we temporary loose the possibility to use this kind of assistants (e.g. no more internet connection, or battery failing) and we are forced to go back to standard paper maps: meanwhile, they became unreadable for us. So the more we use intelligent assistants, the more we loose (unnecessary?) abilities. This is a well known effect of the process of developing new stages of culture by creating helping devices that make invisible some portions of our previous stage. This effect in sustainable if there is still someone, somewhere, which has a comprehension of these portions; but if all humans on the Earth should loose this memory, the risk to loose these abilities permanently is very high. One a day, we will no more able to draw maps, in the same way we can no more read Etruscan texts.

But with LLM, there is also a risk about the reliability of the assistant. Remember, they are not *oracles* , they are *aoidos*. They build models using our knowledge, and they give back to us a story based onto these models. As for Iliad and Odyssey, these new stories become parts of our culture. We will treat them in the same way we treat all well verified parts. If Wikipedia will be written by a LLM, we will no more be able to distinguish between human verified pages and "creative" pages.

The more their user is weak, the more they can be dangerous.

5.2. Weak users

From one point of view, it would be highly desirable to have AI agents capable of some degree of empathy. From small, anthropomorphic robots like NAO[28] to intelligent help desk, users and customers clearly prefer a warm welcome, some attempts to understand even the most confuse

[28] NAO is an humanoid robot created by French company Aldebaran Robotics in 2004. It was tested and deployed in a number of healthcare scenarios, including usage in care homes and in schools. See https://en.wikipedia.org/wiki/Nao_(robot)

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

requests, and a final "Sorry for my... uh... misunderstanding, and see you soon, I hope!" over grammatically perfect sentences strictly limited to cognitive content. Customers prefer a soft female voice or a warm male voice over a synthesized voice without age, gender and inflection.

But from another point of view, if an AI agent pretends to be human, it (or better: s/he) could cheat the user. S/he could ask for personal information, give biased suggestions or even convince the user to do something not correct or legal. Even if we know that humans are prone to errors while machines aren't, we tend to trust humans more than machines.

This is more risky if the users suffer of some form of "weakness".

5.2.1. Age

Age is not a weakness *per se*; or else, it shouldn't be one. But probably in every society, age is highly correlated with weakness, in both direction. Our (occidental) societies are designed to be lived in by adults, aged from 18 years old onwards. Contracts, licenses, votes, are only accessible for adults. This limitation is a form of protection of children from themselves. Children haven't the knowledge and the experience to understand complex problems and could make mistakes. Moreover, they trust adults and don't expect to be cheated by them. In a traditional society, the power of the children was limited accordingly with their lack of knowledge; in modern society, children have much more power, namely the purchasing one, direct or indirect. Influencing children means controlling the purchase of their parents.[29]

At the other end, aged people suffer from a overwhelming knowledge which is old and no more suitable to the real world. Words and concepts change with increasing speed and the ability to update their relationship diminish with the age. This is much more evident in the digital domain, in which the changes are fast and often parallel.[30]

We see today clearly the effect of what Al Gore called "Digital Divide" at the beginning of nineties. In some Countries, if not in all, even to get standard services (a bank account, an electricity contract) citizens are forced to have an email address and a smartphone where to install the official app of the provider. But aged people still have difficulties to get money from a POS; some of them don't understand words like *email client*, *PIN*, *cloud*, *One Time Password*, *token*, *browser*, *URL* and so on. Even if they encounter quite often these words, they soon forget their meaning and their correct usage. They use interchangeably "Google" and "chrome"; hence, they don't know that there are alternatives

[29] See L.A. Flurry, Alvin C. Burns, Children's influence in purchase decisions: a social power theory approach, Journal of Business Research,Volume 58, Issue 5, 2005,Pages 593-601, ISSN 0148-2963,https://doi.org/10.1016/j.jbusres.2003.08.007

[30] "Be mindful of the need to give appropriate policy attention to the needs of older people, especially older women, and to engage them in developing the values and skills needed for living with AI in order to break the barriers to digital life". Beijing Consensus on Artificial Intelligence and Education, UNESCO, 2019

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

to searching with Chrome using Google engine. They don't know exactly where a file is, whether on their personal device or somewhere in the “cloud”; hence, they don't know who can read and use it. They could use an helpdesk service to find an article to buy, but they cannot tell if it is human or machine-driven; hence, they cannot distinguish between information and advertisement.

Another meaning of the expression “digital divide” is even more basic: AI services, given their dimension and requirements in terms of resources, are run remotely and distributed through the Internet. This implies that to access an AI service one has to be connected to Internet all the time. [31] There are still places in Europe in which the Internet is not available as a common facility, nor it is free, but requires some more knowledge - and money.

Generally speaking, while digital services are mostly free, at least at simplest levels, money is still a barrier. Smartphones are the digital devices most used in the World; but the average price for a smartphone which is really usable (in terms of memory, display size and resolution, speed, connecting capacity) is still high.[32]

5.2.2. Cultures and languages

In Europe, we are used to meet people of different cultures just travelling around the corner. On a Mediterranean beach on summer, one can easily hear ten different national languages talked all around his beach umbrella. But cultures are not limited to Countries. Italy is a good example: we have 20 Regions, 100 provinces and 9000 towns. Nearly every town has its own history, its language, its habits… in short: its culture. On a pure right basis, all cultures are equal. But, as a matter of fact, certain cultures are more represented than others in media (literature, movies, songs), and, from 1980 on, in the Web. Since web today is the main environment where people search for an information, for a suggestion or for a solution to a problem, the culture set implicitly the boundaries of what we can access, learn and know.

This is true even for ChatGPT and other chat bot services, that apparently can be prompted in every language: the web is also the main source from where data used to build Large Language Models are extracted. This represents a big bias that makes certain languages (and related cultures) more represented than others.

Moreover, people speaking only their national language or having a limited knowledge of most spoken languages (Chinese Mandarin, Hindi; and, for European languages, English, Spanish, French and Portuguese) are relatively weak and could be an easy target for many kind of fraud.

[31] With some exception: for "techies": see https://www.nomic.ai/gpt4all

[32] "[…] there is also a need to provide low-cost models for developing AI technologies, ensure that the interests of low and middle income countries are represented in key debates and decisions, and create bridges between these nations and countries where the implementation of AI is more advanced." AI and education: guidance for policy-makers, UNESCO, Paris 2021, https://doi.org/10.54675/PCSP7350

5.2.3. Lack of Education

This is a very important point, and has been at the centre of education strategies of Europe Commission since 2006 .The European Framework for Digital Competencies, DigComp (now updated to version 2.2), has a Competence area dedicate to Safety, defined as "ability to ensure that personal and work devices are protected, including personal and work-related data and sensitive information in digital environments, or to understand how technology impacts mental and physical wellbeing and a general awareness of the environmental impact of digital"; but we can find relevant competences also in other areas, like "select a variety of digital technologies to interact", or "to judge the relevance of the source and content".[33]

These competencies are clearly not only technological ones. A digital competence is not about knowing to switch on a computer and to move a mouse; it is about how to use digital devices for self-development, instead of simply multiply speed and quantity of effects. The term "competence" means "having abilities and knowing when and how to apply them to a new situation"; so "digital competence" is not simply about how to use computer programs (or how to write a prompt for an AI service), but also about the *decision* to use it or not and about the *selection* of the better alternative. In a way, these competencies are not only at a personal level (use), but also at a collectivity level (develop). We – as a collectivity - should have these competences; we should know if, when and how to develop and use AI services.

Let take a fast tour on a domain quite different but related: the competence of educational Coding.[34] Coding is not only "move the cat sticking coloured bricks", but is a way to learn by doing. Roughly speaking, it is about digitally simulating a situation, finding and defining the fundamental rules that are behind it, making hypothesis about the evolution of the simulation and then "running" the model to verify the assumptions. The situation to be simulated could be taken from every domain: biology, physics, mathematics, but also linguistics, geography, history, music, and so on. Coding is not limited to information science, nor to STEM disciplines.

While the main reason given by USA President Barack Obama or by European Commission to foster the teaching of Coding in all the school level, starting even with K5, were related to future and to the employment domain ("we will need one million new programmers in the next ten years"), others researchers started to say that learning to code was more related with *culture* than with job. Even for children that would never get a job as a programmer, knowing the basic of programming could be a way to understand the world they live in a experimental, amusing and effective way.

As a secondary effect, playing with code could teach that behind every service there is a program that someone has written with some objectives, some limits, perhaps some errors. There is no magic behind

[33] https://publications.jrc.ec.europa.eu/repository/handle/JRC128415

[34] See for example https://code.org/about

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

digital services. The title of a fifteen years ago Douglas Rushkoff book "Program, or be programmed"[35] refers to the importance to learn some Coding to be autonomous citizens. It was a way to recall that digital technology could be a powerful support, but also a barrier or, even worse, a bunch of magic spells aimed at controlling people that are not smart enough to decipher them.

Going back to AI, the central issue is not to guarantee that every child knows how to write an efficient prompt (since it could be something completely different is two years...), but that every person knows how these services are built, run, which limits they still have and which competencies they are simulating.

We should give students the opportunity to *choose* their future, not simply to be well prepared to it.

[35] https://rushkoff.com/books/program-or-be-programmed/

6. Potential stakeholder needs

6.1. The survey

This chapter is a detailed introduction of the Verfisum Ethics Survey. It is composed by two parts:

1. an explanation of the ratio behind the survey
2. a presentation and discussion of the results

Since Verfisum project is about the design of digital learning tools that will be used by young and young adults, we asked ourselves what the final users of these tools think (and know) about artificial intelligence and ethics. We decided not to do a scientific research with pilots, control groups and so on. We just design a simple tool to have a general picture about sensations, beliefs, conceptions (not really about knowledge) of Verfisum main target.

The survey, along with these explanations, was published and diffused by all Verfisum, partners; it was also used to support discussions about AI end Ethics in small groups. Here follows a brief report on an informal group managed by Life Zone Group in Estonia:

[We] had a small 6 people discussion about the questions nr 9 of the survey.

The discussion was very interesting and quite lively. There were different opinions. All agreed that there should be some kind of regulation on ethics in AI, but agreed that there would anyway be people who would not follow ethics guidelines, the same as in other areas of human activities.

All participants agreed that some more information and clarification is needed to form a strong opinion about any of the questions.
The most arguable question was about AI as a judge. Initially, people were against being judged by AI , however, at the end, they agreed that it would very much depend on the country and its jurisdictional system and also on the type of offense or crime and so on.

Before starting to answer, we asked to people to read this short (and simplified) text about what is ethics, about how it has changed along the time and which could be the relationship with machines.

What is Ethics

When a discussion explodes between two persons or two clans or two cities, each of them thinks to be right. Since these discussions may lead to wars, the role of judge was invented to decide and to

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

stop the discussion. Laws were invented to tell the judges how to decide in different situations. But then a question arises about who write these laws and and how should write them. Moreover, laws only define the right behaviour in general terms; but in everyday life people are confronted with situations which weren't foreseen before.

The word "ethics" was invented by ancient Greek philosophers to define their attempts to determine what is good and what is evil is such a way that all persons could agree about. They strove to give governments a guideline to write laws (and also to give people a rule when laws are not applicable). Why philosophers? Because (they said that) they were impartial, they were only interested in city's good.

Ethics in evolution

The meaning and application of ethics didn't stay unchanged from its invention. Romans were very interested in ethics as a general way to define how a good citizen should behave, also in his private life. Then, for nearly fifteen centuries, (in Europe) ethics was absorbed by religion: in these times the distinction between good and evil came from high. Since the devil was all the time trying to tempt humans, there were the necessity to determine exactly why someone did something and to categorize all possible cases.

From XVII century onward, after the great travels across the oceans and far lands, philosophers started again to discuss about ethics as an earth-level problem. When laws, habits and cultures are so different, could we still refer to a single source of ethics? Ethics is not a close set of laws, they said, but it is an open reasoning about how to define principles in a common way. Ethics is universal, or it is not.

Ethics and machines

From the beginning of the ethics' history, there were attempts to extend the application field of ethics beyond human beings to other categories of subjects: gods, animals, slaves. But these extensions are also useful to better understand the fundamentals of ethics, its principles. Is ethics applicable to 'all mighty' beings, like gods? Or to beings which have no idea of good and evil, like animals? Or to beings that are not able to choose but are forced to execute orders, like slaves?

Digital machines are very curious subjects: they share all these properties. And AI based service are still more intriguing: they are no cleverer than home pets, but they are so able in communicating that they could cheat us and pretend they are humans: "I'm so sorry". They can do some of the things that we supposed reserved to us: walking, talking, writing, driving a car, shooting. They show the ability to decide. They can be harmful: they can kill someone or they can write fake news or fake videos.

But are they guilty for these acts? Can we judge a robot for its bad actions? Or should we judge the robot's master? The programmers which design it? The administrators of the company that owns the factory that produces robots?

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

So, what do you think about computer/AI ethics?

We were very careful about personal data. We collected only the answers. We didn't collect the IPs (the address of the PC of the user), we didn't use tracking cookies. The survey was completely anonymous. Moreover, we used an open source online software (https://yakforms.org/) which was managed by a french association (https://framasoft.org/en/) who declare not to use the collected data (https://yakforms.org/pages/legals) in any form.

At the end, we shared the collected data and the charts with everyone interested, starting from the answering people.

6.2. Part A: the participant profile

1. Age
2. Country / Mother languages
3. Education
4. Do you use AI dialogic services (e.g. ChatGPT, MidJourney, Dall-E, ...) for amusement?
5. Do you use AI dialogic services (e.g. ChatGPT, MidJourney, Dall-E, ...) for study/work?
6. Did you hear about AI before 2021?
7. Did you read something about?
8. Do you know that EU in 2019 published a document in 27 languages about AI and ethics?

This part was intended to try to understand to which degree the participants were aware of the state of the art and of problems posed by AI services. Data could be used to try to cluster the answers or to show dependence of answer from country, age or education. These kind of research is beyond our scope, but raw data are made available to all researchers interested to deepen.

- biographic: 1, 2, 3

- experience in the field: 4, 5

- theoretic knowledge: 5, 6, 7

The 8th question is a little bit tricky: we already knew the answer (young people aren't aware of EU publications...), but we wish to suggest the reading of this document. This could also be an activity to be done in group.

6.3. Part B: the questions

Questions are grouped in 10 sets: ethics, machines, responsibility, fields of application, money and rights, laws and regulation, tricking, trustworthy, privacy and security and health, jobs.

This division is already a way to suggest that application of ethics on AI is a complex field. We start from general, simple questions about freedom and necessity and go forward to questions that require some

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

more analysis. At the same time, the questions try to involve personally the respondent (especially those from 9th set).

The general idea is to foster a critical reflection about AI. There is not a correct answer.

- **Ethics**: the 3 questions are about the necessity, in a multicultural Europe, to take into account diversity,even when we trust in the universality of values.
- **Machines**: the 3 questions are about the freedom needed to behave ethically. Classical machines aren't free enough to make a choice, but what about intelligent ones? Is intelligence related with freedom more than to knowledge and logic?
- **Responsibility**: if machines cannot be taken for liable, who is the human responsible for their actions: the designer, the operator or the owner of the machine?
- **Fields of application**: the aim of these questions is to push the respondent to think about more fields in which there could be ethics problems, behind those that are commonly cited on media.
- **Money and rights**: questions about the authors' right in all those case in which AI generative service seems to produce new contents but on the basis of older ones. The relationship between rights and money is often not so much evident for young people.
- **Laws and regulations**: questions about the difference between laws and guide lines, or between hard limits and soft limits. This difference could be seen at Country level or at European level.
- **Tricking**: these 3 questions use the concept of "guilty" to foster a reflection on a fundamental problem of ethics: we should be able to judge an action independently from the actor.
- **Trustworthy**: the questions try to relay three big source of information on the web: Wikipedia, the search engines and, today, the AI services - which officially aren't source of trustful information.
- **Privacy, security and health**: this is probably the most intriguing set of questions. The three questions could be used as a starting point for a game. They try to push the respondent in a simulated, fictional (by now) situation in which s/he has to do a choice.
- **Jobs**: finally, these questions are more strictly related to the core of Verfisum project. The idea is to stimulate a critical reflection about the impact of artificial intelligence on the job market: there is not a simple future, neither in a positive sense nor in the opposite one. While the

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

questions could seem mutually exclusive, it was perfectly reasonable to answer "yes" to the three.

---

1. ***Ethics***

    1. *Do you think that people should respect some form of ethics beyond laws?*
    2. *Do you think that there ethical principles valid for all of us?*
    3. *Do you think that ethics principles stay unchanged from the beginning of our history?*

2. ***Machines***

    1. *Ethics is about choices. Machines are mechanical devices. Should ethics include machines?*
    2. *Or else, should ethics include at least intelligent agents?*
    3. *If ethics could be strictly defined as a set of rules "do/ don't do", should we program robots to behave "ethically" by design?*

3. ***Responsibility***

    1. *Should the AI driven car programmer be considered responsible for the car's acts?*
    2. *Should the car vendor be considered responsible for the car's acts?*
    3. *Should the car driver be considered responsible for the car's acts?*

4. ***Fields of application***

    1. *Do you think that besides AI-driven cars killing pedestrians there are some more ethics issues in using AI?*
    2. *If yes, check one or more below:*
        1. *intelligent weapons* □
        2. *environmental sustainability* □
        3. *jobs loss* □

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

        4. *privacy risks* □

        5. *fake news* □

        6. *other* □

    3. *Do you think that these aspects should be adapted in different countries/cultures?*

5. *Money, rights, ...*

    1. *Is ethics also concerned with the way AI services are created?*

    2. *Images From Text services (like MidJourney, DALL-E) collect and use a lot of images from around the web. Chatbot agents (like ChatGPT) collect and use a lot of texts from around the web. Do you think that they should cite original authors?*

    3. *Do you think they should pay original authors?*

6. *Laws and regulations*

    1. *Did your school (office, ...) publish some guide-line on using AI services?*

    2. *Do you think that each European country should have a law to limit what an AI service can do?*

    3. *Do you think the EU should have a common regulation about what AI services can do?*

7. *Tricking*

    1. *Do you feel guilty if it happens to you to use an AI service to translate a text you wrote?*

    2. *Do you feel guilty if it happens to you to use an AI service to write a text you should write on your own?*

    3. *Do you appreciate your teacher if you knew that s/he used AI service to assess your assignment?*

8. *Trustworthy*

    1. *Do you think a computer can be wrong?*

2. *Do you think Wikipedia pages are always trustworthy?*

3. *Do you think that search engine answers are always correct?*

4. *Do you think AI service's answers are always true?*

9. ***Privacy, security, health***

1. *A research group of a famous University claims that they found a final cure for Alzheimer disease. Then, they reveal that the therapy was conceived by an AI program. If you were the Health Ministry, would you allow the use of this therapy?*

2. *Messages exchanged via social network services may contain information relevant to prevent a crime. Since it is not possible for humans to read them all, the Police Department proposes the use of an artificial intelligence service to examining them. If you were the Government, would you allow it?*

3. *A police officer halts you and brings you in a prison cell for some reasons you can't even imagine. Would you accept to be judged by an AI Justice Service?*

10. Jobs

1. *Do you think that AI services will delete a lot of jobs?*

2. *Do you think that AI services will create a lot of new jobs?*

3. *Do you think that AI could improve or facilitate the way people work?*

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

6.4. The results

Some general findings:

- 35 questions
- 3045 total answers from 87 participants
- Two third of them were < 26 years old

The experience on the AI field was more deeper than expected:

- >60% of users have some previous experience with ChatGPT, for their job or just for amusement (but since a lot of young people answer so, it is reasonable to understand "job" also as "my school homework");
- 39% declare an experience for both (job and amusement).
- >87% of users have read something about AI
- As expected, only 18% knew about EU 2019 paper on AI and Ethics.

There were three possible answers: yes, no and not sure. We expected much more indecision, given the complexity of the questions,

- only 445 "not sure" answers were given (14,61 %)
- 267 "non sure" answers were given by people < 26 years old (60%, 8,77 % of total answers)

Some questions are more obscure than others - questions with more than 15% of "not sure" answers:

- 27% of participant didn't know if their school/office has official guidelines about using AI

- 26,7 % has not a clear idea whether responsibility for car accidents should be given to software programmer
- Similarly, 25,6 % are not sure about the possibility to program robots to be good in every situation
- 20% of participants has not a clear opinion about the necessity to cite and/or to pay original authors of the documents used to build the Large Linguistic Models.

The top of the uncertainty was reached in the last part of the survey, as expected (see above):

- Two of the "simulation" questions had a "not sure" in 28% of the answers.
- Finally, the 28% of respondents were not sure if AI services will create new jobs.

Some kind of common vision about ethics and AI can be deduced from the fact that nearly all (> 70%) participants answered in the same way to a relevant set of questions.

- 92,2% think that ethics is necessary (this is the answer with the highest percentage)
- 81,1% think that there are universal ethics principles
- 73% think that those principles are also "eternal"
- Ethics should include intelligent agents for 75,6% of participants (only 60,7 % think the same for standard machines)
- 78,4 % of participants think that we should have an European level regulation about AI
- 82% of participants use without feeling guilty the automated translation services
- An average of 83 % of participants knows that answers from search engines, Wikipedia or ChatGPT are not necessarily true or correct, with a small preference accorded to Wikipedia (75,3 %) and a clear diffidence for ChatGPT (89,9 %).
- 88,4% of participants thinks that AI services will improve or facilitate the way people work
- But 77,3 % also thinks that AI service will *delete* a lot of jobs (and only 34,1 % thinks that they will *create* a lot of jobs).

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

The raw data are downloadable in CSV format from https:/www.stefanopenge.it/public/verfisum/verfisum_data.csv.

The complete statistics with related pie charts are reported as Annexe I.

7. How needs should be addressed

In this last chapter we will try to draw some conclusions: what we can do to help younger people to cope with fears, to avoid cons, to understand risks and do the better choice?

7.1. Time

There is a general call for action expressed in sentences like these ones: “We are in hurry, we should no more stay still” or “Now we have to do X, to decide Y, to use Z…”.

This feeling of urgency is correct, since we could not leave the industry to develop and sell AI systems without any control; their developing rhythm is dictated by market and investor’s needs, and we, as civil society, should act rapidly to draw the boundaries.

On the other side, we should not ask people to act without analysis; we should not push youngsters to accept the current situation without first thinking and searching alternatives. We have seen in previous chapter that the respondents to our survey weren’t totally unaware of risks; but they were a little bit confused about some fundamental questions (future job market, copyright, responsibility of criminal acts). We should leave them the time for a deeper reflection.

So, as simple as it may seems, we should ensure that young people have *enough time* for reflection about AI ethics problems. Time is a precious resource that, as we saw before, we are all loosing.

We should create whenever possible occasions to debate, to exchange opinions among peers. The game roles based on fantasy stories about AI applications (like the ones proposed in the final section of the survey, see 9.1, 9.2, 9.3) could help to start thinking outside a rigid, scholastic framework.

But education is also a big issue to face (see before, 5.2.3).

7.2. Literacy

As we saw, weakness is strictly related with lack of *meaningful* knowledge. By “meaningful”, we intend that every European citizen should be put in condition to access relevant information at her level: not too technical, nor too simplistic. As we saw before, the EU AI Act is a fundamental starting point for reasoning about AI Ethics, but it should be adapted to the level of readers.

This kind of education cannot be left to a standard computer science course for beginners. We should design a multi-level curriculum and adapt it to different ages and education levels. This is a project that Universities could lead, giving guidelines and example of learning units, while Associations could collect and interpret the real needs of youngsters and build meaningful educative contexts. Teacher’s should be involved in both activities, to design and to validate learning materials and environments.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Not obvious as it may seems: subjects of these training actions should be trustworthy and ethic themselves to ensure the needed quality. This don't mean excluding private or for profit companies, which can build their legitimate business on critical education about AI ethics.

At the same time, it is important to coordinate this kind of actions at European level.

Learning is a life long task and cannot be limited to formal education. We should ask to industries, offices, companies, to ensure internal training about AI ethics, giving guarantees that this training is not biased by their commercial interests.

7.3. Approach

When reading an advertisement about The Next Big Thing, or a declaration of principles of a company developing or selling AI services, or - even more important - the conditions of use of these services, we should adopt a *critical* approach. In short, this means that

1) we shouldn't take for granted any affirmation and that we should check it;

2) we should have a look to (hidden) conditions that enable a service, and

3) we should try to avoid too abstract and general expressions, translating them to specific and applicable ones.

1. As we tried to show along all this document, the problem with ethics is that everyone agree about its principles ("try to be good, don't be evil!") but in practice we haven't the same idea about what is good.

So, we could try to apply a check list taken from one or more of the list of principles that we saw before:

- Is the economic prosperity created by AI shared broadly? Is it empowering as many people as possible (from Asilomar Principles)?
- Is AI compatible with maintaining social and cultural diversity? Is it restricting the scope of lifestyle choices or personal experiences? (from Montreal Declaration)
- Is AI exploiting any of the vulnerabilities of a person or a specific group of persons due to their age, disability or a specific social or economic situation? (from EU AI Act)
- Is AI deploying subliminal techniques with the effect of materially distorting the behaviour of a person or a group of persons by impairing their ability to make an informed decision? (from EU AI Act)

When we read a declaration of intent by a company proposing an AI service to adopt the beneficence principle, we could use the two dimension schema (3.2) to ask:

- trust vs completeness:
  - are there protocols to verify the quality of the answer?

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

  - are there tools to check the quantity of the answer
- identity vs collectivity:
  - who is going to be the beneficiary and how her identity is protected?
  - who is going to pay for collateral effects and how her safety/security is guaranteed?

With this answers, we could try to place the AI service (ideally) in some point of the two-dimensional schema we introduced before. Is it presenting itself as trustworthy *and* complete? It is declaring to preserve identity of users, while protecting the all community?

As we saw before reading the EU AI Act (4.4), this approach should be adopted particularly when using AI application in *Education and vocational training* and *Employment*, *workers management and access to self-employment*, which are field highly relevant for Verfisum project.

2. A critical approach implies an attempt to find the conditions that enable a certain phenomenon: what makes it possible? Which resources could stop it, if missing? Which boundaries should not be trespassed to keep the service effective? Sometimes, these conditions are not so visible, because they are far or hidden, or simply they pertain to another domain. The questions about sustainability of AI (energy, water consumption) are a good example of this kind of approach: when we ask something to a chat bot we don't see how much energy the answer costs.

3. Finally, a critical approach requires to transform general, abstract terms, to singular and concrete ones, along with conditions that may limit its application. We know that AI is a *portmanteau* word that has different uses: there have been a lot of different meaning of "intelligence", many different problems and completely different applications. This is also true for expressions that become a kind of *motto* that nobody will discuss. For example, when we read "AI will help inclusion", we should try to ask: inclusion for whom? Given through which AI services? Under which conditions and for how much time? And what is supposed to happen after that time?

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

8. Bibliography

8.1. AI Ethics

1. M. Estevez Almenzar, D. Fernandez Llorca, E. Gomez Gutierrez, F. Martinez Plumed, "Glossary of human-centric artificial intelligence". Luxembourg: Publications Office of the European Union, 2022
2. A. Casilli, 'End-to-end' ethical AI. Taking into account the social and natural environments of automation, in Aida Ponce del Castillo. Artificial intelligence, labour and society, ETUI, pp.83-92, 2024, 978-2-87452-707-4.
3. L. Floridi, AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models (February 14, 2023). Philosophy and Technology, 2023, Available at SSRN: https://ssrn.com/abstract=4358789 or http://dx.doi.org/10.2139/ssrn.4358789
4. L.Floridi, The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities , Oxford University Press, 2023, https://doi.org/10.1093/oso/9780198883098.001.0001
5. C. Novelli, F.Casolari, P. Hacker, G.Spedicato, L. Floridi, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity (January 14, 2024). Available at SSRN: https://ssrn.com/abstract=4694565 or http://dx.doi.org/10.2139/ssrn.4694565

8.2. AI Ethics in Education

6. S. Akgun, C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in K-12 settings", in AI Ethics, vol. 2, pp. 431–440, 2022
7. European Commission, Directorate-General for Education, Youth, Sport and Culture, Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2766/153756
8. R.S. Baker, A. Hawn, "Algorithmic Bias in Education", Int J Artif Intell Educ vol. 32, pp. 1052–1092, 2022
9. B. Berendt, A. Littlejohn, M. Blakemore, "AI in education: learner choice and fundamental rights", Learning, Media and Technology, vol. 45:3, pp. 312-324, 2020

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

10. W. Holmes, K. Porayska-Pomsta, K. Holstein, et al., “Ethics of AI in Education: Towards a Community-Wide Framework”, Int J Artif Intell Educ., vol. 32, pp. 504–526, 2022

11. W. Holmes, M. Maya, C. Fadel, “Artificial intelligence in education”, Data ethics: building trust: how digital technologies can serve humanity, pp. 621-653, Globethics Publications, 2023

12. A. Nguyen, H.N. Ngo, Y. Hong, et al., “Ethical principles for artificial intelligence in education”, in Educ Inf Technol vol. 28, pp. 4221–4241, 2023

8.3. AI and Education

13. A. Alam, “Should Robots Replace Teachers? Mobilisation of AI and Learning Analytics in Education”, 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), Mumbai, India, 2021, pp. 1-12

14. A. Alam, “Employing Adaptive Learning and Intelligent Tutoring Robots for Virtual Classrooms and Smart Campuses: Reforming Education in the Age of Artificial Intelligence”. In: Shaw, R.N., Das, S., Piuri, V., Bianchini, M. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Electrical Engineering, vol. 914. Springer, Singapore, 2022

15. H.K. Algabri, K. Kharade, R. Kamat, “Promise, Threats, And Personalization In Higher Education With Artificial Intelligence”, Webology, vol. 18/6, 2021

16. D. Baidoo-Anu, L. Owusu Ansah, “Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning”, Journal of AI, vol. 7(1), pp. 52-62, 2023

17. Digital Education Action Plan: https://education.ec.europa.eu/focus-topics/digital-education

18. L. Chen, P. Chen, Z. Lin, “Artificial Intelligence in Education: A Review,” in IEEE Access, vol. 8, pp. 75264-75278, 2020

19. T.S.K.F. Chiu, Q. Xia, X. Zhou, C. S. Chai, M. Cheng, “Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education”, in Computers and Education: Artificial Intelligence, vol. 4, 2023

20. B. Cope, M. Kalantzis, D. Searsmith, “Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies”, in Educational Philosophy and Theory, vol. 53:12, pp. 1229-1245, 2021

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

21. K.N. Gulson, et al., “Algorithms of Education: How Datafication and Artificial Intelligence Shape Policy”, University of Minnesota Press, 2022

22. C. Perrotta & Neil Selwyn, “Deep learning goes to school: toward a relational understanding of AI in education”, Learning, Media and Technology, vol. 45:3, pp. 251-269, 2020

23. Y. Punie, C. Redecker, “European Framework for the Digital Competence of Educators: DigCompEdu”, EUR 28775 EN, Publications Office of the European Union, Luxembourg, 2017

24. L. Tangi, C. van Noordt, M. Combetto, D. Gattwinkel, F. Pignatelli, “AI Watch. European landscape on the use of Artificial Intelligence by the Public Sector”, Joint Research Centre, 2022

25. R. Vuorikari, S. Kluzer, Y. Punie, Y., “DigComp 2.2: The Digital Competence Framework for Citizens – With new examples of knowledge, skills and attitudes”, Luxembourg: Publications Office of the European Union, 2022

26. B. Williamson, R. Eynon, “Historical threads, missing links, and future directions in AI in education, Learning, Media and Technology”, vol. 45:3, pp. 223-235, 2020

27. S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, H. Drachsler, “Are We There Yet? – A Systematic Literature Review on Chatbots in Education”, Front Artif Intell. 2021

28. S.J.H. Yang, H. Ogata, T. Matsui, N.-S. Chen, “Human-centered artificial intelligence in education: Seeing the invisible through the visible”, in Computers and Education: Artificial Intelligence, vol. 2, 2021

29. K. Zhang, A. Begum Aslan, “AI technologies for education: Recent research & future directions”, Computers and Education: Artificial Intelligence, vol. 2, 2021

30. J. Huang, S. Saleh, Y. Liu, “A Review on Artificial Intelligence in Education”, Academic Journal of Interdisciplinary Studies, vol. 10(3), 206, 2021

31. J. Huang, S. Saleh, Y. Liu, “A Review on Artificial Intelligence in Education”, Academic Journal of Interdisciplinary Studies, vol. 10(3), 206, 2021

32. G.-J. Hwang, H. Xie, B. W. Wah, D. Gašević, “Vision, challenges, roles and research issues of Artificial Intelligence in Education”, in Computers and Education: Artificial Intelligence, vol. 1, 2020

33. G.-J. Hwang & Ching-Yi Chang, “A review of opportunities and challenges of chatbots in education”, Interactive Learning Environments, vol. 31:7, pp. 4099-4112, 2023

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

34. N. Iacono, “Intelligenza Artificiale, perché è urgente cambiare il sistema educativo e come”, AgendaDigitale.eu, 2024

35. H. Khosravi, S. Buckingham Shum, G. Chen, C. Conati, Y. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, D. Gašević, “Explainable Artificial Intelligence in education”, in Computers and Education: Artificial Intelligence, vol. 3, 2022

36. H. Luan, P. Geczy, H. Lai, J. Gobert, SJH Yang, H. Ogata, J. Baltes, R. Guerra, Li P and Tsai C-C, “Challenges and Future Directions of Big Data and Artificial Intelligence in Education” in Front. Psychol., 11:580820, 2020

37. K. Malinka, M. Peresíni, A. Firc, O. Hujnák, F. Janus, “On the educational impact of ChatGPT: Is artificial intelligence ready to obtain a university degree?” In Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE ’23). Association for Computing Machinery, New York, NY, pp. 47–53, 2023

38. C.W. Okonkwo, A. Ade-Ibijola, “Chatbots applications in education: A systematic review, Computers and Education: Artificial Intelligence”, vol. 2, 2021

39. F. Ouyang, P. Jiao, “Artificial intelligence in education: The three paradigms”, in Computers and Education: Artificial Intelligence, vol. 2, 2021

40. F. Ouyang, L. Zheng, P. Jiao, “Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020”, Educ Inf Technol, vol. 27, pp. 7893–7925, 2022

41. J. Q. Pérez, T. Daradoumis, J.M. Marquès Puig, “Rediscovering the use of chatbots in education: A systematic literature review”, in Computer Applications in Engineering Education, 28(6), 1549–1565. https://doi.org/10.1002/cae.22326